Sistemas de **C**onhecimento **B**aseados em **D**ata **M**ining:

Aplicação à **A**nálise da **E**stabilidade de **E**struturas **M**etálicas

Hélder Quintela

Manuel Filipe Santos

Paulo Cruz

À Carla.

À memória do meu pai.

Agradecimentos

Realizar um trabalho como este implica empenhamento, concentração, rigor, e o acompanhamento e estímulo das pessoas que estão mais próximas. Na conclusão desta etapa da minha formação académica fico grato: ao Professor Manuel Filipe Santos, orientador e amigo, por ter influenciado o meu interesse pela área de Inteligência Artificial, pelos ensinamentos, motivação e aconselhamento durante este processo; ao Professor Paulo Cruz pela amizade e pela paciência na transmissão de conhecimentos da área de aplicação que me era totalmente desconhecida, pelo acompanhamento e disponibilidade manifestada em todos os momentos. Ao Professor Paulo Cortez um agradecimento pelas ajudas e conselhos.

Um agradecimento especial a um "*companheiro de jornada*", o João Pereira, que mais do que um colega, é um amigo, que partilhou comigo muitas e longas conversas: umas vezes de partilha de ensinamentos, outras de desabafos e motivação. Um agradecimento ao António Torres pelos auxílios fornecidos e pela amizade, e um desejo de felicidades e sucesso na prossecução dos seus trabalhos de Mestrado.

Não posso deixar neste momento de agradecer à minha família: à minha mãe e à minha irmã Sara, ao meu tio José Quintela.

Resumo

O ritmo de construção de estradas em Portugal, nunca foi tão intenso como nos últimos anos, conduzindo a que em muitas situações as obras de arte de Engenharia Civil se deteriorem a um ritmo mais acelerado, muitas vezes não dando tempo a que sejam reparadas ou substituídas.

A ausência de cuidado com este aspecto pode estar na origem de uma redução da vida útil das estruturas e no agravamento dos custos de manutenção.

Uma estrutura, poderá ser considerada durável se ao longo da sua vida mantiver a integridade, funcionalidade e qualidade estética, sem necessidade de grandes intervenções. A durabilidade não é contudo, uma propriedade intrínseca dos materiais, mas sim uma função relacionada com o desempenho destes sob determinadas condições ambientais.

A área da manutenção de estruturas e obras de arte da Engenharia Civil assume, assim, aspectos cada vez mais importantes permitindo evitar a sua deterioração e potenciando uma intervenção pró-activa.

A existência de dispositivos de recolha automática de dados (e.g., sensores, câmaras) com o auxílio de computadores controlados de forma remota permitiu nos últimos tempos a criação de repositórios de dados contendo valores da medição de um conjunto alargado de parâmetros. A partir destes dados, torna-se possível a criação de modelos de previsão e diagnóstico que suportam a peritagem e suporte à decisão no planeamento de acções de manutenção e vistoria.

Vários trabalhos foram realizados utilizando abordagens clássicas oriundas da estatística. Apesar destes avanços, os modelos gerados não captam convenientemente o conhecimento implícito nem revelaram as acuidades desejáveis.

As técnicas de Data Mining, surgem assim como alternativa a explorar, uma vez que apresentam características que permitem o estudo de problemas complexos, de difícil resolução através das abordagens mais convencionais, sendo por isso cada vez mais utilizadas nas diferentes áreas da engenharia.

Neste trabalho, é apresentado o processo de desenvolvimento e avaliação do desempenho de modelos de Inteligência Artificial em dois casos práticos: Previsão da Carga Crítica em estruturas metálicas com secções muito esbeltas sujeitas a cargas concentradas e, Previsão da Tensão Crítica de vigas com perfil em I, de inércia variável, tendo em vista o desenvolvimento futuro de sistemas de suporte à decisão para a manutenção de estruturas de Engenharia Civil. Uma primeira abordagem para o desenvolvimento de um Sistema de Conhecimento baseado em Data Mining para Análise da Estabilidade de Estruturas de Engenharia Civil é apresentada e testada, embebendo algoritmos e técnicas de Data Mining, através da utilização e implementação de uma especificação genérica e independente da plataforma de utilização.

Abstract

Nowadays, the construction effort in Portugal is very intense, and the absence of care conduces to the deterioration in the structures.

This absence of care implies the reduction of the service time of the structure, and increments the maintenance costs.

A Civil Engineering Structure can be considered stable if, during the service time, the features like integrity, functionality and aesthetic quality are preserved, without necessity of great maintenance interventions. The durability is not intrinsic to the materials, but a function related to the performance in several environmental constraints.

The interest in Civil Structures maintenance assumes aspects each time more relevant, in order to avoid a high level of deterioration, and enabling pro-active actions.

The existence of automatic devices for data collection (e.g., sensors and video cams) complemented with remote controlled computers, led to the creation of databases containing the registers of several attributes important to check the security of the civil structures. This enables the development of diagnostic models to support the audit and the maintenance actions.

Several works using classical approaches have been developed during the last years. Although this advances, the generated formulation uses inadequate security coefficients and presents large errors.

Data mining techniques are now a complementary option, due to the characteristics of these techniques that enables the resolution of complex problems, difficult of solving using classical approaches.

In this work, is presented the process of development and assessment of the performance of Artificial Intelligence models in two case studies: forecasting of the ultimate resistance of steel beams subjected to concentrated loads, and forecasting of shear strength of tapered steel plate girders. This opens room to the future development of decision support systems for Civil Structures maintenance. The prototype of the SCAE System, a Data Mining based Knowledge System for Study of the Stability of Steel structures is developed and tested in this work.

Conteúdo

Índice de Figuras

Índice de Tabelas

Índice de Algoritmos

Capítulo 1

Introdução

Na concepção e projecto de qualquer estrutura de engenharia civil devem ser ponderados factores da mais variada índole, tais como: estética, funcionalidade, deformabilidade, estabilidade, durabilidade, resistência e custo. Na generalidade dos casos esse exercício está condicionado à busca da solução mais segura e mais económica. Esta preocupação, aliada à evolução das propriedades dos materiais e dos meios de cálculo, tem conduzido à utilização de critérios de dimensionamento cada vez mais refinados.

No caso particular de estruturas metálicas com secções muito esbeltas, o erro apresentado pelas fórmulas de Previsão de Cargas Críticas de vigas de aço sujeitas a cargas concentradas é significativo, ficando a dever-se ao grande número de parâmetros que influenciam o comportamento e ao número insuficiente de dados experimentais que permitam efectuar uma análise paramétrica completa e calibrar, convenientemente, modelos simplificados.

No caso das vigas metálicas soldadas com perfil em I (Figura 1.1), de alma esbelta e inércia variável, foram propostos vários modelos para determinação da capacidade máxima resistente [Cruz e Guimarães, 2003], baseados nos modelos desenvolvidos para vigas de inércia

constante [Falby e Lee, 1976], [Davies e Mandal, 1979], [Takeda e Mikami, 1972]. Contudo, verifica-se que estes modelos apresentam algumas lacunas que fazem com que se tornem bastante conservadores para o dimensionamento deste tipo de estruturas. Perante este cenário, Zárate desenvolveu um modelo para o dimensionamento de vigas com perfil em I de alma esbelta e inércia variável [Zárate, 2002]. Baseando-se nos resultados obtidos por um modelo de elementos finitos e tomando como ponto de partida a teoria clássica da instabilidade, estabeleceu uma expressão analítica que permite determinar a tensão crítica de enfunamento.

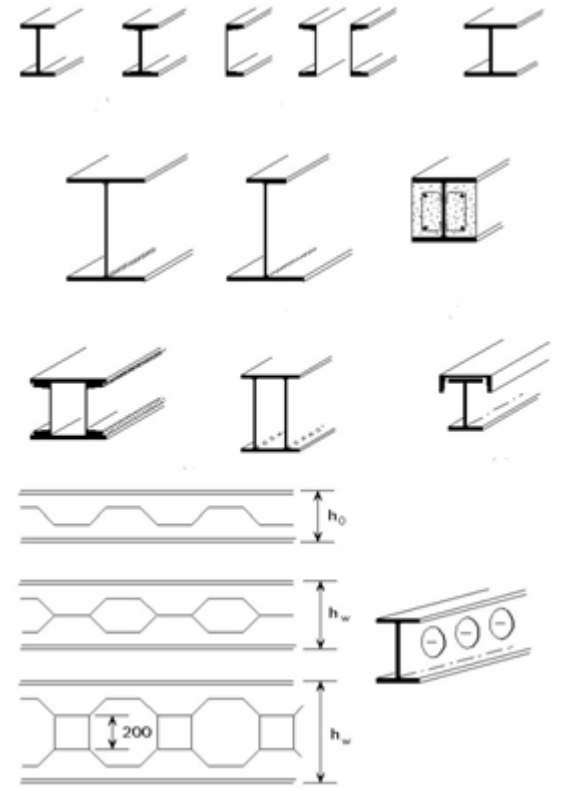


Figura 1.1 Diferentes tipos de vigas metálicas (perfil em I, H, T).

As técnicas de Data Mining (DM) são apresentadas aqui como uma alternativa a explorar para a Previsão da Carga Crítica de vigas de aço sujeitas a cargas concentradas e da Tensão Crítica de vigas com perfil em I de inércia variável, uma vez que apresentam

características que permitem o estudo/resolução de problemas complexos, de difícil resolução através de abordagens mais convencionais, sendo por isso cada vez mais utilizadas nas diferentes áreas da engenharia.

1.1 Motivação

O interesse pela aplicação de técnicas de Aprendizagem Automática (AA) para resolução de problemas na área da Engenharia Civil remonta aos anos 60, aquando da publicação do artigo "Artificial Intelligence and Structural Design" [Spillers, 1966], no qual era apresentada uma demonstração da aplicabilidade de regras elementares de aprendizagem na concepção de estruturas, concretamente a possibilidade de geração de regras através de exemplos anteriores. A inspiração para o trabalho de Spillers adveio da área médica, onde é possível estabelecer uma relação do tipo causal entre sintomas e doença, através da observação do historial clínico dos pacientes. Posteriormente, Jadid e Fairbairn [Jadid et al, 1994] utilizaram Redes Neuronais Artificiais para a análise estrutural, de forma a demonstrar a possibilidade de utilização deste tipo de modelos alternativamente ao cálculo numérico intensivo.

As técnicas de Inteligência Artificial (IA) têm sido utilizadas, entre outros, para a previsão de sismos [Bento et al., 1997], projecto de estruturas [Fonseca, 1999] [Hadi, 2003], detecção de anomalias em obras de arte [Bien, 2004]. Este interesse/utilização crescente é motivado pelas características íntrinsecas a estas técnicas, que permitem a resolução de problemas complexos, comuns na área da Engenharia Civil. Um dos modelos mais utilizados têm sido as Redes Neuronais Artificiais (RNA) devido a características como: aprendizagem não-linear, classificação multi-dimensional e tolerância ao ruído.

Os problemas de previsão constituem uma classe de problemas alvo de estudo na área de DM. Assumem particular importância em áreas como: economia [Cielen et al., 2004], saúde

[Silva et al., 2003a] [Silva et al, 2003b] [Silva et al., 2004], engenharia [Hadi, 2003], previsão de fenómenos raros [Ribeiro, 2003].

A previsão, surge associada normalmente a problemas de vital importância, motivada pela necessidade e pelos efeitos benéficos a que pode conduzir. Por exemplo, a previsão meteorológica permite adoptar medidas de defesa e prevenção capazes de evitar tragédias e antecipar soluções. No caso das estruturas de Engenharia Civil a previsão da carga crítica e da tensão crítica através de modelos de IA com acuidades superiores às fórmulas actuais, possibilitará o desenvolvimento de sistemas automáticos para a monitorização e planeamento de manutenção, permitindo evitar acidentes como os ocorridos recentemente em Portugal (e.g., queda da Ponte de Entre-os-Rios e da Ponte do IC 19) com consequências trágicas nos planos humano e económico, assim como reduzir os custos de manutenção.

Na maioria dos casos a descrição completa do fenómeno que se pretende prever é completamente conhecida, sendo a previsão governada por regras absolutamente determinísticas usadas para simulações do fenómeno. O desejo de compreender o passado e prever o futuro impulsiona a procura de leis que expliquem o comportamento de certos fenómenos ou acontecimentos. Se são conhecidas as equações analíticas que os explicam, estas podem ser utilizadas para prever o resultado de uma dada experiência desde que sejam conhecidas as condições iniciais. No entanto, na ausência de regras que definam o comportamento de um sistema, procura-se determinar o seu comportamento futuro a partir de observações concretizadas no passado.

Por exemplo, as marés constituem um fenómeno resultante da atracção gravitacional exercida pela Lua e pelo Sol sobre a Terra. Pelo seu carácter sistemático torna-se fácil de prever, mesmo por pessoas que desconhecem as razões científicas para este fenómeno, sabe-se que durante aproximadamente um dia o nível das águas atinge por duas vezes uma altura máxima na praia-mar, e por outras duas uma altura mínima na baixa-mar. A acrescentar a este conhecimento empírico é conhecida alguma relação entre as marés e as diferentes fases da Lua. Com Isaac Newton (1643 - 1727) as marés tornaram-se num fenómeno com uma descrição perfeitamente conhecida e explicada através de uma ciência - a física. Apesar de este ser um fenómeno natural governado por factores bem determinados, existe contudo um grau de indeterminação devido ao número de factores que nele influem. Existe um certo atraso na periodicidade das marés devido à inércia exercida pelo solo do oceano. A água é puxada pela

Lua ao mesmo tempo que é puxada pela Terra, interferindo aí as irregularidades do solo oceânico próximo de onde se observam e medem as marés.

Existem outros fenómenos que dada a sua grande irregularidade são denominados caóticos. A previsão meteorológica é um fenómeno desse tipo. Os modelos numéricos não conseguem reproduzir a enorme diversidade de factores que podem influenciar a evolução das condições atmosféricas. O que se sabe nem sempre é o que se pratica e a imprevisibilidade de certos acontecimentos introduz situações não contempladas pela descrição conhecida e dada como válida. Isso não coloca em causa a validade da descrição do fenómeno, estando à componente de imprevisibilidade associados fenómenos como: enchentes, secas, furacões, etc. Apesar da previsão poder não ser exacta, o resultado de computação da previsão é sempre um dos comportamentos possíveis da atmosfera.

O mesmo acontece com o mercado bolsista, um sistema altamente influenciável por todo o tipo de factores do meio envolvente (e.g., conjuntura económica, política, social). Esta característica tão íntriseca deste sistema, é o que contribui para que se tenha tornado tão desafiador.

A previsão de sismos também se apresenta como um fenómeno com uma componente indeterminística, uma vez que ocorrem de forma irregular, variando a localização geográfica, a frequência e a magnitude.

Na área da Engenharia Civil, a previsão de parâmetros como a tensão crítica ou a carga crítica estão condicionados pelo grande número de parâmetros que influenciam o comportamento das estruturas e número insuficiente de dados experimentais que permitam efectuar uma análise paramétrica completa. Estes factores, conduzem a fórmulas de previsão bastante conservadoras, que se traduzem na utilização de coeficientes de segurança inadequados que geram desperdícios de material e o consequente aumento do custo de obra.

Em contextos como estes referenciados anteriormente, as técnicas de IA, onde foram desenvolvidas um conjunto de técnicas para a resolução de problemas, inspiradas na Natureza (e.g., Redes Neuronais Artificiais, Algoritmos Genéticos e Evolucionários) assumem um papel relevante e fundamental na resolução deste tipo de problemas, pelo recurso a métodos de aprendizagem a fim de descrever o conceito do fenómeno com base nos seus registos históricos

e de fazer previsões com base na descrição obtida, contemplando componentes de difícel resolução através de fórmulas de cálculo. Na maior parte das situações pretende-se que o modelo de previsão funcione como um *perito* no painel de *conselheiros* no processo de tomada de decisão. Um *perito* que construiu o seu conhecimento a partir do histórico de dados, o que conduz normalmente à concepção de previsões mais fundamentadas que acabam por se reflectir na tomada de decisões.

A existência de dispositivos de recolha automática de dados instalados em estruturas de engenharia civil e a realização de alguns estudos [Fonseca, 1999] [Zárate, 2002], permitiram a criação de repositórios de dados que viabilizam o desenvolvimento de processos de Descoberta de Conhecimento em Bases de Dados, para geração por exemplo de modelos de previsão.

Os modelos de previsão desenvolvidos podem posteriormente ser incorporados em Sistemas de Conhecimento baseados em DM para o planeamento e manutenção de Estruturas de Engenharia Civil, possibilitando o prolongamento da sua vida útil e a melhoria das condições de serviço.

1.2 Objectivos

Este trabalho tem por objectivos: (i) desenvolver modelos para a previsão da carga crítica de vigas de aço sujeitas a cargas concentradas, e da tensão crítica de vigas em I de inércia variável, através da utilização técnicas de AA, com validade superior às formulações actualmente utilizadas, no âmbito de um processo de Descoberta de Conhecimento em Bases de Dados, e (ii) arquitectar e prototipar um Sistema de Conhecimento que incorpore os modelos de DM gerados (SCAE – Sistema de Conhecimento para a Análise da Estabilidade de Estruturas de Engenharia Civil), utilizando a plataforma JAVA [Sun, 2005], com incorporação da biblioteca Xelopes [Prudsys, 2003].

Os modelos de previsão a desenvolver neste trabalho deverão estar mais ajustados à realidade (i.e., maior precisão da previsão), o que possibilitará o dimensionamento mais eficaz

das estruturas em estudo pelos técnicos. O sistema SCAE deverá permitir a análise da estabilidade de estruturas metálicas de Engenharia Civil através da utilização dos modelos de previsão da carga crítica e da tensão crítica, devendo dispôr de um interface com o utilizador constituído por um painel de controlo, no qual o especialista de Engenharia Civil terá a possibilidade de ajustar os diferentes parâmetros que influenciam o comportamento de uma estrutura metálica.

1.3 Organização da Dissertação

A dissertação encontra-se estruturada em 5 capítulos, organizados da seguinte forma (excluindo o capítulo introdutório):

Capítulo 2

Sistemas de Conhecimento e Previsão da Carga e da Tensão Crítica em Estruturas de Engenharia Civil

É efectuada uma apresentação dos conceitos fundamentais sobre Sistemas de Conhecimento, bem como dos conceitos gerais e revisão de trabalhos anteriores da área de aplicação das técnicas de Data Mining neste trabalho (i.e., Engenharia Civil).

Capítulo 3

Descoberta de Conhecimento em Bases de Dados e Data Mining

São apresentados os objectivos, conceitos, tipos de abordagem, áreas relacionadas, metodologias e especificações, para a Descoberta de Conhecimento em Bases de Dados.

Capítulo 4

Modelos e Técnicas de Data Mining

São apresentados os principais modelos e técnicas utilizados no processo de Data Mining, evidenciando as propriedades associadas, com particular ênfase naqueles que foram usados na aquisição de conhecimento para o sistema SCAE.

Capítulo 5

SCAE – Sistema de Conhecimento baseado em Data Mining para Análise da Estabilidade de Estruturas Metálicas

Neste capítulo é apresentado o protótipo de um Sistema de Conhecimento para a Análise de Estabilidade de Estruturas Metálicas de Engenharia Civil (SCAE), baseado em técnicas de Data Mining, e uma abordagem para o desenvolvimento de modelos para a previsão da Carga Crítica em vigas de aço sujeitas a cargas concentradas e da Tensão Crítica de vigas em I de inércia variável.

Capítulo 6

Conclusões e Trabalho Futuro

São apresentas as conclusões do trabalho desenvolvido, identificando-se e as principais contribuições para as áreas de Tecnologias e Sistemas de Informação e de Engenharia Civil, sendo lançadas algumas linhas orientadoras para trabalho a desenvolver futuramente.

Capítulo 2

Sistemas de **C**onhecimento e **P**revisão da **C**arga e da **T**ensão **C**rítica em **E**struturas de **E**ngenharia **C**ivil

É efectuada uma apresentação dos conceitos fundamentais sobre Sistemas de Conhecimento, bem como dos conceitos gerais e revisão de trabalhos anteriores da área de aplicação das técnicas de Data Mining neste trabalho (i.e., Engenharia Civil).

2.1 Sistemas de Conhecimento

Durante as últimas décadas tem-se assistido a um interesse crescente nos Sistemas Baseados em Conhecimento (SC). Este sistemas, usados quando a formulação genérica do problema a ser resolvido computacionalmente é complexa e quando existe uma grande quantidade de conhecimento específico do domínio sobre como resolvê-lo, representam um importante avanço tecnológico na resolução de problemas complexos por computadores, que antes só eram resolvidos por seres humanos.

2.1.1 Conceitos

Os SC são programas de computador que usam o conhecimento (representado de forma simbólica ou sub-simbólica) para resolver problemas. No desenvolvimento de um SC, conhecimento e processo de resolução de problemas são pontos fulcrais. Estes sistemas manipulam conhecimento e informação de forma inteligente e são desenvolvidos para serem utilizados na resolução de problemas para os quais é necessária uma quantidade considerável de conhecimento humano e de especialização.

Segundo [Motta, 1998] um Sistema de Conhecimento deve possuir as seguintes características:

- o conhecimento actual sobre o problema deve estar explicitamente representado na Base de Conhecimento do sistema;
- a Base de Conhecimento deve ser usada por um agente (i.e., mecanismo de inferência) capaz de a interpretar;
- os problemas resolvidos pelos Sistemas de Conhecimento são aqueles sobre os quais não é conhecido um procedimento determinístico que garanta uma resolução efectiva em termos de recursos e de tempo.

Os SC diferem dos sistemas convencionais na forma como são organizados, como incorporam o conhecimento, como se executam e o grau de usabilidade.

No contexto dos Sistemas Inteligentes (Figura 2.1), os termos Sistemas de Conhecimento e Sistemas Especialistas (SE) são usados de forma indiferenciada. No entanto, é importante estabelecer uma distinção. De uma forma geral, os SC são sistemas capazes de resolver problemas usando conhecimento específico sobre o domínio de aplicação, enquanto os SE são SC que resolvem problemas resolvidos quotidianamente por um especialista humano. Os SC podem ser classificados como SE quando o desenvolvimento do mesmo é voltado para aplicações nas quais o conhecimento a ser manipulado se restringe a um domínio específico, apresentando um elevado grau de especialização.

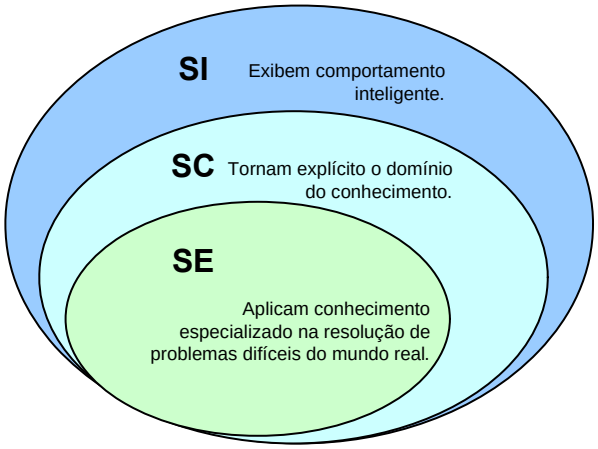


Figura 2.1 Sistemas Inteligentes – Adaptado de [Rezende, 2003].

Existem exemplos de aplicações de SC nas mais variadas áreas (e.g., medicina, ciência, engenharia), utilizados na resolução das mais variadas classes de tarefas (e.g., interpretação, classificação, monitorização, planeamento, projecto).

2.1.2 Estrutura

Apesar de nem todos os SC terem a mesma estrutura, a maior parte apresenta uma estrutura semelhante, composta pelos seguintes módulos: Núcleo do Sistema de Conhecimento ou *shell*, Base de Conhecimento, Memória de Trabalho, Base de Dados e Interface com o utilizador [Laudon et al., 2002].

O núcleo do Sistema de Conhecimento desempenha as funções principais, sendo responsável: pelo controlo da interacção com o utilizador ou com entidades externas, pelo processamento do conhecimento utilizando uma linha de raciocínio, pelos mecanismos de inferência do sistema, pela justificação/explicação das conclusões obtidas a partir do raciocínio.

O núcleo do SC é composto por 3 sub-módulos interdependentes: o módulo de recolha de dados, o motor de inferência e o módulo de explicações.

A Base de Conhecimento contém a descrição do conhecimento necessário para a resolução do problema. Isto inclui informação sobre: domínio de aplicação, heurísticas e métodos de resolução de problemas. A descrição do conhecimento pode estar representada de forma simbólica (e.g., regras de decisão), ou de forma sub-simbólica (e.g., redes neuronais).

A Memória de Trabalho representa a área de trabalho de um SC, na qual são registadas todas as respostas fornecidas durante as interacções realizadas com o sistema, evitando que o utilizador responda à questão várias vezes, em diferentes ocasiões. A memória de trabalho do sistema armazena condições iniciais, conclusões intermédias, bem como soluções finais.

O sistema pode interagir com uma Base de Dados para leitura/escrita de dados e/ou informações.

A interface de um SC é responsável pela interacção entre o Sistema de Conhecimento e o utilizador, proporcionando a comunicação em ambas as direcções.

2.1.3 Desenvolvimento

O desenvolvimento de um SC é um processo iterativo, estando altamente dependente dos recursos disponíveis e de como estes recursos estão organizados e são geridos. Durante este processo deparam-se obstáculos relacionados com a complexidade dos problemas a resolver, a forma de representação do conhecimento e a aquisição do conhecimento. Derivado da complexidade dos problemas resulta a motivação para a aplicação de técnicas com origem na área da Inteligência Artificial.

As principais fases do desenvolvimento de um SC (Figura 2.2) são:

- Planeamento;
- Aquisição do Conhecimento;
- Implementação;
- Validação e manutenção.

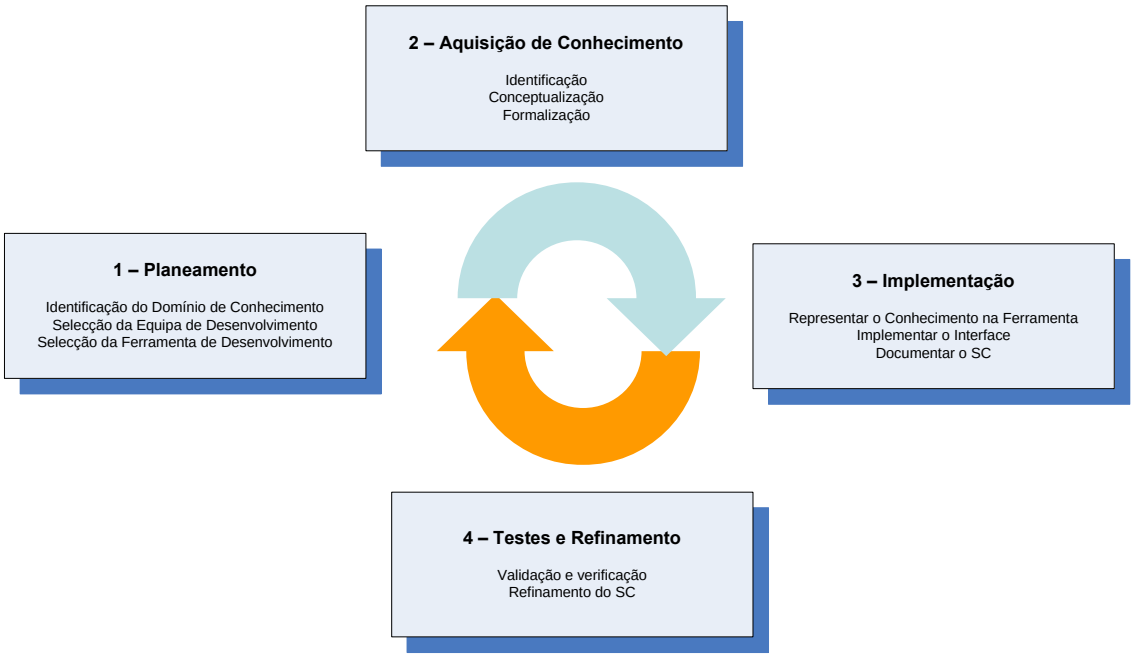


Figura 2.2 Fases de Desenvolvimento de um Sistema de Conhecimento.

Na fase de Planeamento do SC, que integra as tarefas de: **(i)** identificação do domínio de conhecimento, **(ii)** selecção da equipa de desenvolvimento, **(iii)** selecção das tecnologias de desevolvimento, são descritos e identificados os seguintes itens: domínio de conhecimento, terminologia, referências, e conceitos relacionados com o domínio de conhecimento, com o objectivo de facilitar a compreensão das pessoas que estão envolvidas no processo de desenvolvimento do SC. Nesta fase de planeamento é realizada a análise funcional, identificando os módulos do sistema, as entradas e saídas, os elementos da equipa de desenvolvimento e as tecnologias a utilizar.

A segunda fase – Aquisição do conhecimento -, que compreende as tarefas de: **(i)** identificação, **(ii)** conceptualização e **(iii)** formalização, tem como objectivo adquirir o conhecimento armazenado na base de conhecimento.

Na terceira fase que corresponde à Implementação, o objectivo principal é a representação formal do conhecimento adquirido na fase anterior do processo de desenvolvimento do SC. Para a representação utiliza-se a estrutura de Representação do Conhecimento seleccionada na primeira fase. É realizada nesta fase de implementação: a codificação do sistema através de linguagens ou ferramentas adequadas, documentação do sistema, criação de manuais e implementação da interface do sistema com o utilizador.

A última fase do desenvolvimento de um Sistema de Conhecimento é a fase de Testes e Refinamento, que envolve a validação e verificação do sistema, de modo a assegurar que: o sistema funciona correctamente, fornece resultados verdadeiros e satisfaz os requisitos do cliente. Para além destes objectivos, envolve a realização de eventuais alterações nos requisitos do sistema, realçando-se por exemplo a aquisição contínua de conhecimento e a monitorização do funcionamento do SC.

Uma das fases mais críticas no desenvolvimento de um Sistema de Conhecimento é a Aquisição de Conhecimento para o desenvolvimento da Base de Conhecimento, devendo ainda considerar-se uma outra fase que não está representada na figura 2.16 de manutenção do SC, uma vez que a aplicação e evolução temporal vai implicar alterações no sistema para que este continue a executar de forma eficaz as tarefas que estiveram na origem do seu desenvolvimento.

2.1.4 Aquisição de Conhecimento

O processo de Aquisição de Conhecimento pode ser definido segundo duas abordagens: uma mais antiga que postula que a Aquisição de Conhecimento consiste na transferência e transformação do conhecimento especializado com potencial para a resolução de problemas de uma fonte de conhecimento para um programa, enquanto a mais actual advoga que se trata de um processo de modelação de problemas e soluções pertinentes para um domínio específico.

Figura 2.3 Fases do processo de Aquisição de Conhecimento.

Buchanan, Barstow e Bechtel [Buchanan et al., 1983] identificaram cinco fases no processo de Aquisição de Conhecimento (Figura 2.3): identificação, conceptualização, formalização, implementação e teste, que apesar de estarem representadas de forma sequencial são realizadas de forma iterativa e evolucionária. Neste processo a figura central é o Engenheiro do Conhecimento que observa e interpreta conhecimento sobre o domínio, o problema e estratégias de resolução.

A primeira fase – **Identificação** – assemelha-se à Análise de Requisitos em Engenharia de Software e tem como objectivo primário procurar elementos do domínio que permitam identificar: a classe do problema que o SC deverá resolver; os dados sobre os quais irá operar; quais os critérios de classificação das soluções nos contextos de funcionamento; e a forma como o problema deve ser resolvido. Para atingir os objectivos desta fase procede-se à recolha de bibliografia sobre o domínio, a entrevistas não-estruturadas (brainstorming[1]) para apurar necessidades, complexidade da tarefa, terminologia utilizada, capacidade de cada especialista em tornar explícito o conhecimento do domínio, e a entrevistas com utilizadores para formular o modelo de interacção utilizador-sistema.

Na fase de **Conceptualização** são formulados os conceitos importantes do problema e identificados os relacionamentos existentes entre esses conceitos e elaborada uma ontologia[2] de forma informal, recorrendo-se a entrevistas estruturadas a partir do material recolhido anteriormente, a casos para modelação e teste do SC e à observação do especialista no seu trabalho.

Uma vez efectuada a selecção dos conceitos e formalizadas as relações importantes para o problema, na fase de **Formalização** é realizada a modelação computacional do problema e a formalização da ontologia através de uma linguagem formal (e.g., Lógica) de primeira ordem e suas extensões, frames, e regras de produção envolvendo: definição do modelo de tarefa a ser adoptado; escolha da linguagem de representação para modelar o sistema; definição do espaço de pesquisa do problema; definição do espaço de soluções do problema; definição dos métodos de pesquisa de soluções; e identificação das limitações do sistema.

[1] Técnica usada para maximizar a geração de ideias provenientes de um grupo de pessoas, ideias relacionadas com as causas ou soluções de um problema, ou, direccionadas para a criação de novos produtos ou inovações.

[2] Conjunto de termos que designam conceitos e relações, juntamente com as suas definições.

Após a fase de formalização é realizada a **Implementação** que consiste na modelação computacional do problema, envolvendo as seguintes actividades: selecção da linguagem de programação (e.g., Prolog); selecção da ferramenta (e.g., ART); programação do SC; prototipação; e validação pelo especialista do domínio de aplicação.

Estando o SC desenvolvido é necessário proceder à realização de **Testes** para verificação da usabilidade e utilidade do sistema, verificando-se se este corresponde aos objectivos que estiveram presentes na sua construção, através da submissão ao sistema de um conjunto representativo de casos de teste, simulando a utilização do SC no ambiente real durante um determinado período de tempo.

Apesar de não se encontrar identificada no modelo proposto por [Buchanan et al., 1983], que representa o processo desde a necessidade à produção do SC, deve ser considerada uma fase de Manutenção, para que seja possível implementar modificações ao SC que permitam a incorporação de novos requisitos provocados por alterações no ambiente ou problema e de conhecimento situado (i.e., adquirido à medida que o problema é resolvido), e a manutenção da Base de Conhecimento.

2.1.5 Linguagens para Aquisição de Conhecimento

O processo de Aquisição de Conhecimento descrito na secção anterior, envolve a comunicação do conhecimento do especialista para o Engenheiro do Conhecimento, e deste para um agente computacional. Durante as fases de aquisição o conhecimento pode ser descrito de várias formas: na fase de identificação o conhecimento é usualmente transferido através de linguagem natural, enquanto na fase de implementação o conhecimento se encontra normalmente expresso numa linguagem de programação.

Na fase de Identificação é normalmente utilizada Linguagem Natural devido à facilidade de comunicação e registo, apresentando contudo desvantagens tais como grande número de vocábulos que podem resultar na ocorrência de ambiguidades e/ou imprecisões.

Devido a estas limitações, na fase de Conceptualização são utilizadas **Linguagens Diagramáticas**, que recorrendo a gestos, imagens, figuras, esquemas e diagramas, facilitam a identificação de inconsistências, apresentando um elevado poder de comunicação, e facilitando um entendimento amplo e global sobre uma Representação do Conhecimento.

À medida que aumenta a necessidade do conhecimento estar representado de uma forma mais consistente, aumenta o grau de formalização das linguagens utilizadas. Na fase de Formalização são utilizadas **Linguagens Semiformais** (e.g., HTML [W3Ca, 2004], XML [W3Cb, 2004]) - que combinam notações formais com representação informal como a linguagem natural -, e **Linguagens Formais** que expressam o conhecimento de forma precisa, consistente e não ambígua. Existem três tipos comuns de linguagens formais: lógica (i.e., o conhecimento é representado através de um conjunto de fórmulas em lógica de predicados, difusa e modal); sistemas de produção (i.e., regras do tipo Se <condição> Então <acção>); e estruturados (i.e., grafos onde os nodos e arcos possuem semântica fixa – frames e objectos -, ou variada – redes semânticas).

Na fase de Implementação são utilizadas Linguagens de Programação[3] (linguagens formais) que apresentam como principais características a precisão e não ambiguidade. Para a utilização deste tipo de linguagem formal são necessários conhecimentos sobre programação e engenharia de software. O SC pode ser especificado através da linguagem UML (Unified Model Language) [OMG, 2004].

2.1.6 Técnicas para Aquisição de Conhecimento

A Aquisição de Conhecimento é uma das tarefas mais críticas no desenvolvimento de um SC. Ao longo do tempo têm sido realizados esforços no sentido de sistematizar e/ou automatizar este processo através da utilização de: **técnicas manuais** oriundas da Psicologia e da Análise de Sistemas, baseadas em entrevistas, em acompanhamento ou em modelos; **técnicas** de aquisição **semi-automáticas** baseadas em técnicas oriundas da Psicologia (e.g., AGR – Análise de Grades de Relatos), na reutilização de modelos (e.g., KIF – Knowledge Interchange Format, linguagem universal para expressar e transportar conhecimento entre bases de conhecimento), e em ontologias reutilizáveis (e.g., Ontolingua, OCML); e **técnicas automáticas** baseadas na aprendizagem automática e em tecnologias de Data Mining.

Neste trabalho, a Aquisição de Conhecimento para construção da Base de Conhecimento do sistema SCAE será realizada através da utilização de técnicas de Aquisição de Conhecimento automáticas.

[3] As linguagens de programação são descritas por gramáticas livres de contexto, o que as torna mais simples e mais fáceis de interpretar relativamente a outros tipos de linguagens formais.

2.1.7 Trabalhos Relacionados

A aplicação de Sistemas de Conhecimento tem conhecido um interesse crescente demonstrado pela diversidade de áreas (ver Tabela 2.1) em que têm sido aplicados (e.g., planeamento de sistemas hidrotérmicos, fusão de sensores, diagnóstico em isolamento de transformadores de potência, diagnóstico de falhas de transformadores, classificação de imagens de satélite [Rezende, 2003]). Na área da Engenharia Civil existem exemplos de aplicação para definição da constituição de cimentos (ver figura 2.8) [Foo et al., 1993] [Akhras et al., 1994] [Zain et al., 2005].

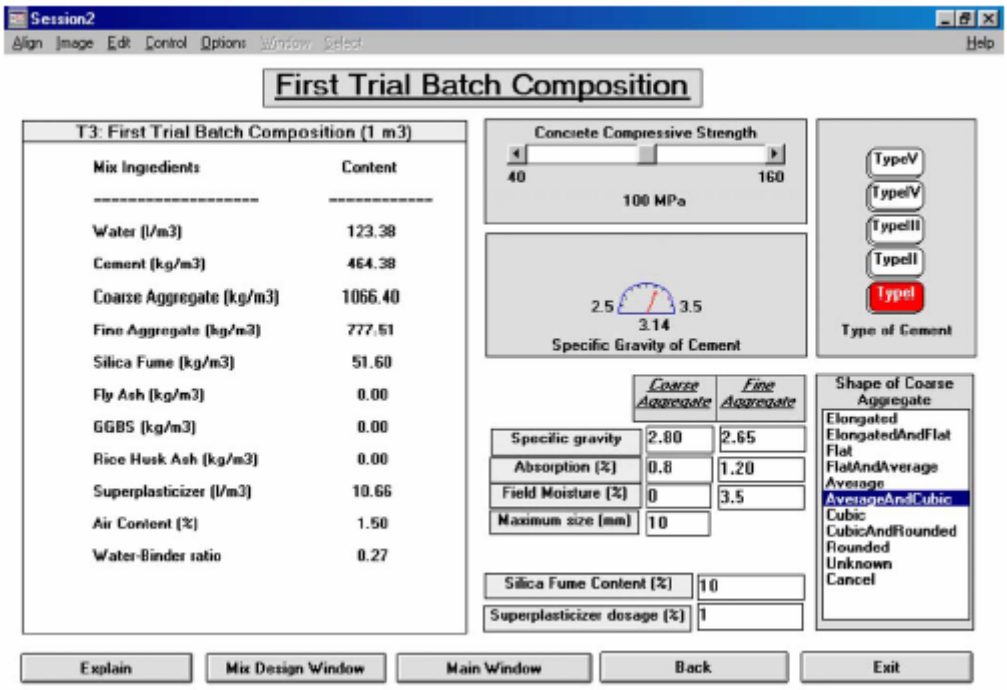


Figura 2.4 Interface do Sistema Inteligente HPCMIX
para determinação da composição óptima do cimento [Zain et al., 2005].

Tabela 2.1 Sistemas baseados em Conhecimento – Áreas de Aplicação (adaptado de [Liao, 2005]).

Ano	Área de Aplicação
1995	Análise de falhas em actividades de Engenharia
1997	Análise Financeira
1997	Gestão do Conhecimento
1997	Representação do Conhecimento
1999	Previsão de condições meteorológicas
1999	Definição da composição do aço
1999	Gestão estratégica
2001	Controlo de processos químicos
2002	Planeamento de actividades agrícolas
2004	Planeamento urbano
2005	Definição da composição do cimento

O estudo das características e das áreas de aplicação de SC abre perspectivas para o desenvolvimento de um Sistema de Conhecimento baseado em Data Mining para a Análise da Estabilidade de Estruturas Metálicas.

Neste projecto um dos objectivos primordiais e simultaneamente crítico, é o desenvolvimento de modelos de Previsão da Carga Crítica de vigas de aço com perfil em I, e da Tensão Crítica de vigas em I de inércia variável, alternativos às formulações analíticas actualmente existentes. Estes modelos de previsão deverão ser desenvolvidos na fase de aquisição do conhecimento através de técnicas automáticas, nomeadamente, Redes Neuronais Artificiais.

2.2 Previsão da Carga Crítica e da Tensão Crítica em Estruturas de Engenharia Civil

Uma estrutura de Engenharia Civil (e.g., ponte, edifício) deve ser projectada para suportar a carga a que vai ser submetida (e.g., tráfego, fenómenos meteorológicos), da forma mais segura e utilizando a menor quantidade de material possível (Figura 2.9). As estruturas de aço têm-se revelado cada vez mais eficientes e económicas, o que contribuiu para a sua utilização crescente.

A procura da melhor solução, requer a utilização de perfis mais leves e esbeltos. Para que esse objectivo seja alcançado é necessário uma boa calibração das fórmulas de previsão da carga crítica, de forma a evitar a utilização de coeficientes de segurança inadequados, que quando utilizados por excesso geram desperdícios de material e o consequente aumento do custo da obra, e quando utilizados por defeito podem provocar situações de rotura dos elementos. Em construção metálica é relativamente frequente que os elementos estruturais não apresentem as dimensões geométricas definidas na fase de projecto, devido a potenciais erros de fabrico ou a efeitos causados pelas condições ambientais durante a sua exploração.

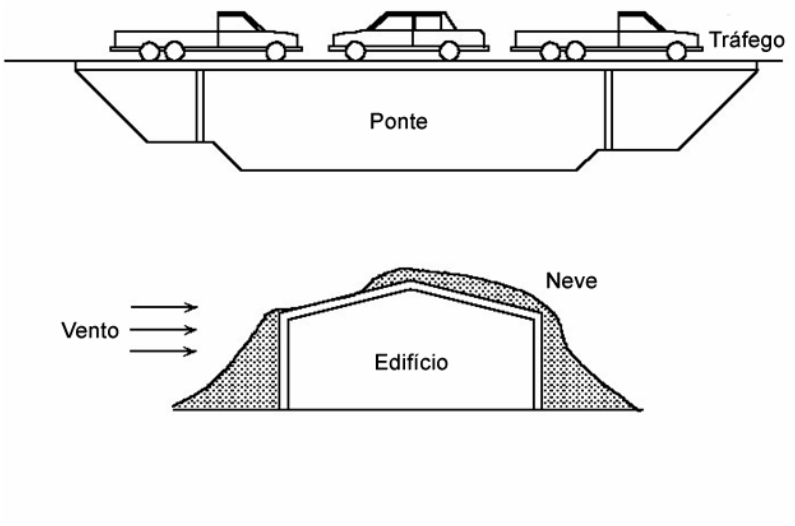


Figura 2.5 Diferentes tipos de cargas suportadas por Estruturas de Engenharia Civil.

Ao longo dos anos têm sido desenvolvidas diversas teorias no sentido de procurar uma melhor solução, no entanto, o erro médio das fórmulas de previsão da carga crítica de vigas sujeitas a cargas concentradas é ainda superior a 20% [Fonseca, 1999].

As principais causas desta dificuldade de previsão devem-se ao grande número de parâmetros que influenciam o comportamento de uma viga sujeita a cargas concentradas, e ao número insuficiente de dados experimentais presentes na literatura que permitam efectuar uma análise paramétrica completa.

Uma das alternativas, para a previsão da carga crítica, seguida no contexto deste trabalho, reside na utilização de modelos de previsão baseados em algoritmos de Aprendizagem Automática (AA) (e.g., Árvores de Regressão, Redes Neuronais Artificiais), gerados no contexto de um processo de Descoberta de Conhecimento em Bases de Dados (vd., Capítulo 3).

Uma das desvantagens, referenciadas na utilização de RNAs relativamente a outras abordagens (e.g., Árvores de Decisão, Árvores de Regressão) é a dificuldade na sua interpretação, contudo tal afigura-se como possível [Gallant, 1995], embora não faça parte dos requisitos deste projecto.

Neste trabalho as técnicas de AA são aplicadas a dois casos de estudo – previsão da carga crítica de uma viga de aço sujeita a cargas concentradas e previsão da tensão crítica de vigas em I de inércia variável, objecto de detalhe nas secções seguintes.

2.2.1 Carga Crítica de uma Viga de Aço Sujeita a Cargas Concentradas

Quando é aplicada uma carga concentrada no plano da alma de perfis de aço (Figura 2.10), pode ocorrer um fenómeno de instabilidade local. Os exemplos mais comuns são as vigas secundárias descarregando sobre vigas principais ou a carga de compressão descarregando no banzo de uma coluna. Nestes casos, onde a localização da carga é conhecida, podem ser usados reforços transversais para aumentar a resistência, mas devem ser evitados por razões económicas. Quando são consideradas cargas móveis, como no caso das pontes rodoviárias ou ferroviárias (Figura 2.11), torna-se imprescindível conhecer a resistência de almas não reforçadas para uma carga localizada de compressão.

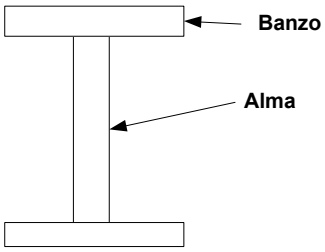


Figura 2.6 Representação de uma Viga em I.

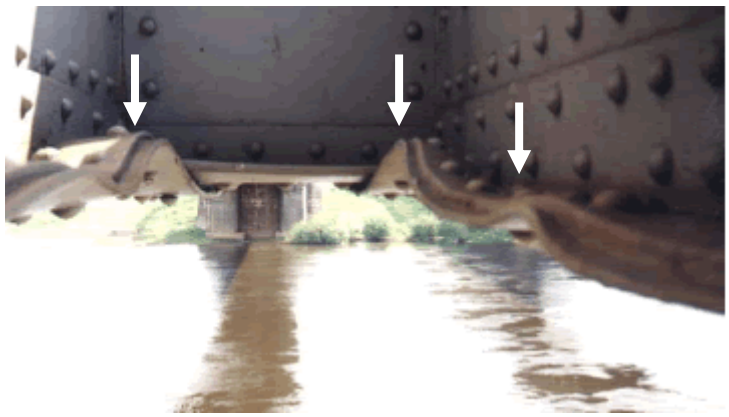


Figura 2.7 Vigas de aço deformadas numa ponte [Bien, 2004].

As cargas concentradas podem ocorrer perto dos apoios ou no interior das vigas, podendo estar aplicadas num dos banzos, ou nos dois, como é o caso de duas vigas descarregando a compressão em cada um dos banzos da coluna.

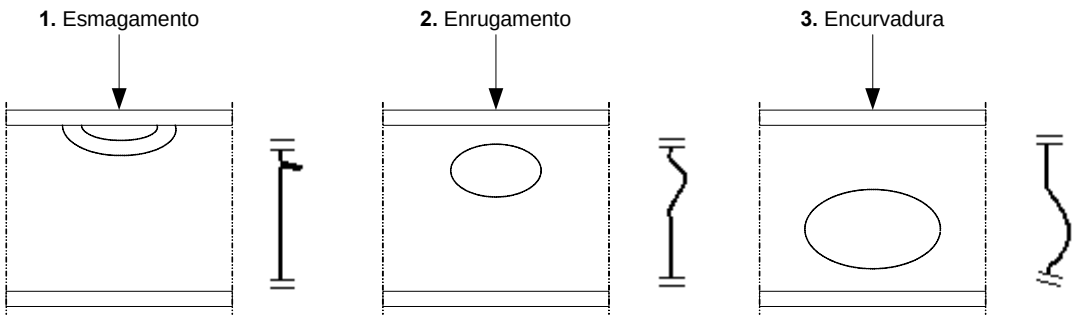


Figura 2.8 Modos de colapso devido à introdução da carga.

A resistência de uma alma não reforçada a forças transversais, aplicadas através de um banzo, é condicionada por um dos seguintes modos de colapso (Figura 2.12):

1. **Esmagamento** da alma junto ao banzo, acompanhado de deformação plástica do banzo;
2. **Enrugamento** da alma sob a forma de encurvadura localizado da alma junto ao banzo, acompanhados de deformação plástica do banzo;
3. **Encurvadura** da alma abrangendo a maior parte da altura da peça.

A carga crítica de uma viga de aço sujeita a cargas concentradas depende de vários parâmetros geométricos e materiais (Figura 2.13): distância entre os reforços (*a*), altura da alma (*h*), espessura da alma (*tw*), largura do banzo (*bf)*, espessura do banzo (*tf*), comprimento da zona de aplicação da carga (*c*), tensão de cedência do banzo (σ_{yfl}) e tensão de cedência da alma (σ_{yw}).

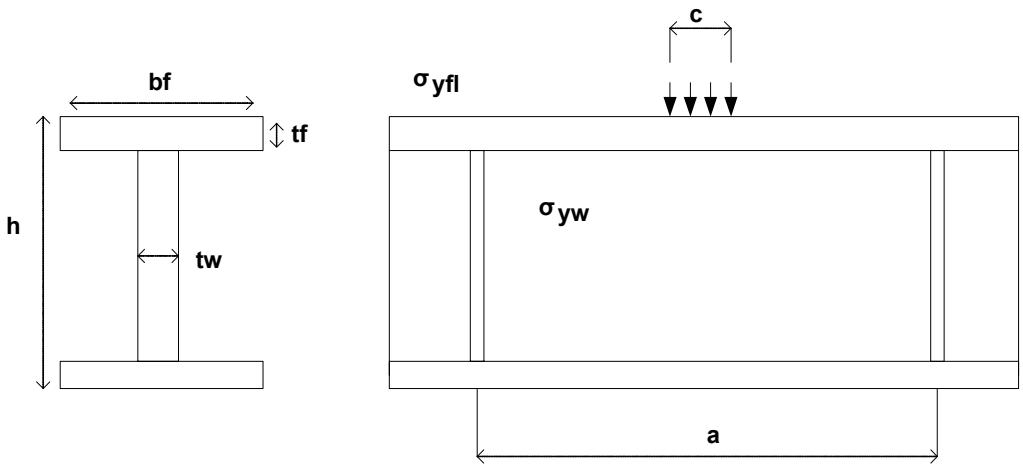


Figura 2.9 Parâmetros Geométricos e Materiais de uma viga de aço com perfil em I.

Trabalhos relacionados

Diversos trabalhos teóricos, numéricos e experimentais foram desenvolvidos para uma melhor formulação deste problema, no entanto, o erro das expressões matemáticas de previsão da carga crítica de vigas sujeitas a cargas concentradas é ainda considerável, fornecendo erros superiores a 20% [Fonseca, 1999].

Um dos primeiros trabalhos experimentais foi desenvolvido por Lyse e por Godfrey [Lyse, 1935], que identificaram as razões para o colapso das vigas e observaram um problema de instabilidade local da alma, a partir de testes em 6 perfis laminados com esbeltez da alma (*h*/*tw*) próximo de 52. Em função das altas tensões observadas nos testes, concluiram que o efeito de enrugamento obtido era um efeito localizado. A equação proposta para modelar este fenómeno de encurvadura ainda hoje é a base de projecto em diversas normas internacionais. Contudo, essa equação não considera a instabilidade provocada pelas características geométricas do elemento estrutural que pode gerar problemas de encurvadura ou esmagamento da alma.

Entre as pesquisas mais relevantes encontram-se os trabalhos de Roberts [Roberts et al., 1978] [Roberts, 1981] [Roberts et al., 1988] [Roberts et al., 1997]. Em 1978, Roberts e Rockey criaram um método para prever a carga última baseando-se num mecanismo de rótulas plásticas. Roberts, em 1981, efectou uma uma nova série de testes para estudar a influência das dimensões do banzo e da altura da alma na carga última da viga, propondo uma nova equação, modificada pela primeira vez em 1988, e por último em 1997 (Fórmula 2.1), que transforma as soluções por mecanismos em equações mais simples. Os resultados foram comparados com ensaios experimentais.

(Fórmula 2.1)

$$P_{cr} = \left[\; 1.1 tw^2 \left(E \sigma_y^w \right)^{0.5} \; \left(\frac{tf}{tw} \right)^{0.25} \left(1 + \frac{ce * tw}{h * tf} \right) \right] \frac{1}{F}$$

onde $ce = c + 2\ tf$ e F (factor de segurança) = 1.45

De acordo com os trabalhos mais recentes, a carga última é aproximadamente proporcional ao quadrado da espessura da alma. As dimensões do banzo, a altura da viga, as propriedades do material e o comprimento carregado têm importância secundária e a sua influência está ligada à espessura da alma.

Recentemente [Fonseca, 1999] desenvolveu um trabalho com o objectivo de analisar e investigar o comportamento estrutural de cargas concentradas, através de uma análise paramétrica para identificar a influência dos diversos parâmetros no problema, com utilização de Redes Neuronais Artificiais. Os resultados obtidos permitiram concluir que:

- A não consideração da influência do factor *a/h* e da largura do banzo nas fórmulas existentes foram a principal razão da diferença encontrada entre os resultados da aplicação destas fórmulas matemáticas e os resultados obtidos no estudo. Desta forma, parte do erro presente nas fórmulas existentes de previsão da carga crítica deve-se à não consideração destes parâmetros;

- A carga última é mais afectada pelo factor *a/h* quando este é inferior a 2.5;

- Os resultados confirmaram que a altura da alma tem uma influência secundária na carga última, e que este parâmetro não deve ser avaliado isoladamente;

- A equação proposta por Lyse e Godfrey [Lyse, 1935] é muito conservadora na previsão da carga crítica de vigas sujeitas a cargas concentradas.

2.2.2 Tensão Crítica de Vigas em I de Inércia Variável

Desde a primeira utilização por Brunel no século XIX de vigas de aço na construção de pontes, têm sido desenvolvidos diversos estudos para a compreensão do seu comportamento de forma a obter soluções mais seguras e económicas.

As vigas metálicas soldadas de alma esbelta e inércia variável são apropriadas para vencerem grandes vãos e resistirem a cargas elevadas, apresentado um reduzido peso próprio

devido à variação da inércia. Por este facto estas são muitas vezes utilizadas em zonas onde existe uma forte variação do momento flector. Normalmente, a variação da inércia consegue-se através da redução da altura da alma, mantendo horizontal o painel do banzo superior. A optimização das soluções consegue-se através da variação das dimensões dos banzos e/ou da alma.

A variação da inércia, obtida normalmente através da redução da altura da alma mantendo horizontal o painel do banzo superior, conduz a uma diminuição do peso próprio e por conseguinte a um dimensionamento mais eficaz e económico, ficando o sistema estrutural e esteticamente beneficiado, com aparência esbelta (Figura 2.14).

Figura 2.10 Obras de arte com vigas metálicas soldadas de alma esbelta e inércia variável.

No entanto, o aumento da esbelteza torna os painéis da alma das vigas mais susceptíveis aos fenómenos de instabilidade para valores de carga inferiores aos que originaria a plastificação da secção (Figura 2.15 e Figura 2.16).

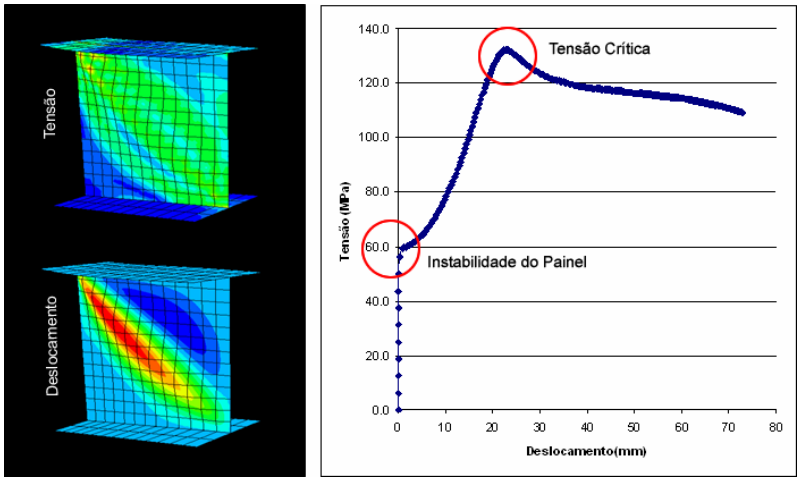


Figura 2.11 Comportamento da alma de um elemento estrutural submetido a tensões tangenciais.

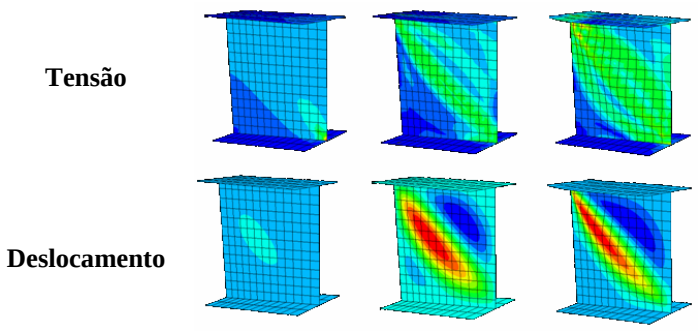


Figura 2.12 Comportamento da alma de um elemento estrutural submetido a tensões tangenciais analisada através de três mecanismos de resistência.

Para a Previsão da Tensão Crítica são considerados os seguintes parâmetros geométricos (Figura 2.17): relação entre a distância dos reforços transversais (*a*) e altura da alma (h_1), esbeltez (largura do banzo (*bf*)/ espessura do banzo (*tf*), relação largura do banzo (*bf*) / altura da alma (h_1), pendente do banzo inferior e espessura da alma (*tw*).

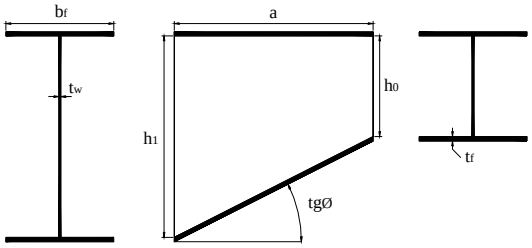


Figura 2.13 Parâmetros geométricos considerados nas vigas com perfil em I de alma esbelta e inércia variável.

Para este tipo de vigas foram propostos vários modelos para determinação da carga máxima resistente [Falby e Lee, 1976] [Davies e Mandal, 1979] [Takeda e Mikami, 1972], baseados nos modelos desenvolvidos para vigas de inércia constante. Contudo, estes modelos apresentam algumas lacunas que fazem com que se tornem bastante conservadores para o dimensionamento deste tipo de estruturas.

Os estudos de capacidade última de vigas metálicas de inércia variável realizados por Falby e Lee permitiram concluir que os modelos de capacidade última de vigas de inércia constante podiam ser utilizados em situações de vigas com painéis de alma com altura variável, desde que o ângulo de inclinação do banzo inferir fosse pequeno. Para ângulos de inclinação do banzo inferior consideráveis, Falby & Lee propuseram a utilização de um modelo simplificado de cálculo (Figura 2.18), que considera que a tensão crítica de enfunamento da alma pode ser calculada mediante as expressões da teoria clássica de chapas rectangulares, adoptando como valor da altura o valor médio das alturas maior e menor do painel de alma com altura variável. Considera-se que a capacidade pós crítica do painel da alma se deve ao desenvolvimento de um campo diagonal de tracções. Falby e Lee concluíram que a distribuição do campo diagonal de tracções na alma é conservadora, considerando, contudo, que através do modelo proposto se obtinham melhores resultados que ao utilizar as expressões de vigas com inércia constante, no caso de vigas de inércia variável.

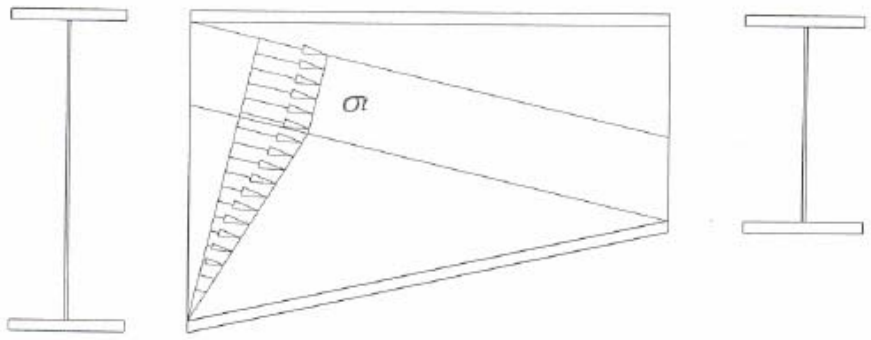
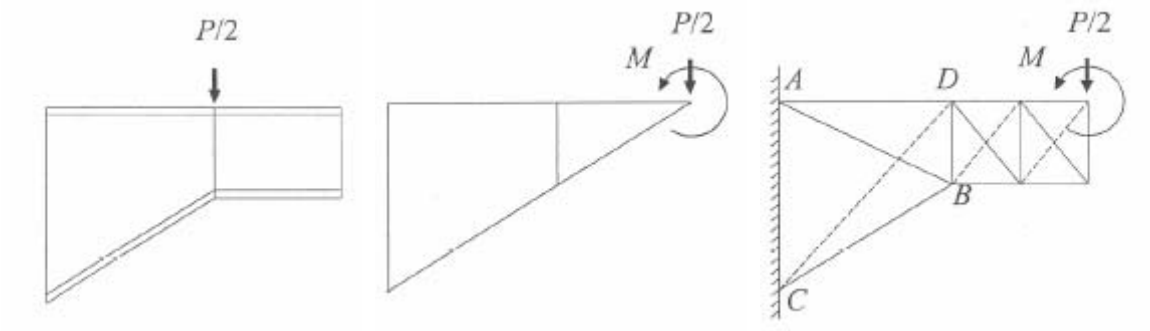


Figura 2.14 Modelo proposto por Falby e Lee.

O modelo de capacidade última para vigas de inércia variável de alma esbelta proposto por Davies e Mandal (Figura 2.19), que teve por base ensaios experimentados, levou a que considerassem que a capacidade última do painel da alma das vigas de inércia variável pode ser definida através da sobreposição de dois estados de tensão. O primeiro, correspondente à tensão crítica de enfunamento do painel da alma e, o segundo, referente à capacidade pós crítica da alma. Tal como no modelo proposto por Falby e Lee, a tensão crítica de enfunamento do painel da alma obtém-se mediante as expressões da teoria clássica.



a) viga de inércia variável solicitada por uma carga P/2 **b)** discretização através de carregamento num ponto fictício **c)** treliça utilizada para o desenvolvimento do modelo

Figura 2.15 Modelo proposto por Davies e Mandal [Zárate, 2002].

Um dos modelos mais recentes de capacidade última de vigas de inércia variável foi proposto por Takeda e Mikami (Figura 2.20), baseado na teoria proposta por Chern e Ostapenko [Chern & Ostapenko, 1969] para vigas metálicas com painéis de alma com, altura constante. A capacidade última está definida pela sobreposição de dois estados de tensão, correspondentes às fases pré e pós crítica. Contudo, ao contrário dos modelos propostos por Falby e Lee, e por Davies e Mandal, a tensão crítica de enfunamento é determinada com base na formulação resultante da aplicação dos modelos de elementos finitos ao estudo da instabilidade de uma placa trapezoidal.

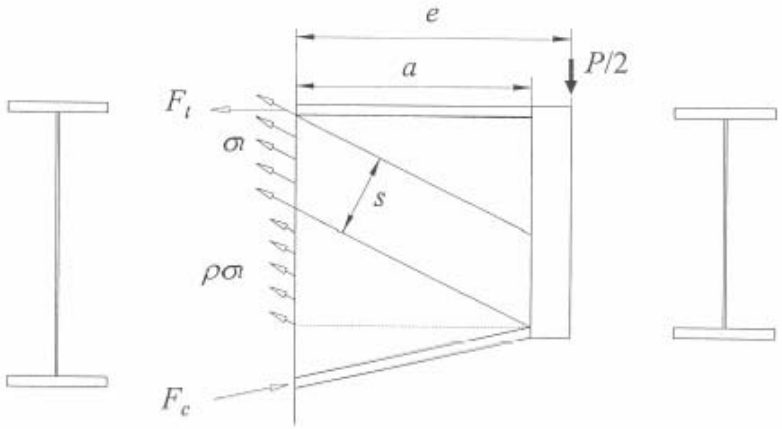


Figura 2.16 Modelo proposto por Takeda e Mikami [Zárate, 2002].

Perante este cenário, Zárate [Zárate, 2002] desenvolveu um modelo para o dimensionamento de vigas com perfil em I de alma esbelta e inércia variável (Figura 2.21). Baseando-se nos resultados obtidos por um modelo de elementos finitos e tomando como ponto de partida a teoria clássica de instabilidade, estabeleceu uma expressão analítica (Fórmula 2.2) que permite determinar a tensão crítica de enfunamento. Paralelamente desenvolveu um método de dimensionamento e/ou verificação da segurança de vigas com perfil em I de inércia variável, mediante a adaptação do método do campo diagonal de tracções a este tipo de elemento estrutural.

(Fórmula 2.2)

$$\tau cr = k\frac{\pi^2 E}{12(1-v^2)}\left(\frac{tw}{h_0}\right)^2$$

onde E é o módulo de elasticidade, v o racio de *Poisson*,

tw / h_0 o inverso da esbeltez da alma e k o coeficiente de enfunamento.

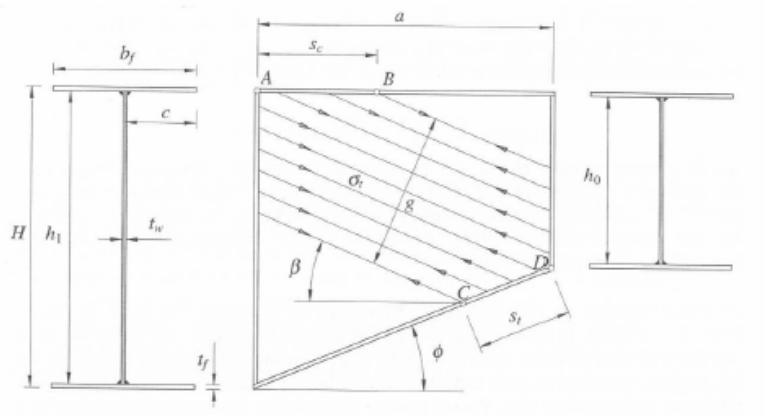


Figura 2.17 Modelo proposto por Zárate [Zárate, 2002].

Recentemente [Cruz e Guimarães, 2003], desenvolveram um trabalho com o objectivo de realizar uma análise numérica do efeito da redução de espessura da alma na estabilidade de vigas em I de inércia variável (Figura 2.22), baseando-se no trabalho de Zárate.

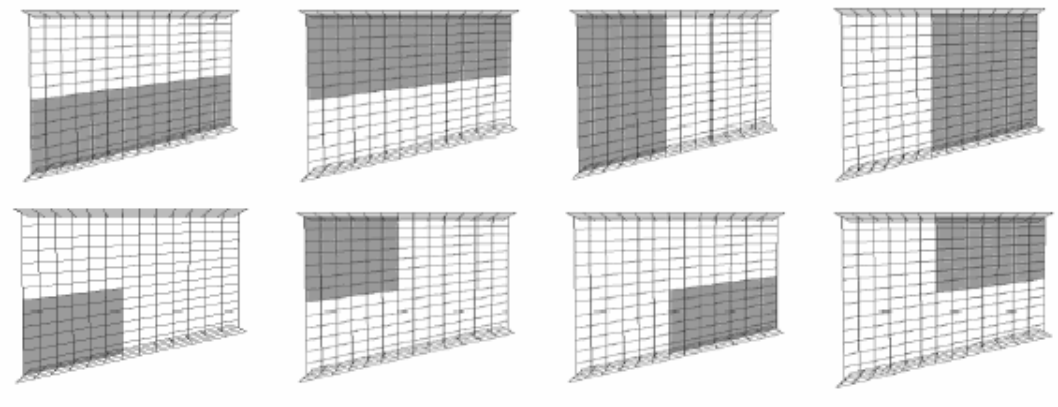


Figura 2.18 Casos de análise em [Cruz e Guimarães, 2003].

A avaliação do comportamento estrutural até à rotura das vigas em estudo, contemplou os fenómenos da não-linearidade geométrica (enfunamento dos painéis) e os que são consequência da plastificação do material em zonas localizadas do elemento (rótulas plásticas nos banzos e plastificação do painel de alma. As análises numéricas realizadas por Cruz e por Guimarães, permitiram extraír as seguintes conclusões quanto ao comportamento estrutural das vigas em I de alma esbelta e inércia variável:

- Os elementos estruturais nem sempre apresentam as características geométricas definidas em projecto. Desta forma a resistência efectiva das vigas podem ficar aquém da prevista na fase de projecto. Assim, por questões de segurança, é recomendável que sejam efectuados cálculos para prever a capacidade resistente dos elementos estruturais com as espessuras mínimas regulamentares;

- A diminuição da espessura em todo o painel que constitui a alma provoca a diminuição do valor da carga crítica de enfunamento e da carga máxima resistida pela viga. No entanto, o decréscimo do valor da carga crítica é bastante superior à diminuição do valor da carga máxima, o que se traduz em que a reserva de resistência seja maior em painéis de alma de maior esbelteza;

- O comportamento da viga não depende directamente da percentagem de área do painel da alma com espessura reduzida, mas sim da zona onde a referida redução se verifica. Uma redução da espessura do painel da alma no quadrante

esquerdo superior tem efeitos significativos na diminuição da capacidade resistente das vigas em I de alma esbelta e inércia variável (carga crítica, carga máxima e reserva de resistência);

- Se a redução da espessura se localizar no quadrante esquerdo inferior, a resposta da viga face à acção de esforços de corte é praticamente igual à da viga de referência;

- Quando a alma apresenta uma redução de espessura na zona onde a altura da viga é menor, então o deslocamento correspondente à carga máxima a que a viga resiste é superior ao da viga com espessura de alma constante.

Este modelo proposto por Cruz e Guimarães está mais adaptado à realidade, com a previsão da carga crítica de enfunamento elástico e da carga última. No entanto, o número de análises foi reduzido, e a divisão do painel em quatro zonas poderá dificultar a aplicação em casos práticos.

2.3 Conclusões

A análise das características intrínsecas dos Sistemas de Conhecimento e respectivas áreas de aplicação, abre perspectivas para o desenvolvimento do Sistema SCAE - Sistema de Conhecimento baseado em Data Mining para Análise da Estabilidade de Estruturas Metálicas -, para aplicação na área da Engenharia Civil. Trata-se de uma área em que a necessidade de utilização de Sistemas Inteligentes é premente para apoiar os especialistas na concepção e no planeamento da manutenção de estruturas, permitindo assim o prolongamento da vida útil e segurança das mesmas (e.g., Pontes, Viadutos).

Neste trabalho o foco de aplicação está direccionado para a análise da Carga e da Tensão Crítica de Estruturas Metálicas, devido ao elevado número de estruturas deste tipo existentes em Portugal que apresentam sinais de desgaste, necessitando de monitorização permanente que possa desencadear mecanismos de manutenção, e pela necessidade de

existência de um sistema que permita aos especialistas a simulação do comportamento de estruturas metálicas na fase de concepção/planeamento, uma vez que as fórmulas analíticas actualmente existentes ainda não se encontram devidamente calibradas devido ao elevado número de factores que influenciam o comportamento deste tipo de estruturas, e à dificuldade de realização de ensaios experimentais.

Capítulo 3

Descoberta de **C**onhecimento em **B**ases de **D**ados e **D**ata **M**ining

São apresentados os objectivos, conceitos, tipos de abordagem, áreas relacionadas, metodologias e especificações, para a Descoberta de Conhecimento em Bases de Dados.

3.1 Introdução

A banalização da utilização de tecnologias de Bases de Dados (BD) originou um crescimento exponencial dos dados armazenados, criando uma janela de oportunidades no que diz respeito à análise, sintetização e extracção de conhecimento a partir desses dados, uma vez que a quantidade de dados armazenados superou a capacidade humana de realização de tarefas de análise de dados sem recurso a ferramentas de computação.

O desenvolvimento de processos de análise de dados implica a utilização de aplicações com funcionalidades que permitam a automatização do processo de descoberta de conhecimento. Essa necessidade originou o surgimento de aplicações vocacionadas para a tarefa da Descoberta de Conhecimento em Bases de Dados e uma evolução dos Sistemas de Gestão de Bases de Dados (SGBD) existentes através da incorporação de técnicas de Aprendizagem Automática (e.g., MS SQL Server [Microsoft, 2004], Oracle [Oracle, 2004]).

3.2 Princípios

O termo Descoberta de Conhecimento em Bases de Dados (DCBD) surgiu para referenciar o processo de descoberta de conhecimento a partir de dados armazenados em BDs, processo que culmina na aplicação de técnicas de Data Mining (DM). Por vezes os termos DCBD e DM são usados de forma indiferenciada. Ao longo deste documento DCBD será usado para referenciar todo o processo de descoberta de conhecimento (e.g., estudo do domínio, processamento dos dados) enquanto DM referenciará a aplicação de algoritmos de identificação de padrões a partir dos dados.

Descobrir conhecimento significa extraír, de grandes conjuntos de dados, sem uma formulação prévia de hipóteses, informações genéricas, relevantes e previamente desconhecidas, que podem ser utilizadas para a tomada de decisões.

Assim DCBD pode ser definido como o ***processo*** *não trivial de identificação de* ***padrões*** *válidos e potencialmente úteis, perceptíveis a partir dos* ***dados*** [Fayyad et al., 1996].

O termo **processo** encontra-se associado à execução de diversos passos interativos (uma vez que requer a intervenção do utilizador sempre que é necessária a tomada de decisão), e iteractivos (uma vez que se podem verificar retrocessos a etapas anteriores), apresentados na Figura 3.1 e descritos na secção 3.3, assumindo-se como não trivial, uma vez que pode envolver a procura de estruturas, modelos, padrões ou parâmetros.

Os **dados** representam um conjunto de factos *F*, armazenados numa BD, na qual subconjuntos do mesmo são responsáveis pela caracterização de diversos padrões.

Um **padrão** pode ser caracterizado através de modelos, relações ou estruturas existentes nos dados, que devem ser perceptíveis, se não imediatamente, após determinado período de processamento, devendo ser válido, desconhecido e útil. Um padrão é uma expressão *E* numa linguagem *L* que descreve um subconjunto de factos *FE* do conjunto *F*. Por exemplo, em

relação aos dados sobre empréstimos bancários, o padrão *E1="Se Salário<T Então a pessoa faltou ao pagamento"* poderia ser um padrão para uma escolha apropriada de T.

A **novidade** de um padrão pode ser avaliada em relação às alterações verificadas ao nível dos dados ou do conhecimento e é representada por uma função *N(E,F)*, que pode ser do tipo lógico ou real. **A utilidade** representa o grau de utilidade de um padrão, isto é, até que ponto o padrão contribui para os objectivos inerentes ao processo, como por exemplo, o esperado aumento de lucro de um banco por aplicação da regra de decisão *E1*. A utilidade pode ser definida pela função *u=U(E,F)*.

Um dos objectivos da DCBD é gerar padrões que sejam compreendidos pelos humanos na perspectiva de contribuir para uma melhor compreensão dos dados. Assume-se que o grau de **interpretação** de um padrão é definido pela função *s=S(E,F)*.

O conhecimento descoberto pode também ser quantificado, seja *i=I(E,F,C,N,U,S)* o grau de interesse num dado padrão *E*, diz-se que o padrão *E* é conhecimento se para um dado valor de *i*, *I(E,F,C,N,U,S)>i*.

A aquisição de conhecimento nos seres humanos é feita através do processo de aprendizagem. Numa perspectiva sistémica, a aprendizagem pode ser definida como as alterações do sistema, que lhe permitem refazer as mesmas tarefas de uma forma mais eficaz e eficiente no futuro. Por outro lado, do ponto de vista da matemática, a aprendizagem pode ser vista como a percepção de conjuntos de dados [Adriaans e Zantinge, 1996]. A aprendizagem, torna-se viável ao basear-se na experiência e tendo em vista a compreensão do conjunto de dados.

O processo de DCBD depende de uma nova geração de ferramentas e técnicas de análise de dados, que envolve diversas etapas. A principal, que forma o núcleo do processo, e que muitas vezes se confunde com ele como referido anteriormente, chama-se Data Mining.

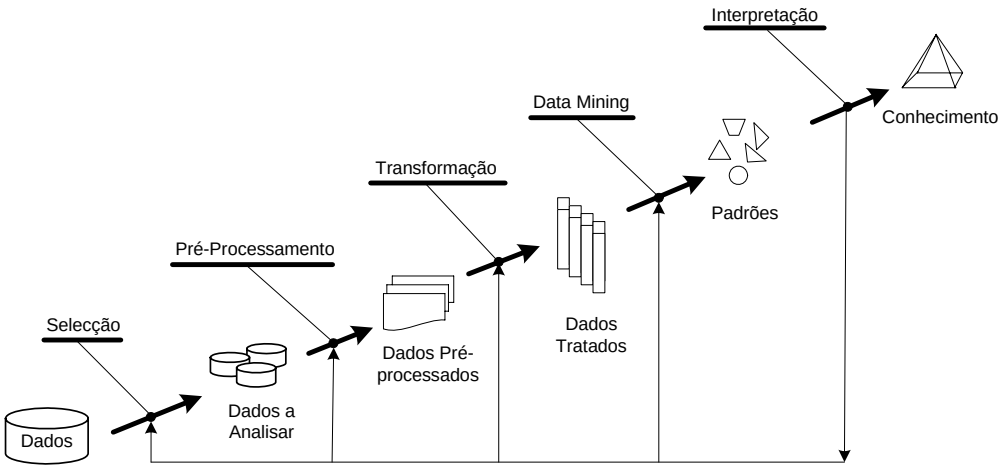


Figura 3.1 Fases do processo de DCBD (adaptado de [Fayyad et al., 1996]).

3.3 Fases do Processo de Descoberta de Conhecimento em Bases de Dados

As diferentes fases do processo de DCBD, apresentados na Figura 3.1, incluem:

- A selecção dos dados, que compreende o estudo/compreensão do domínio de aplicação e a selecção dos dados a analisar;
- O pré-processamento dos dados, onde são realizados procedimentos para a correcção de anomalias no conjunto de dados;
- A transformação dos dados, onde são realizadas transformações no conjunto de dados para que estes fiquem na forma correcta para aplicação de técnicas de Data Mining;
- O Data Mining, em que são aplicados os algoritmos ao conjunto de dados;
- A interpretação dos resultados alcançados, que engloba o estudo e a avaliação dos resultados alcançados na fase anterior de Data Mining.

Estudos anteriores demonstraram que a maior parte do esforço no desenrolar de processos de DCBD está concentrada nas fases de selecção, tratamento e pré-processamento dos dados, uma vez que continua bem válida na área da computação a máxima GIGO: "garbage in, garbage out" [Feelders, 2002]. A forma como são desenvolvidas, e tomadas as decisões nas fases iniciais têm um impacto significativo nos resultados obtidos.

A fase **Selecção dos Dados** compreende duas actividades: o estudo/compreensão do domínio de aplicação, e a selecção dos dados a analisar.

No estudo do domínio, são adquiridos conhecimentos sobre domínio de aplicação, através da apreensão de conceitos fundamentais e da definição clara dos objectivos do processo em curso. O conhecimento do domínio constitui um recurso essencial em qualquer processo de DCBD, sendo utilizado para conduzir o processo, podendo o conhecimento existente ser complementado com o conhecimento obtido no processo de descoberta. Por este facto torna-se necessária a presença, na equipa de desenvolvimento do processo de DCBD, de especialistas na área de aplicação. Estas equipas são por norma multi-disciplinares, integrando especialistas da área de aplicação, técnicos de bases de dados, especialistas em técnicas de data mining, etc…

Posteriormente, são seleccionados os dados armazenados nos diversos repositórios, desde sistemas de transacções, e.g., Data Warehouses, Data Marts, necessários para a geração de padrões pelos algoritmos de DM.

A selecção dos dados tem como principal objectivo limitar o espaço de pesquisa, direccionando o foco para subconjuntos de variáveis ou de dados, onde será realizada a descoberta de conhecimento.

Na fase de **Pré-Processamento e Transformação dos Dados**, são realizados procedimentos para a correcção de anomalias verificadas no conjunto de dados, e transformação dos dados para que estes fiquem na forma correcta para a aplicação dos algoritmos de aprendizagem na fase de DM, e.g., verificação de integridade dos dados, transformação da estrutura relacional – "desnormalização" da BD (Figura 3.2) –, fusão de bases de dados, redução do número de variáveis.

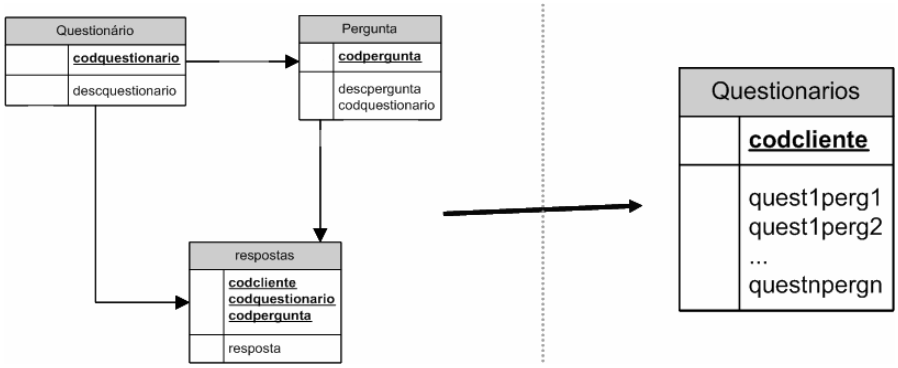


Figura 3.2 Exemplo de "desnormalização" de uma BD.

O exemplo apresentado na Figura 3.2 [Pinto e al, 2004], respeitante à "desnormalização" de uma BD, por forma a colocar os dados relativos a um cliente na mesma linha (e.g., dados pessoais e respostas a questionários), não é sempre necessária, uma vez que para determinados esquemas de BD e determinadas tarefas, as aplicações de Data Mining, incorporam funcionalidades que permitem o relacionamento de tabelas (e.g., nodo merge do SPSS Clementine Data Mining System).

As anomalias nos dados podem ser significativas quando se está na presença de dados relativos a bases de dados do "*mundo real*", em contraponto com aquelas construídas com base em experiências laboratoriais onde existe maior controlo e rigor dos dados registados.

No ponto 3.3.1 são identificados os problemas mais comuns e as técnicas mais usadas para a sua resolução.

Após a fase de Pré-Processamento e Transformação dos Dados, estes encontram-se na forma correcta para aplicação dos algoritmos de **Data Mining**.

Esta é a fase em que através da utilização de algoritmos de DM, os dados previamente seleccionados, pré-processados e transformados são utilizados pelo algoritmo de DM escolhido com base no objectivo do processo de DCBD: regressão, classificação, previsão, etc.

As tarefas e os algoritmos de DM são objecto de uma descrição mais detalhada na secção 3.3.2.

Após a aplicação dos algoritmos de DM procede-se ao estudo e avaliação dos resultados obtidos (i.e., **Interpretação dos Resultados**) pela utilização dos algoritmos de DM, na fase anterior. Os modelos gerados durante a fase de treino, são aplicados a novos conjuntos de dados de teste, de modo a avaliar o desempenho dos mesmos com dados desconhecidos, isto é, dados não utilizados durante a fase de geração do modelo.

Quando, ao longo do processo de DCBD, ocorrem falhas originadas por decisões que se revelam inapropriadas, estas vão reflectir-se na validade e utilidade dos modelos obtidos, não satisfazendo os objectivos delineados, ou retratando apenas o comportamento dos dados analisados, não podendo ser aplicados a dados desconhecidos por uma dificuldade de generalização. Quando se verificam estes casos, retrocede-se no processo, de modo a alterar as decisões tomadas, sendo o processo posteriormente retomado, permitindo identificar novos modelos que resultam das alterações efectuadas, verificando o grau de validade e utilidade dos mesmos e analisando o incremento ou decremento relativamente aos modelos anteriormente gerados.

Apesar dos algoritmos disponíveis possuirem critérios objectivos de avaliação da qualidade das regras, a introdução de medidas de interesse subjectivas tem como propósito limitar o conjunto de resultados a apresentar ao utilizador. A definição de medidas de interesse subjectivas, e que dependem de utilizador para utilizador, tendem a aumentar o grau de envolvimento do utilizador no processo de descoberta de conhecimento, tendo como contrapartida o aumento do interesse das diversas regras encontradas. Duas medidas de interesse subjectivas são o grau de surpresa, salientando que um padrão é interessante se ele é inesperado pelo utilizador, e a utilidade do padrão, sendo este interessante se o utilizador ou a organização poder usufruir do mesmo em sua vantagem.

Seguidamente são apresentadas de forma mais detalhada cada uma das fases do processo de DCBD, sendo enunciados alguns dos problemas mais comuns enfrentados em cada uma delas e técnicas utilizadas.

3.3.1 Pré-Processamento e Transformação dos Dados

No desenvolvimento do processo de DCBD, um dos factores críticos de sucesso é a qualidade dos dados. A forma como estes estão estruturados condiciona o sucesso e a prossecução de uma análise inteligente dos dados.

As técnicas de pré-processamento de dados têm por objectivo a melhoria da qualidade dos dados a utilizar, contribuindo para que a fase de Data Mining seja efectuada da forma mais eficiente e precisa.

Existem vários tipos de problemas, nomeadamente: tamanho e dimensionalidade das BD, volatilidade dos dados e do conhecimento (i.e., a alteração rápida dos dados pode tornar os padrões anteriormente encontrados inválidos), ruído e dados omissos, atributos relevantes não considerados. Os mais comuns correspondem a informação insuficiente e ainda a dados corrompidos caracterizados por possuírem ruído ou estarem incompletos. Na Tabela 3.1 encontra-se retratado um exemplo de dados com ruído, correspondente a respostas possíveis a uma pergunta de um questionário. Como pode verificar-se apesar da existência de uma opção para ausência de resposta ou resposta considerada inválida, cerca de 22% dos registos do atributo considerado estão em branco.

O crescimento registado no tamanho das BD, tem subjacente um efeito de aumento da probabilidade de existencia de dados omissos [Brown, 2003]. Quando se tratam de dados do "*mundo real*" os problemas são ainda maiores, causados por erros de transcrição e/ou por linguagem de descrição insuficiente.

Tabela 3.1 Exemplo de uma tabela de frequências do valor de um atributo de uma *BD*.

Valor	**Frequência**	**Percentagem**
Em branco	1952	21.8%
Não responde/resposta inválida	492	5.5%
Não	2052	22.9%
Sim	4464	49.8%
Total	*8960*	*100%*

A utilização de determinados algoritmos de DM, implica também alguns cuidados adicionais na fase de pré-processamento e transformação dos dados, uma vez que alguns métodos de aprendizagem admitem apenas atributos com valores simbólicos, enquanto outros lidam com valores numéricos, apresendo alguns uma maior sensibilidade a inconsistências nos dados (e.g., dados omissos ou desconhecidos, incoerências).

Valores Omissos ou Desconhecidos

Existem diversas técnicas para lidar com dados que contêm valores omissos ou desconhecidos com origens na estatística e matemática, e.g., Métodos Bayesianos [Gelman et al., 1995], Métodos de Imputação [Little, 1992], [Schafer, 1997], [Rubin, 1996].

Um dos métodos de imputação consiste na eliminação dos registos com dados omissos, no entanto, este método, pode conduzir a alterações significativas na representação dos dados relativamente ao universo em estudo e conduzir a amostras muito pequenas.

Existem métodos mais refinados de imputação, que incluem a substituição de casos pelo:

- valor mais comum do atributo[1];
- pelo seu valor médio[2] ou mediana[3];
- por um valor resultante da aplicação do método do vizinho mais próximo.

Quando existem atributos com valores omissos ou desconhecidos as técnicas mais comuns consistem em eliminar os dados da amostra ou substituir cada valor omisso pelo:

- Valor mais comum do atributo (moda);
- Valor médio ou mediana do atributo;
- Valor neutro.

Teoricamente, por exemplo num universo de 10k registos a eliminação de 1% dos casos, não irá afectar o desempenho dos algoritmos durante a fase de aprendizagem, no entanto, se o número de registos com atributos contendo valores omissos ou desconhecidos for de 10% torna-se necessária a utilização de uma técnica menos radical. A determinação da técnica adequada resulta da análise do caso em estudo.

As razões mais frequentes para a inconsistência nos dados são resultado de factores processuais, e quando se tratam de dados recolhidos a partir de questionários, os problemas podem também advir de recusa de resposta , e/ou opções de resposta inadequadas.

[1] A **moda** de um conjunto de números é o valor que ocorre com maior frequência: ou seja, é o valor mais comum. A moda pode não existir, e caso exista pode não ser única.
[2] Uma **média** é um valor típico, ou representativo, de um conjunto de dados. Como os valores representativos têm tendência a estar no centro do conjunto de dados, as médias são muitas vezes denominadas medidas de tendência central.
[3] A **mediana** de um conjunto de números ordenados relativamente à sua grandeza é o valor central (no caso de o número de observações ser ímpar) ou a média aritmética dos dois valores centrais (quando o número de observações é par).

Dados Dispersos

A construção de regras para classificação dos dados é condicionada pelos valores disponíveis na BD e a partir das quais as várias classes são definidas. A determinação do limite das classes, na construção de regras de classificação, está condicionada pelos valores verificados na *BD* para um dado atributo. A existência de muitos valores para o referido atributo (Tabela 3.2) conduz a que a tarefa de determinação das regras seja grandemente dificultada.

Em muitas situações verifica-se uma tendência na amostra, o que se traduz na dificuldade de aprendizagem por parte dos algoritmos de aprendizagem automática. Para ultrapassar esta dificuldade, as metodologias recomendadas são:

a) redução do número de classes através de agrupamento (Figura 3.3);

b) criação de um novo atributo resultado da aplicação de *log(x)*, em casos de atributos com dados do tipo numérico, em que *x* corresponde ao valor do atributo.

Tabela 3.2 Exemplo de frequências da distribuição inicial de um atributo de uma BD exemplo (*N.º de Filhos*)

Valor	**Frequência**	**Percentagem**
0	64	1.32%
1	1676	34.69%
2	2118	43.83%
3	448	9.27%
4	106	2.19%
5	44	0.91%
6	10	0.21%
7	2	0.04%
8	0	0.00%
9	2	0.04%
10	362	7.49%
Total	*4832*	*100%*

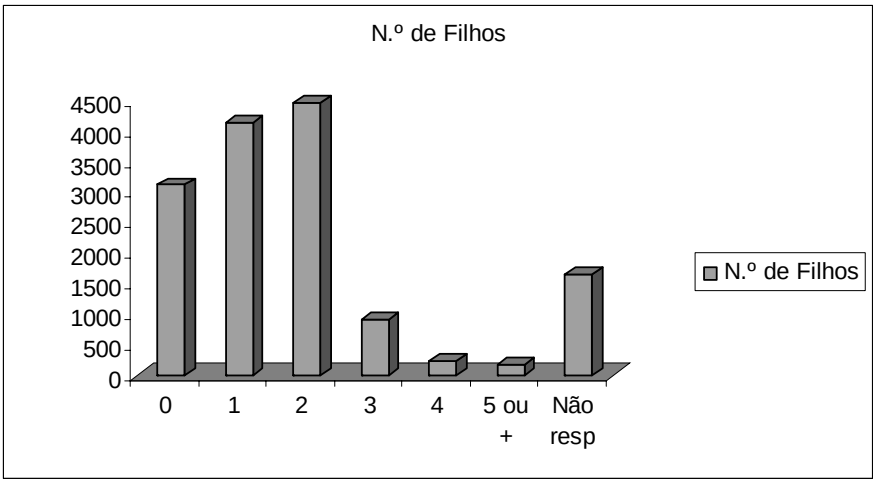


Figura 3.3 Gráfico de frequências de um atributo de uma BD exemplo (*N.º de Filhos*) após agrupamento.

Weiss e Provost [Weiss e Provost, 2001] realizaram um estudo sobre a influência da distribuição de classes na aprendizagem e de que forma isso afecta a performance do classificador gerado. Através da realização de testes comparativos constataram que tipicamente os classificadores gerados a partir de um conjunto de dados com uma desproporção evidente de classes apresentavam um pior desempenho na classificação da classe minoritária, em comparação com os classificadores gerados a partir do mesmo conjunto de dados, mas com uma proporção mais equilibrada de classes.

As razões para esta situação são duas. A primeira deriva do facto das regras geradas para a classe minoritária serem baseadas em menos exemplos e consequentemente mais sobre-ajustadas. O classificador tenderá a aprender limites mais rígidos do conceito. Este comportamento está relacionado com um problema já bem referenciado na área da aprendizagem: *small-disjuncts*. A segunda razão, tem a ver com o facto de dadas as características do domínio existirem mais exemplos de teste da classe minoritária. A classe mais frequentemente prevista será a maioritária, existindo uma maior probabilidade de classificar incorrectamente exemplos da classe minoritária. Existem dois métodos básicos para tornar a distribuição de classes mais equilibrada: *under-sampling* (criar uma amostra mais pequena do conjunto de exemplos da classe maioritária) e *over-sampling* (que consiste em gerar casos a

partir dos casos iniciais do conjunto, de forma a aumentar o número de casos da(s) classe(s) minoritária(s)). Estes dois métodos têm associadas desvantagens, como o desprezar de dados potencialmente úteis no primeiro método, o aumento do tamanho do conjunto de treino e portanto o tempo de computação, proporcionando um maior sobre-ajustamento aos dados no segundo método. Este balanceamento acontece apenas no conjunto de treino, devendo ser respeitada a distribuição original no conjunto de teste.

Escalonamento dos Dados

O objectivo deste procedimento é realizar uma transformação nos dados de modo a acelerar/melhorar o processo de aprendizagem. Os atributos considerados podem estar numa escala/domínio diferente, o que provoca problemas em métodos de aprendizagem, que podem dar demasiada importância a um atributo com um domínio alargado. Por exemplo, se o atributo a1 $\in$ *{0…10}* e o atributo a2 $\in$ *{2100…4010}*, o algoritmo de aprendizagem utilizado pode atribuir uma importância inadequada ao atributo a2 em consequência deste apresentar um domínio alargado de valor. O escalonamento depende do tipo de dados:

Entradas - o escalonamento das variáveis de entrada tem diversos efeitos conforme os algoritmos de aprendizagem considerados. De um modo particular, algoritmos de gradiente descendente (e.g., Back-Propagation), são bastante sensíveis ao escalonamento.

Saídas - existem duas fortes razões para escalonar as saídas. Em primeiro lugar, quando se usa mais do que uma saída e se a função de erro é sensível à escala, como acontece no caso da aprendizagem do gradiente descendente, então a diferença de escalas entre as saídas pode afectar a forma como, por exemplo, uma rede neuronal aprende. Se uma saída tem valores entre 0 e 1 enquanto outra tem valores entre 0 e 1000000, o algoritmo irá dispender a maior parte do esforço de aprendizagem na segunda saída. Assim, as saídas que têm a mesma importância devem ser transformadas para a mesma escala de valores.

Filtragem de Ruído

Em determinadas circunstâncias torna-se útil o uso de técnicas de filtragem para a eliminação de ruído, uma vez que este ocasiona problemas de dois tipos que se reflectem na construção de modelos a partir de amostras com ruído e na utilização destes na classificação de dados com ruído.

3.4 Data Mining

A fase de Data Mining consiste na procura de relacionamentos e padrões que se encontram ocultos no volume de dados armazenado. Estes relacionamentos representam conhecimento acerca do conjunto de dados explorado e das entidades nela contidas. Neste trabalho, a fase de DM engloba o processo de visualização de dados, com o objectivo de inferir automaticamente modelos e regras que representam conhecimento implícito acerca dos dados analisados, e o processo de geração de modelos de previsão.

O termo DM pode ser referenciado através de outras formas como extracção de conhecimento, arqueologia de dados, colheita de informações (*information harvesting*) e Data Dredging [Fayyad et al., 1996]. As definições encontradas para o termo DM variam de acordo com o autor, a abordagem e a área de especialização.

A principal diferença entre DM e outras ferramentas de análise de dados reside na forma como estas exploram as relações entre os dados. Enquanto nas diversas ferramentas de análise disponíveis, o utilizador constrói hipóteses sobre relações específicas e então corrobora-as ou refuta-as através das saídas da ferramenta utilizada, o processo de DM é o responsável pela geração de hipóteses, o que garante maior rapidez, aperfeiçoamento, autonomia e fiabilidade aos resultados. Contudo, o DM é uma etapa de um processo mais abrangente denominado de DCBD, sendo sustentado por três pilares fundamentais, dos quais depende o sucesso do projecto de DM [Berry e Linoff, 2001]: os modelos e as técnicas, os dados e a modelação de dados.

Os termos DM e AA tendem a ser confundidos uma vez que as técnicas de Data Mining têm a sua origem em métodos de diferentes áreas (e.g., estatística, reconhecimento de padrões, base de dados, inteligência artificial, sistemas distribuídos, visualização) [Fayyad e Stolorz, 1997], no entanto existem dois critérios que os diferenciam. DM diz respeito ao número e qualidade dos dados alvo e o segundo aos algoritmos usados. Os algoritmos aplicados em DM gozam da propriedade da escalabilidade: conhecidos os recursos do sistema (e.g., memória, velocidade de processamento) o tempo de execução do algoritmo deve crescer linearmente com o tamanho do conjunto de dados. Deve no entanto referir-se que grande parte dos algoritmos usados no domínio do DM tiveram a sua origem na AA e alguns deles sofreram alterações ao longo do tempo por forma a se tornarem escaláveis [Ramakrishnan, 1998].

Numa perspectiva histórica, o período de tempo que antecedeu o aparecimento do DM compreende três épocas revolucionárias, marcadas por novas abordagens de exploração dos dados dados armazenados em BD (Tabela 3.3).

Tabela 3.3 Épocas na exploração de dados.

ÉPOCA	TECNOLOGIAS	CARACTERÍSTICAS
Colecção de dados (década de 60)	Computadores, tapes, discos.	Resultado retrospectivo de dados estatísticos.
Acesso aos dados (década de 80)	Bases de Dados Relacionais (RDBMS), Structured Query Language (SQL), On-Line Database Connectivity (ODBC).	Resultado retrospectivo e dinâmico dos dados ao nível dos registos.
Data Warehousing e Suporte à Decisão (década de 90)	On-Line Analytic Processing (OLAP), Bases de Dados Multidimensionais, Data Warehouses.	Resultado retrospectivo e dinâmico dos dados em níveis múltiplos.
Data Mining (actualidade)	Algoritmos de Inteligência Artificial, multiprocessamento, Bases de Dados de grande dimensão.	Resultado em perspectiva da informação.

A área de DM é mutidisciplinar captando o interesse de investigadores e cruzando-se com áreas como: Aprendizagem Automática e Inteligência Artificial, Reconhecimento de Padrões, BD e Data Warehouse, Estatística e Matemática, Sistemas Periciais e Sistemas Baseados em Conhecimento. A abordagem protagonizada pelos algoritmos de DM permite encontrar descrições lógicas ou matemáticas, eventualmente de natureza complexa, de padrões num conjunto de dados [Decker e Focardi, 1995].

Nas últimas décadas foram desenvolvidos vários tipos de algoritmos de aprendizagem, que se distinguem na forma como traduzem a informação descoberta e no processo como é realizada essa descoberta, sendo alguns mais adequados a determinados tipos de problemas e de dados.

3.4.1 Objectivos de DM

Existem vários objectivos de Data Mining, apresentados na tabela 3.4. Normalmente enquadram-se nas categorias de classificação ou regressão. O objectivo de DM afecta a escolha dos algoritmos de aprendizagem a utilizar.

Tabela 3.4 Objectivos de Data Mining.

Objectivo
Previsão
Classificação
Regressão
Descrição
Associação ou Dependência
Sumariação
Segmentação
Visualização

Classificação

A *classificação* é um dos objectivos mais comuns de Data Mining, correspondendo à descoberta de uma função que associa um caso a uma classe dentro de diversas classes discretas de classificação, de forma a classificar um novo objecto de acordo com um padrão de classificação. A maior parte das técnicas utilizadas na classificação consideram conjuntos de treino com exemplos pré-classificados (i.e., aprendizagem supervisionada) com a finalidade de construir modelos adequados à descrição das classes, que posteriormente serão aplicados a dados não classificados.

Regressão

A *regressão*, muitas das vezes referida como previsão, consiste em prevêr valores futuros ou desconhecidos de uma variável dependente, a partir de exemplos. Por exemplo, nos casos de estudo deste trabalho, com base em dados experimentais, pretendeu-se criar modelos de previsão da carga crítica e da tensão crítica de elementos de Engenharia Civil (e.g., vigas de aço) recorrendo à utilização de *Redes Neuronais Artificiais*. A ideia subjacente corresponde à concepção de um modelo capaz de mimetizar uma função desconhecida que se aproxime da função dada por um conjunto de vectores etiquetados.

Segmentação

A *segmentação* permite identificar um conjunto finito de categorias ou segmentos para descrever os dados (e.g., identificação de grupos homogéneos de objectos em que cada grupo pertence a uma classe). Na situação ideal os objectos de cada grupo devem ter a menor distância entre eles, e a maior entre os objectos pertencentes a outros grupos. De uma forma geral, a segmentação é um objectivo intermédio de Data Mining, sendo realizado numa fase inicial para encontrar segmentos homogéneos de dados para posterior aplicação a cada um deles de algoritmos de aprendizagem para classificação ou previsão.

Associação ou dependência

Quando o objectivo é a *associação ou dependência*, pretende-se encontrar um modelo que descreva dependências significativas entre variáveis, através da identificação de grupos de dados fortemente associados. As associações surgem quando várias ocorrências estão ligadas num único evento, podendo surgir a nível estrutural (i.e., o modelo é representado de uma forma gráfica e com variáveis localmente dependentes em relação a outras) ou quantitativo (i.e., o modelo especifica o peso das dependências segundo uma escala numérica).

Sumariação

A *sumariação* utiliza métodos para encontrar uma descrição compacta para um subconjunto de dados. Os métodos de sumariação mais sofisticados derivam de regras de resumo e descobertas de relações funcionais entre variáveis. As técnicas de sumariação são sempre aplicadas à análise exploratória de dados e à geração automática de relatórios.

Visualização

A *visualização* trata da apresentação dos resultados (finais ou intermédios) de DM através de uma forma visual, geralmente através de gráficos. Pretende-se na visualização descrever informações complexas através de diagramas, o que permite uma melhor representação de padrões e tendências. Quanto melhor for a descrição de um conjunto de dados maior é a possibilidade de o entender e de compreender o domínio em que está inserido.

3.4.2 Metodologias e Especificações

Para o processo de DM estão actualmente disseminadas duas metodologias: a metodologia *CRISP-DM – C*ross-*I*ndustry *S*tandard *P*rocess for *D*ata *M*ining – (apresentada em

detalhe no Anexo A), e a metologia *SEMMA – S*ample, *E*xplore, *M*odify, *M*odel, *A*ssessement -, (apresentada em detalhe no Anexo B) que se encontram completamente especificadas e desenvolvidas. Estas metodologias foram desenvolvidas em ambientes diferentes, a primeira por um consórcio composto por organizações de diferentes sectores de actividade (e.g., indústria, serviços, fornecedores de tecnologia), e a segunda por uma organização fornecedora de soluções de suporte à decisão e *Business Intelligence*.

Tabela 3.5 Fases do ciclo de vida da metodologia CRISP-DM.

Fase	**Descrição**
Estudo do Negócio	Análise dos objectivos e dos requisitos funcionais, técnicos e temporais na perspectiva do negócio.
Estudo dos Dados	Recolha e análise dos dados.
Preparação dos Dados	Conjunto de actividades com a finalidade de construção do conjunto de dados que será usado na criação e validação do modelo na fase de DM.
Modelação	Selecção e aplicação de técnicas de DM de acordo com os objectivos definidos.
Avaliação	Revisão das actividades realizadas na construção do modelo e verificação da sua contribuição para o alcance dos objectivos de negócio.
Implementação	Actividades que conduzem à organização do conhecimento e à sua disponibilização.

Atendendo ao facto de em termos de processos para desenvolvimento de um projecto de Data Mining a metodologia *CRISP-DM* ser mais completa que a *SEMMA* - pela incorporação das fases de Estudo do Negócio, Estudo dos Dados e Implementação -, e se encontrar melhor documentada - focalizando todo o processo no estudo do negócio, i.e., orienta as suas etapas nos objectivos de negócio especificados, traduzindo-se numa forma segura e directa de resolução do

problema de Data Mining, ao apresentar uma visão mais ampla -, os casos práticos apresentados neste trabalho foram desenvolvidos segundo a metodologia CRISP-DM, cujas principais fases do ciclo de vida estão apresentadas na tabela 3.5.

3.4.2.1 Cross-Industry Standard Process for Data Mining (CRISP-DM)

Introdução

A metodologia *CRISP-DM*, foi desenvolvida no final da década de 90, motivada por um interesse crescente e generalizado pelo mercado *Data Mining* e pela necessidade de um processo padronizado, por um consórcio formado pelas seguintes organizações: NCR (Estados Unidos da América e Dinamarca), Daimler-Chrysler AH (Alemanha), SPSS Inc. (Estados Unidos da América) e OHRA (Grupo Bancário Holandês) [Chapman et al, 2000].

Os fundamentos da metodologia *CRISP-DM* estão assentes em sólidos princípios teóricos, e na experiência prática daqueles que desenvolver projectos de *Data Mining*. Desta forma foi incorporado conhecimento prático, de forma a responder aos requisitos dos utilizadores, não se centrando os princípios unicamente na tecnologia mas também nas necessidades dos utilizadores na resolução de problemas de negócio [Han & Kamber, 2001].

Fases do Processo

Esta metodologia é descrita como um modelo de processo hierárquico, iteractivo e interactivo, com um ciclo de vida (Figura 3.4) que se desenvolve ao longo de seis fases: *Estudo do Negócio*, *Estudo dos Dado*s, *Preparação dos Dados*, *Modelação*, *Avaliação e Implementação*. Cada tarefa genérica encontra-se subdividida em sub-tarefas especicializadas cuja execução depende do tipo de problema a resolver. A sequência de fases não é rígida, dependendo dos resultados alcançados e do desempenho das outras fases ou das tarefas

particulares de determinada fase [Chapman et al, 2000]. No final de cada uma das fases do ciclo de vida desta metodologia é produzida documentação sobre o processo.

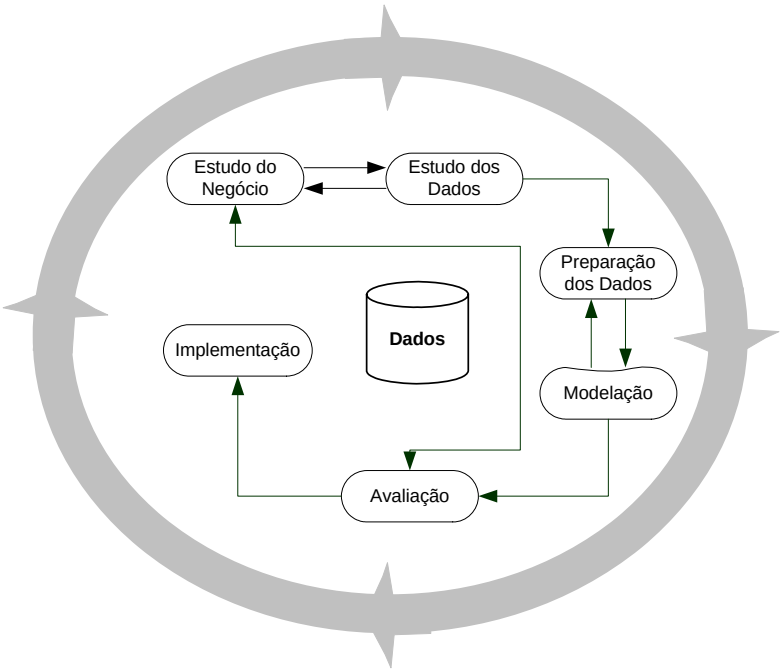


Figura 3.4: Ciclo de vida da metodologia CRISP-DM.

Estudo do Negócio

O *Estudo do Negócio* focaliza-se na análise dos objectivos dos projectos e nos requisitos funcionais, técnicos e temporais, na perspectiva do negócio. Este conhecimento é posteriormente utilizado na definição do problema de *Data Mining* e na definição do plano preliminar para alcance dos objectivos. O primeiro passo no processo de *Data Mining* consiste na compreensão da envolvente do problema a resolver, i.e., no estudo da necessidade de realização do projecto de *Data Mining*, na compreensão da perspectiva do problema, na

definição dos objectivos a atingir, e na descoberta de factores importantes que influenciam os resultados.

A fase de *Estudo do Negócio* compreende as seguintes tarefas:

1. *Determinação dos objectivos de negócio*: no início do projecto é fundamental determinar todos os detalhes acerca da situação do negócio, conhecer o histórico e antecedentes (e.g., descrever os objectivos primários do cliente segundo a perspectiva do negócio, identificar os critérios específicos ou gerais e subjectivos para avaliação do sucesso e utilidade dos resultados).
2. *Avaliação da situação actual*: inventariação dos recursos (e.g., humanos, dados, tecnológicos), listagem de todos os requisitos, pressupostos e restrições do projecto (i.e., programa de realização, compreensibilidade, qualidade dos resultados, segurança, aspectos legais e restrições na disponibilidade dos recursos), identificação dos riscos, ameaças ou eventos que possam comprometer o projecto e respectivos planos de contingência, elaboração de um glossário de termos relevantes para o projecto (e.g., terminologia do negócio e de *Data Mining*), elaboração de uma análise de custos/benefícios para o projecto.
3. *Definição dos objectivos de Data Mining*: descrição dos objectivos de *DM* e dos critérios para avaliação do processo (e.g., Classificação, Regressão/Previsão, Segmentação, Associação ou Dependência, Sumariação, Visualização).
4. *Concepção de um plano para o projecto*: elaboração de um plano para o projecto que deve incluir a definição da duração, dos recursos, das fases e sub-fases, das interacções entre os processos, entradas, saídas e dependências; e elaboração de um pressuposto inicial das tecnologias e técnicas (e.g., requisitos ao nível das tecnologias, *BD*, Sistemas Operativos) onde é seleccionada uma aplicação de *Data Mining* capaz de suportar os vários métodos a utilizar durante as diferentes fases do projecto.

No final desta fase, obtém-se como resultado um plano do projecto que inclui a informação, objectivos e critérios de sucesso do negócio, bem como os vários recursos, os requisitos, as restrições, os custos, os benefícios, os objectivos de Data Mining e os pressupostos relativamente a tecnologias e técnicas a utilizar.

Estudo dos Dados

Esta fase inicia-se com a recolha dos dados e prossegue com a sua análise de forma a identificar problemas de qualidade nos mesmos. Aos dados seleccionados é aplicada uma ou mais técnicas com o objectivo de descoberta de subconjuntos de dados para formular hipóteses para a informação escondida.

Tal como mencionado anteriormente, antes da aplicação das técnicas de *DM*, é necessário ter em conta algumas tarefas de pré-processamento e transformação dos dados que se traduzem no seguinte:

1. *Recolha inicial dos dados*: consiste na aquisição e compreensão dos dados. Esta tarefa tem como resultado uma lista dos dados adquiridos, a sua localização, métodos de aquisição, problemas encontrados e soluções para os mesmos.
2. *Descrição dos dados*: após a tarefa de recolha dos dados é necessário efectuar a sua descrição, verificar o formato em que se encontram, o número de registos disponíveis, e outras características identificadas.
3. *Exploração dos dados*: consiste na elaboração de uma lista inicial de hipóteses e o seu impacto no projecto. Para a obtenção de melhores resultados nesta fase utilizam-se, por exemplo,gráficos que permitam a identificação de características dos dados.
4. *Verificação da qualidade dos dados*: elaboração de uma listagem que inclui os problemas de qualidade e possíveis soluções a implementar, para correcção dos problemas identificados nos dados.

Preparação dos dados

A *Preparação dos Dados* consiste num conjunto de actividades com a finalidade de construção do conjunto de dados que será utilizado na criação e validação do modelo na fase de DM, sofrendo contudo várias optimizações. Inclui a selecção de tabelas, registos e atributos, bem como a transformação e limpeza dos dados a utilizar na fase seguinte de DM, envolvendo a seguintes actividades:

1. *Selecção dos dados*: escolha dos dados a utilizar na análise, seguindo critérios que incluem a relevância dos objectivos de *Data Mining* e restrições técnicas e de qualidade, assim como os limites no volume e tipo de dados, e listagem dos dados incluídos e excluídos e as razões das decisões tomadas.
2. *Limpeza dos dados*: esta actividade complementa a anterior, consistindo na aplicação de técnicas de optimização da qualidade dos dados.
3. *Construção de dados*: esta tarefa compreende a derivação de novos atributos, a criação de novos registos e a transformação dos dados.
4. *Integração dos dados*: utilização de métodos para criação de novos registos ou valores, cuja informação é uma combinação de múltiplas tabelas ou registos.
5. *Formatação de dados*: esta actividade envolve modificações sintácticas dos dados que não modifiquem o significado, mas que são um requisito da tecnologia de modelação.

Modelação

Nesta fase de modelação são seleccionadas e aplicadas as técnicas de *Data Mining* mais apropriadas, de acordo com os objectivos definidos. São seleccionadas várias técnicas de modelação (e.g., Árvores de Decisão, Redes Neuronais Artificiais, Algoritmos Genéticos) e os seus parâmetros são ajustados de forma a optimizar os resultados. De uma forma geral, existem várias técnicas para o mesmo tipo de problema, sendo que algumas têm requisitos específicos

para a forma como os dados são apresentados, implicando por vezes, um retrocesso à fase de preparação dos dados.

No início do processo são especificados os problemas e os objectivos do *Data Mining*. No entanto, é nesta fase que os dados são submetidos para a modelação, devendo seleccionar-se as técnicas que melhor se adequam aos objectivos e proceder à modelação.

A fase de modelação inclui as seguintes tarefas:

1. *Selecção das técnicas de modelação*: na escolha da técnica mais apropriada deve ter-se em consideração os seguintes aspectos: tipo de problema, tecnologias e objectivos do *Data Mining*.
2. *Criação de uma concepção de teste*: antes da construção do modelo é necessário definir um procedimento ou mecanismo para testar o desempenho do modelo.
3. *Construção do modelo*: escolhida a tecnologia de modelação, esta é aplicada ao conjunto de dados preparados anteriormente, por forma a criar um ou mais modelos. Os vários parâmetros devem ser ajustados, e os modelos resultantes devem ser convenientemente interpretados e o seu desempenho explicado. A criação do modelo representa a fase central do *Data Mining*.
4. *Revisão do modelo*: os modelos gerados devem ser interpretados de acordo com o domínio de conhecimento, critérios de sucesso do projecto de *Data Mining* e com o procedimento de teste definido. Deve ser avaliado o sucesso da aplicação do modelo e discutidos os resultados do processo de *Data Mining*.

Avaliação

Esta fase compreende as tarefas de avaliação do(s) modelo(s), revisão das actividades realizadas na sua construção e verificação da sua contribuição para o alcance dos objectivos de negócio. Nesta fase de avaliação são desenvolvidas as seguintes actividades:

1. *Avaliação dos resultados*: é efectuada uma análise aos resultados, avaliando-se se o modelo atingiu os objectivos de negócio e de *Data Mining*, procurando determinar se o modelo apresenta alguma deficiência.
2. *Revisão do processo*: integra a revisão de todas as fases de forma a realçar actividades que não foram realizadas e/ou que necessitam de ser repetidas. Esta actividade, tem como objectivo a validação de forma a determinar a existência de factores importantes que tenham sido omitidos.
3. *Determinação dos próximos passos*: quando os passos anteriores apresentam resultados considerados satisfatórios, o projecto deve ser concluído, devendo inciar-se a fase de implementação. Caso se verifique a situação contrária, é necessário proceder a uma nova iteração, percorrendo as fases do ciclo de vida da metodologia. Avalia-se, a necessidade de eventuais correcções no processo através do retrocesso a fases anteriores, ou inclusivé, o reinício do processo de *Data Mining*.

Implementação

Na *Implementação*, são empreendidas actividades que conduzem à organização do conhecimento e à sua disponibilização, para que possa ser utilizado no negócio. Por exemplo, se num projecto de *Database Marketing* for desenvolvido um modelo do tipo *Árvore de Decisão*, de onde seja possível extraír um conjunto de regras explicativas do comportamento dos consumidores considerados, essas regras devem ser incorporadas nos processos de marketing, uma vez que podem conduzir a estratégias com um grau de sucesso superior ao conseguido anteriormente [Santos et al, 2004].

O grau de dificuldade desta fase está relacionado com o output: geração de um relatório, ou implementação de todo o processo de *Data Mining*.

Na maioria dos casos é o cliente, e não o analista, que executa as actividades de implementação. Todavia, é importante que as acções a serem executadas para uso dos modelos

criados sejam devidamente definidas e compreendidas. As actividades envolvidas nesta fase são as seguintes:

1. *Planeamento da avaliação dos resultados*: definição da estratégia a seguir na implementação dos resultados do processo de *Data Mining*, que deve incluir os passos a executar e a forma como devem ser executados.
2. *Planeamento da monitorização e manutenção*: quando os resultados do processo de *Data Mining*, i.e., os modelos, forem implementados no domínio do problema como parte da rotina do dia-a-dia, é aconselhável uma estratégia de monitorização e manutenção. Os resultados da monitorização e manutenção podem indicar se os modelos são usados de forma correcta.
3. *Produção do relatório final*: é a actividade que conclui o projecto de *Data Mining*, e que consiste na elaboração de um relatório final onde devem ser apresentados, de forma resumida, os pontos mais importantes, a experiência adquirida, e a explicação dos resultados alcançados sobretudo daqueles considerados de maior importância.
4. *Revisão do projecto*: avaliação dos pontos que foram efectuados de forma correcta, do que correu bem e do que necessita de ser melhorado. Compreende a revisão das experiências mais importantes do projecto, assumindo relevância pelo facto de apresentar contributos para projectos futuros e situações similares, com apresentação dos constrangimentos, aproximações erradas, selecção das técnicas de *Data Mining*.

3.4.2.2 SEMMA – Sample, Explore, Modify, Model

Introdução

A metodologia *SEMMA* foi desenvolvida pelo Instituo SAS (SAS Institute Inc.), organização cuja missão é o desenvolvimento de soluções para as áreas de Estatística, Análise de Dados, Business Intelligence, Data Mining e Suporte à Decisão [SAS, 2004].

O Instituto SAS define DM como o processo de extraír informação valiosa e relações complexas de grandes volumes de dados. Com base neste conceito genérico, o processo de *Data Mining* foi dividido em 5 etapas, que compõe o acrónimo *SEMMA*: *S*ample – Amostragem; *E*xplore – Exploração; *M*odify – Modificação; *M*odel – Modelação; e *A*ssessment – Avaliação [Groth, 2000]. Estas etapas, distintas, correspondem a um ciclo, em que as tarefas internas podem ser executadas de forma repetida sempre que se verifique necessário (Figura 3.5).

Mais do que uma metodologia de *Data Mining*, é considerada um auxiliar na condução de um projecto em todas as suas etapas, desde a especificação do problema até à sua implementação, disponibilizando uma estrutura para a concepção, criação e evolução dos projectos de Data Mining, de forma a apresentar soluções para os problemas do negócio.

Etapas do Processo

A metodologia SEMMA pode ser descrita como um processo composto por 5 etapas, que se inicia com a selecção da amostra (i.e. conjunto de dados) e que termina com a avaliação do desempenho do modelo gerado na fase de modelação no conjunto de teste da amostra.

Amostragem

Nesta etapa é seleccionada uma amostra representativa do universo em estudo, que deve corresponder a um subconjunto de dados que pertencem ao universo em que cada elemento tem

as mesmas hipóteses de ser incluído, devendo ter uma dimensão ajustada, de forma a optimizar: custos, rentabilidade e desempenho da metodologia (e.g., rapidez de processamento).

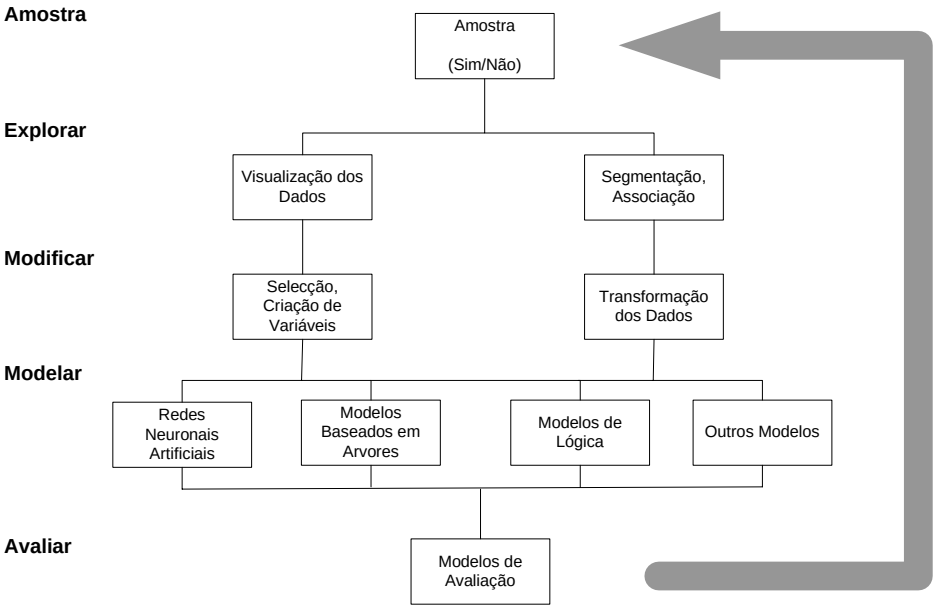


Figura 3.5 Etapas da metodologia SEMMA.

Exploração

A etapa é marcada pela procura de tendências imprevistas e por anomalias nos dados, com o objectivo de obter um conhecimento aprofundado sobre os mesmos e sobre relações implícitas.

O conjunto de dados seleccionados é nesta etapa explorado através da utilização de técnicas de análise visuais ou numéricas, verificando-se tendências e/ou agrupamentos inerentes nos dados. Quando as técnicas estatísticas mais simples não revelem tendências claras, poderá recorrer-se à utilização de técnicas mais refinadas (e.g., distribuição de Poisson, Minímos Quadrados, Qui-Quadrado, Regressão Linear).

Modificação

A etapa de modificação envolve todos os processos que visam realizar as transformações necessárias aos dados, identificadas na etapa de Exploração. As transformações podem ser de inclusão de novos atributos através do agrupamento de subgrupos significativos identificados nos dados, de selecção ou introdução de novos atributos, para que o conjunto de dados a utilizar na etapa de Modelação inclua os atributos mais significativos.

Modelação

Após as etapas iniciais de exploração e preparação do conjunto de dados, nesta etapa são definidas e aplicadas as técnicas de Data Mining (e.g., técnicas de AA, modelos estatísticos) adequadas aos objectivos definidos, de forma a obter os resultados desejados. Na selecção da técnica deve ter-se em conta que de uma forma geral existem várias técnicas para o mesmo tipo de problema, sendo que cada uma tem propriedades e características singulares, bem como requisitos específicos para a forma como os dados são apresentados.

Avaliação

Na metodologia SEMMA, a última etapa é a de Avaliação do modelo gerado de forma a aferir o seu desempenho. Normalmente corresponde à apresentação de um conjunto de dados (i.e., conjunto de teste) ao modelo, avaliando a resposta do mesmo perante casos que não foram utilizados na fase de treino.

Capítulo 4

Modelos e **T**écnicas de **D**ata **M**ining

São apresentados os principais modelos e técnicas utilizados no processo de Data Mining, evidenciando as propriedades associadas, com particular ênfase naqueles que foram usados na aquisição de conhecimento para o sistema SCAE.

4.1 Introdução

Um modelo é definido como uma função (mapa) que atribui a cada exemplo possível no domínio definido pelos atributos de entrada, um valor contido no domínio do atributo de saída, contendo cada modelo um conjunto de parâmetros que têm de ser ajustados (ou estimados) a partir de um conjunto de dados, através de um algoritmo, na fase de aprendizagem. Após a aprendizagem, é possível extrapolar novas saídas, alimentando o modelo com novas entradas (utilização de um modelo) [Cortez, 2004].

Como técnica compreende-se o conjunto de processos baseados em conhecimentos científicos, de cálculo ou experimentação, utilizados para a obtenção de um resultado.

Na construção de um modelo definem-se as principais características do sistema, que devem representar o mais fielmente possível a realidade, recolhem-se os dados necessários para a construção do modelo e para a consequente validação, sendo necessária uma divisão do conjunto de dados em dois subconjuntos, um para geração do modelo, chamado de conjunto de treino, e outro para validação do modelo, chamado de conjunto de teste. Regra geral, o maior número de exemplos da amostra é colocado no conjunto de treino, contudo em proporções variáveis dependente de vários factores (e.g., natureza do problema, número de casos da amostra, técnica a utilizar). Ao modelo são aplicados algoritmos para a identificação de padrões e relacionamentos.

Deve realçar-se que não existe um modelo universal de Data Mining que resolva, de forma eficiente, todos os problemas [Harrison, 1998]. A escolha de um determinado algoritmo é de certa forma uma arte [Fayyad et al, 1996], uma vez que existem diferentes modelos para as mesmas tarefas de DM com vantagens e desvantagens íntrinsecas. Na tabela 4.1 é apresentado um mapeamento entre as diversas tarefas e as técnicas de DM.

Tabela 4.1 Tarefas e Técnicas de Data Mining.

Técnicas / Tarefas	*Classificação*	*Segmentação*	*Visualização*	*Sumariação*	*Associação*	*Previsão*
Árvores de Decisão	√	√	√	√		√
Indução de Regras	√	√			√	√
Redes Neuronais	√	√			√	√
Algoritmos Genéticos	√	√		√		√
Aprox- Vizinhanças		√		√		

4.2 Árvores de Decisão

As Árvores de Decisão tiveram a sua origem na área da Aprendizagem Automática, com base na análise designada por Automatic Interaction Detection percursionada pela Universidade de Michigan. Esta análise testa automaticamente todos os valores de um determinado dado de forma a identificar aqueles que têm uma forte associação com os registos de saída seleccionados para o teste. Os valores com fortes associações são os de previsão chave ou factores explicativos, chamados de regras dos dados. Outro algoritmo pioneiro foi o CHAID=(Chi-Quadrado + Automatic Interaction Detection), desenvolvido com base nas capacidades da análise Automatic Interaction Detection com a adição da fórmula estatística do Chi-Quadrado [Lamb, 1997].

O teste Chi-Quadrado permite verificar a semelhança entre categorias discretas e mutualmente exclusivas (e.g., diferenças de comportamento entre homens e mulheres). Cada item deve pertencer a uma e somente a uma categoria.

No entanto o *"pai"* das Árvores de Decisão é Ross Quinlan, da Universidade de Sidney na Austrália. A ele se deve o desenvolvimento da tecnologia que permitiu o seu aparecimento, através do algoritmo ID3 (Algoritmo 4.1) em 1983. Desde essa data têm sido introduzidas novas funcionalidades (Tabela 4.2) que resultaram no surgimento de evoluções do ID3 (e.g., algoritmos C4.5 e C5.0).

Formalmente uma árvore de decisão consiste numa estrutura arboriscente em que cada nó define uma condição lógica sobre um atributo duma instância. Denominando um conjunto de instâncias por S e o conjunto de atributos considerado por $A = \{a_1,...,a_m\}$, então para $x \in S$ tem-se $x \equiv \langle a_1(x),...,a_2(x) \rangle$, sendo $a_i(x)$ o valor assumido pelo atributo a_i na instância x. Assim o nó duma árvore contém uma condição sobre algum elemento de *A*, por exemplo $a_k>3,7$ *o*u $a_k=alto$. Cada ramo derivado dum nó consiste num possível valor do atributo considerado no nó. Cada folha da árvore representa um elemento duma classe[1].

[1] Na aprendizagem supervisionada, cada instância possui um atributo especial denominado classe, que descreve o fenómeno de interesse. Em casos de classificação as classes pertencem a um conjunto discreto nominal de valores, enquanto que nos casos de regressão pertencem a um conjunto de valores reais.

Cada caminho, desde a raiz até uma folha corresponde a uma regra de decisão ou classificação. Uma árvore de decisão é traduzível numa disjunção de conjunções lógicas de condições sobre os valores de A, sendo cada ramo da árvore uma conjunção de condições e o conjunto dos ramos disjuntos.

As Árvores de Decisão, essencialmente utilizadas em problemas de classificação, são uma forma de representação de um conjunto de regras que seguem uma hierarquia de classes ou valores, expressando uma lógica simples condicional. Graficamente, são semelhantes a uma árvore, consistindo numa estrutura que interliga um conjunto de nós através de ramos resultantes de uma partição recursiva dos dados, desde o nó raiz até aos nós terminais (folhas), que fornecem a classificação para a instância (Figura 4.1).

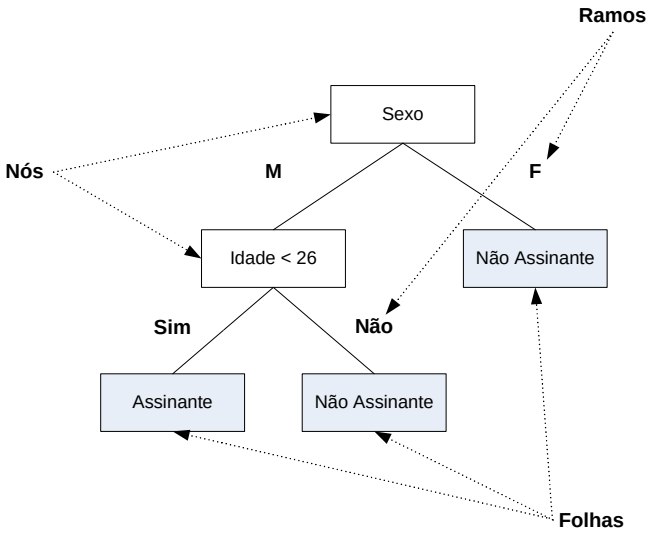


Figura 4.1 Árvore de Decisão para um problema de classificação de indivíduos como assinantes de uma revista de automóveis.

Existem dois tipos de Árvores de Decisão [Berry Linoff, 2000][Quinlan, 1996]:

- *Árvores de Classificação;*
- *Árvores de Regressão* (Figura 3.5).

As primeiras têm por objectivo qualificar os registos e associá-los com determinada classe e garantir que essa classificação esteja correcta. As árvores de regressão realizam a estimativa do valor de uma determinada variável (e.g., calcular o valor da carga crítica de uma viga de aço).

Estes dois tipos de árvores apresentam a mesma estrutura. Quando os dados são apresentados ao modelo, cada registo percorre um determinado caminho para uma série de testes, até que o registo alcança uma folha ou nó puro da árvore sendo-lhe atribuído uma classe. Com base no conjunto de treino são classificados os registos numa classe, ou no caso das árvores de regressão, um valor baseado numa função matemática dos valores que alcançaram essa folha no conjunto de treino [Berry & Linoff, 2000].

Na Figura 4.2 é apresentado um exemplo de uma árvore de regressão para a previsão do valor médias de casas, sendo os nós da árvore compostos por: número médio de quartos (*RM*), percentagem da população de baixo estrato social (*LSTAT*), distância pesada aos grandes centros de emprego (*DIS*), concentração de óxido de nitrogénio (*NOX*) e o valor da aplicação da fórmula $100*(b-0.632)^2$ em que *b* é a percentagem da população negra (*B*).

A partir de uma árvore de decisão é possível extraír um conjunto de regras representativas do modelo através da técnica de Indução de Regras (secção 4.3), como demonstrado pelas regras associadas ao modelo apresentado na Figura 4.1:

```
Regra #1
If sexo='M'
and idade>=26
then Não Assinante.
```

Regra #2
If sexo='F'
then ***Não Assinante.***

Regra #3
If sexo='M'
and idade<26
then ***Assinante.***

Os algoritmos de indução de Árvores de Decisão, constroem os padrões a partir dos dados de treino, de uma forma recursiva subdivindo o conjunto de dados até que este seja apenas composto por nós "puros", i.e., sempre que possível que cada nó represente apenas uma única classe, ou satisfaça um critério. A estrutura das árvores geradas é composta por ***folhas*** (nós puros) que correspondem às classes/objectos, ***nós internos***, que correspondem aos atributos (especifica algum teste efectuado num único atributo, com duas ou mais sub-árvores que representam saídas possíveis) e ***ramos***, que correspondem aos valores dos atributos.

O algoritmo de indução de árvores de decisão é definido por:

seja T a arvore de decisão a induzir
seja S o conjunto de exemplos para a aprendizagem
se todos os exemplos de S pertencem à mesma classe C
    então a árvore tem um só nodo C
senão seleccionar o atributo A mais "informativo" cujos valores são $v_1,\ldots,v_n$
particionar S em n subconjuntos $S_1,\ldots,S_n$ um para cada valor v_i de A
construir (recursivamente) sub-árvores $T_1,\ldots,T_n$ para cada $S_1,\ldots,S_n$

Algoritmo 4.1 Indução de Árvores de Decisão (ID3).

Esta técnica funciona através da criação e treino de subconjuntos de informação para os quais é inferida uma ou mais regras. Tendo como ponto de partida uma grelha de dados, com inúmeras colunas e linhas, de acordo com a escolha de saída mostra o factor mais correlacionado com este objecto, ou seja, o primeiro nó da Árvore de Decisão. Os restantes factores são subsequentemente classificados como nós e relacionados com os nós anteriores, possibilitando uma visualização fácil e rápida do factor com um factor de relacionamento mais elevado com o objecto de saída.

Na construção de uma árvore de decisão deparam-se dois problemas: **(i)** que atributo seleccionar para teste num nó, e **(ii)** quando parar a divisão dos exemplos.

Existem várias medidas para avaliar a capacidade de um dado atributo para discriminar as classes (i.e., atributo mais informativo), no entanto todas convergem em dois pontos: uma divisão que mantém as proporções de classes em todas as partições é inútil, e uma divisão onde em cada partição todos os exemplos são da mesma classe tem utilidade máxima.

As medidas de partição dividem-se em três tipos:

1. Medida da diferença dada por uma função baseada nas proporções das classes entre o nó corrente e os nós descendentes, valorizando a pureza das partições, e.g., índice de gini, entropia;
2. Medida da diferença dada por uma função baseada nas proporções das classes entre os nós descendentes, que valoriza a disparidade entre as partições;
3. Medida de independência, que mede o grau de associação entre os atributos e a classe.

A construção de uma árvore de decisão é guiada pelo objectivo de diminuir a entropia. A entropia é uma medida da aleatoriedade, i.e., a dificuldade de previsão do valor correcto para a variável objectivo, sendo uma medida aplicável à partição de um espaço de probabilidade.

Se tivermos um conjunto de várias instâncias *S*, e um conjunto de n classes $C=\{C_1,...,C_n\}$, sendo p_i a probabilidade da classe C_i em *S*, então a entropia conjunto S, é uma medida de homogeneidade deste:

(Fórmula 4.1)

$$Entropia(S) = \sum_i p_i * log_2 p_i$$

Considerando um atributo *A* das instâncias de *S*, com $A \in \{v_1,v_2,...,v_r\}=V$, então a consideração de $A=v$, com $v \in V$ separa um subconjunto de elementos de *S*. Denominando esse subconjunto de elementos por S_v, então podemos recalcular a entropia deste novo conjunto: *Entropia* (S_v). Realizando esta operação para cada elemento de *V*, podemos determinar o quanto é esperado que seja reduzida a entropia, considerando que os valores de A são conhecidos. O ganho de informação do atributo A, é dado pela Fórmula 4.2.

(Fórmula 4.2)

$$ganho(S,A) = Entropia(S) - \sum \frac{|S_v|}{|S|} Entropia(S_v)$$

em que $|S|$ e $|S_v|$ designam a cardinalidade dos conjuntos *S* e *Sv*, respectivamente.

Dado um conjunto de exemplos, que atributo usar para teste?

1. Os valores de um atributo definem partições do conjunto de exemplos;
2. O ganho de informação mede a redução da entropia causada pela partição dos exemplos de acordo com os valores do atributo.

Quanto à paragem da divisão dos exemplos deve acontecer quando:

1. Todos os exemplos pertencem à mesma classe, ou;
2. Todos os exemplos têm os mesmos valores dos atributos, mas classes diferentes, ou;
3. O número de exemplos é inferior a um certo limite, ou;
4. O mérito de todos os possíveis testes de partição dos exemplos é muito baixo.

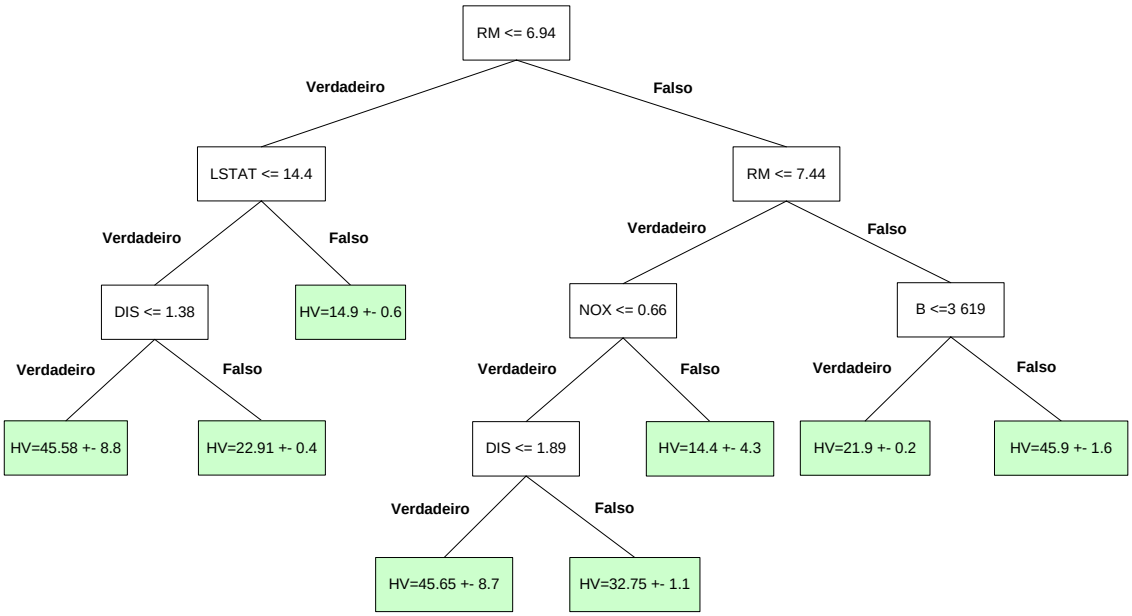


Figura 4.2 *Árvore de Regressão* relativa a um exemplo de *previsão do valor médio das casas (HV)* (adaptado de [Torgo, 1998]).

Uma das limitações inerentes ao algoritmo ID3 consiste na possibilidade de gerar árvores demasiado ajustadas aos dados de treino (i.e., *overfitting*) com um desempenho quase perfeito neste conjunto de dados, mas com um baixo desempenho nos novos dados de teste. O algoritmo C4.5 [Quinlan, 1993] é um método melhorado do ID3, que permite ultrapassar o obstáculo do sobreajustamento, através da utilização da poda da árvore (i.e., podar significa

reduzir algumas subárvores a folhas através de um teste estatístico que tem em conta os erros num nó e a soma dos erros nos nós que descendem desse nó). A poda de uma árvore pode ter ainda o objectivo de diminuição do tamanho das árvores, existindo duas formas de poda; pré-pruning e post-pruning. O pre-pruning envolve decidir durante o processo de crescimento da árvore quando é que este deve parar, enquanto no post-pruning a árvore é concebida em duas fases: (i) construção da árvore e (ii) eliminação dos testes demasiado específicos ao conjunto de treino e cuja existência não se traduz num acréscimo de pureza do modelo, como por exemplo um decréscimo da estimativa do erro do modelo. A estratégia de post-pruning permite obter, regra geral, melhores resultados

Além do sobreajustamento o algoritmo C4.5 permite ultrapassar problemas concretos e comuns do mundo real (e.g., atributos com valores numéricos, valores omissos, dados com ruído), disponibilizando a possibilidade de realizar uma validação cruzada, incrementando assim a qualidade da estimativa do erro cometido pelo classificador.

Tabela 4.2 Evolução do algoritmo ID3.

ALGORITMO	CARACTERÍSTICAS
ID3	Variáveis discretas; Critério da Entropia; Variável mais informativa em cada nodo.
C4.5	Possibilitou: . uso de valores contínuos; . utilização de dados com valores omissos; . poda Árvores de Decisão; . derivação de Regras.
C5.0	Boosting

O algoritmo C5 [Quinlan, 1997] é o sucessor mais recente do C4.5, melhorado para lidar com as exigências do mundo real, através do incremento da eficiência ao nível do tempo de processamento e de memória utilizada. Uma das características mais importantes do C5 é a

utilização da técnica de *Boosting* [Schapire, 2002], que consiste em gerar vários classificadores a partir do mesmo conjunto de treino e depois combiná-los num único classificador final no qual cada classificador inicial participa votando com um certo peso. Este peso é ajustado durante o processo de treino [Quinlan, 1996]. Segundo Quinlan [Quinlan, 1997], em determinados casos a redução dos erros de classificação pode atingir os 40%.

Na maioria das situações, as Árvores de Decisão são aplicadas conjuntamente com a tecnologia de indução de regras. A apresentação das regras é realizada segundo a prioridade estabelecida da seguinte forma: a regra mais importante é apresentada na árvore como o primeiro nó e as regras menos relevantes são mostradas nos nós subjacentes de acordo com o grau de relevância ou prioridade. A principal desvantagem advém da necessidade de utilização de uma quantidade considerável de dados quando se trata de estruturas complexas. As vantagens baseiam-se no facto de levarem em consideração as regras mais relevantes, apresentado estas um elevado índice de legibilidade e compreensão, possibilitando identificar de forma expedita os factores mais influentes.

As Árvores de Decisão são uma ferramenta surpreendentemente versátil, sendo uma das metodologias de aprendizagem indutiva mais utilizadas na actualidade, quer ao nível de aplicação, quer ao nível de trabalho e investigação académica.

4.3 Indução de Regras

A Indução de Regras é outra das técnicas de Data Mining muito divulgada [Berson et al., 2000], que permite a detecção de tendências e padrões em grupos de dados, i.e., regras sobre os dados [Sousa, 2004], consistindo na descoberta de regras de previsão, do tipo *Condição…Acção*, onde a *Condição* da regra especifica alguns atributos, e a *Acção* da regra prevê um valor para um determinado atributo cuja previsão é desejada [Fayyad et al., 1996]. A Indução de Regras surge associada às Árvores de Decisão, sendo usadas para representar o conhecimento representado nas Árvores de Decisão. Dado que uma árvore de decisão se

encontra na forma normal disjuntiva, é relativamente fácil traduzir este classificador para um conjunto de regras de decisão.

O objectivo da Indução de Regras é a descoberta de dependências entre os atributos ou valores através da análise das probabilidades condicionais, sendo os resultados apresentados sob a forma de regras $X \rightarrow Y$, que significa que "*se X está presente, então Y também tem probabilidade de estar presente*". O elemento X pode ser uma combinação de atributos e valores, formando assim regras mais complexas. Estas regras têm dois graus associados: o grau de confiança e o nível de suporte. A primeira é a probabilidade da condição da regra se verificar, e a segunda o número de casos onde a regra se verifica.

Algumas vantagens desta técnica é o modo directo de lidar com os dados, o desempenho, a facilidade de explicação e compreensão das regras, a fácil identificação dos passos para a solução do problema. O facto de as regras serem altamente heurísticas, pouco profundas, a dificuldade de manuseamento da informação incompleta ou valores inesperados, a explicação baseada na prova e não nos fundamentos teóricos constituem as principais desvantagens desta técnica [Langley e Simon, 1995].

4.4 Redes Neuronais Artificiais

A maior parte da investigação em Redes Neuronais Artificiais (RNA) foi inspirada e influenciada pelos sistemas nervosos dos seres vivos, em particular do ser humano. As RNA são modelos sub-simbólicos, e muitos investigadores acreditam que estas oferecem a aproximação mais promissora para a construção de verdadeiros sistemas inteligentes, com capacidade para ultrapassar a explosão combinatória associada à computação simbólica baseada em arquitecturas de *Von Newmann*.

Os princípios que ainda hoje vigoram sobre as RNA foram apresentados pela primeira vez em 1940 com o trabalho de Warren McCulloch e Walter Pitts, que demonstraram que as

redes de neurónios artificiais podiam calcular qualquer função aritmética ou lógica [Hagan et al., 1996]. Em 1949, Hebb sugeriu um método para que o modelo de neurónios de McCulloch e Pitts se ajustasse de forma automática, dando origem aos fundamentos das RNA.

O cérebro humano é uma estrutura altamente complexa, não linear e paralela. Possui uma capacidade de organizar os seus constituintes, denominados neurónios, por forma a executarem certas tarefas complexas (e.g., reconhecimento de padrões ou de voz), de uma forma intangível pelo computador mais potente até hoje concebido. Em termos de velocidade de processamento, um neurónio é cerca de 5 a 6 vezes mais lento do que uma porta lógica de silício. Contudo, esta limitação é ultrapassada pela estrutura maciçamente paralela do cérebro. Estima-se que o córtex humano possui cerca de 10 biliões de neurónios e 60 triliões de sinapses.

Um neurónio (Figura 4.4) é uma célula complexa que responde a sinais electro-químicos, sendo composto por um núcleo, um corpo celular, um numeroso conjunto de dendrites (entidades que recebem sinais de outros neurónios via sinapses), e por um axónio, que transmite um estímulo a outros neurónios através das sinapses. Um único neurónio pode estar ligado a centenas, milhares, ou mesmo dezenas de milhares de neurónios. No cérebro existem estruturas anatómicas de pequena, média e alta complexidade com diferentes funções, sendo possíveis parcerias.

Os neurónios tendem a agrupar-se em camadas, existindo três tipos principais de conexões (Figura 4.3): **(i)** *divergentes*, onde um neurónio pode estar ligado a outros neurónios através de uma arborização do axónio, **(ii)** *convergentes*, onde vários neurónios podem estar conectados a um único neurónio, e **(iii)** *encadeadas* ou *cíclicas*, as quais podem envolver vários neurónios e formarem ciclos.

As conexões podem corresponder a um de dois tipos de sinapses: *inibitórias* e *excitatórias*. Uma sinapse excitatória contribui positivamente para a activação de um neurónio. Por outro lado, uma sinapse inibitória, influencia a desactivação de um neurónio. Sinapses diferentes possuem diferentes intensidades, que influenciam em escala diferente o comportamento de outros neurónios.

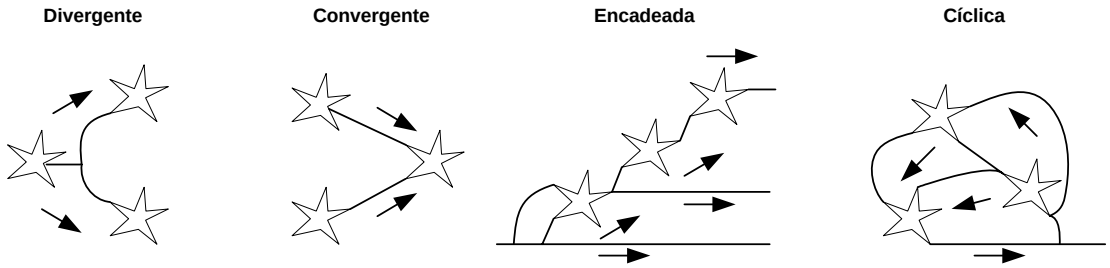


Figura 4.3 Tipos de conexões.

As RNA, também designadas por sistemas conexionistas são modelos simplificados do sistema nervoso central do ser humano. Uma RNA é um processador eminentemente paralelo, composto por unidades de processamento, designadas por neurónios ou nodos, que possuem uma capacidade natural para armazenar conhecimento empírico e torná-lo acessível ao utilizador. Assemelha-se ao comportamento do cérebro humano em dois aspectos: **(i)** o conhecimento é adqurido a partir de um ambiente, através de um processo de aprendizagem e, **(ii)** o conhecimento é armazenado nas conexões, também designadas por ligações ou sinapses, entre os nodos.

As RNA apresentam características únicas tais como:

- *Aprendizagem e generalização*: conseguindo descrever o todo a partir de algumas partes, constituindo-se como formas eficientes de aprendizagem e armazenamento de conhecimento;
- *Processamento maciçamente paralelo*: permitindo que tarefas complexas sejam realizadas num curto espaço de tempo;
- *Não linearidade*: atendendo a que a maioria dos problemas reais são de natureza não linear;
- *Adaptabilidade*: podendo adaptar a sua topologia de acordo com mudanças do ambiente;

- *Robustez e degradação suave*: permitindo processar o ruído ou informação incompleta de forma eficiente, sendo capazes de manter o seu desempenho quando acontece a desactivação de algumas conexões e/ou nodos;
- *Flexibilidade*: com um grande domínio de aplicabilidade.

O poder computacional de uma RNA alicerça-se em dois aspectos fundamentais: **(i)** uma topologia que privilegia o paralelismo, e **(ii)** a capacidade de aprendizagem e generalização. São estas duas características que tornam possível a resolução de problemas de elevada complexidade, de difícil resolução através de outros métodos.

Durante o processo de aprendizagem, dado por um algoritmo de aprendizagem ou de treino, os pesos das conexões são ajustados de forma a se atingir um dado objectivo; i.e., o estado de conhecimento da rede. Embora esta seja a forma tradicional de construir RNA, também é possível modificar a sua própria estrutura interna (ou topologia), à semelhança do que se passa no cérebro, onde os neurónios (Figura 4.4) podem morrer e novas sinapses, e mesmo neurónios, se podem desenvolver. Para se construir uma RNA tem que se determinar o número de neurónios, definir o seu tipo, como é que estes vão estar ligados, iniciar os pesos da rede e proceder ao treino da rede por aplicação de um algoritmo [Groth, 2000].

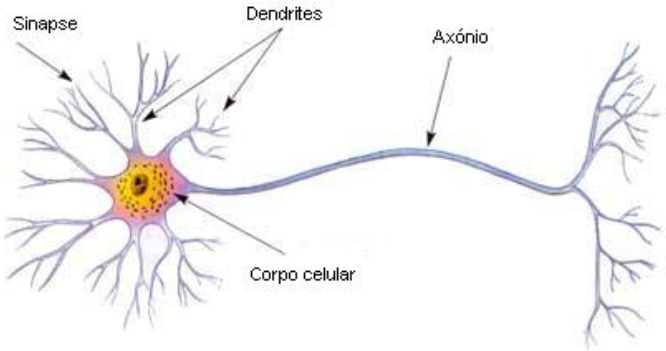


Figura 4.4 Estrutura de um neurónio natural.

Um neurónio artificial, denominado nodo (Figura 4.5), é a unidade de processamento chave para a operação de uma RNA. Embora existam diversos tipos de nodos, em princípio, comporta-se como um comparador que produz uma saída quando o efeito cumulativo das entradas excede um dado valor limite. Um nodo é constituído por três elementos fundamentais:

- Um conjunto de conexões que representam as sinapses ou conexões entre neurónios. Cada conexão tem associado um peso, i.e., um número real ou binário (w_{ij}). que tem um efeito excitatório (valores positivos) e inibitório (valores negativos). Assim, o sinal ou estímulo (x_j) como entrada da conexão é multiplicado pelo correspondente peso w_{ij}, onde i representa o nodo objecto de estudo e j o nodo emissor do sinal. Em algumas situações pode ainda existir uma conexão extra, denominada de bias, cuja entrada é fixada no valor +1, que estabelece uma certa tendência ou inclinação no processo computacional (i.e., adiciona uma constante para que se estabeleçam as correctas condições operacionais para o nodo).

- Um integrador (g), que reduz os n argumentos de entrada (estímulos) a um único valor. Frequentemente, é utilizada a função adição (Σ), pesando todas as entradas numa combinação linear.

- Uma função de activação (f), que pode condicionar o sinal de saída, introduzindo uma componente de não linearidade no processo computacional.

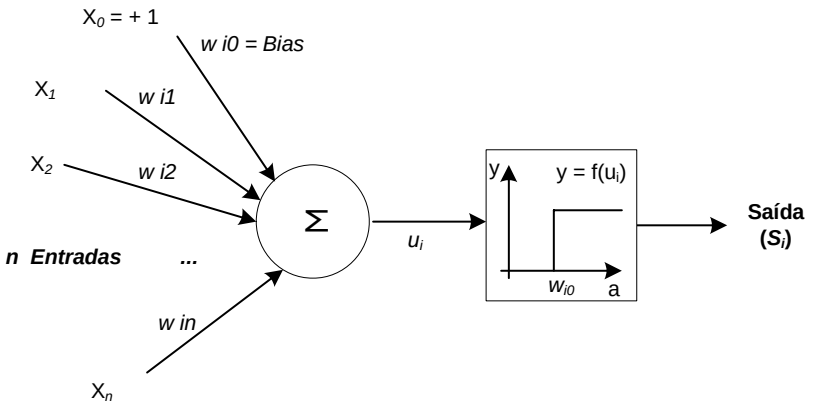


Figura 4.5 Estrutura do neurónio artificial de McCulloch e Pitts.

Em termos formais tem-se que este neurónio artificial ou nodo, é descrito pelas seguintes equações:

(Fórmula 4.3)

$$u_i = g(w_{i0}, x_1 \times w_{i1}, x_2 \times w_{i2}, ..., x_n \times w_{in})$$

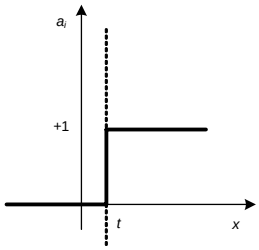
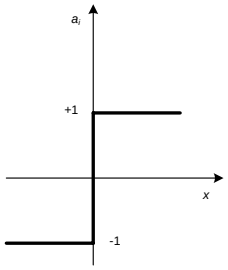
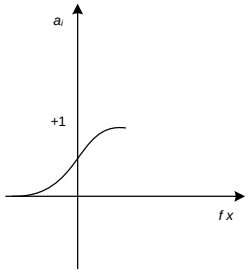
(Fórmula 4.4)

$$S_i = f(u_i)$$

Para um nodo *i* com *n* entradas e uma saída, onde u_i representa o ganho do nodo *i* e S_i a saída do nodo.

As três funções de activação (Tabela 4.3) mais utilizadas são: **(i)** *degrau* ou *heavisidade*, **(ii)** *sinal* e **(iii)** *logística* ou *sigmóide*.

Tabela 4.3 Funções de Activação.

FUNÇÃO	REPRESENTAÇÃO	EQUAÇÃO
DEGRAU		1 se x >= t 0 se x < t
SINAL		1 se x >= 0 -1 se x < 0
LOGÍSTICA/SIGMÓIDE		$\frac{1}{1+e^{-kx}}$

A função **(i)** é normalmente utilizada nos nodos do tipo *McCulloch-Pitts*, em que a saída toma o valor +1 apenas se o ganho for não-negativo, de acordo com a filosofia tudo ou nada. Em seguida, aparecem duas outras funções lineares. A função **(iii)** cuja forma é similar a um *S*, é a mais utilizada no uso de RNA. Trata-se de uma função crescente que exibe um balanceamento

gracioso entre um comportamento linear e não linear. Quando se varia a inclinação (*k*) obtêm-se funções com diferentes declives.

Os nodos interligam-se numa estrutura de rede denominada por *arquitectura* ou *topologia*. Existem vários tipos de arquitecturas ou topologias de RNA, organizando-se em três categorias:

1. *Redes Feedforward de uma Só Camada* (Figura 4.6). Uma RNA feedforward pode ser organizada por camadas, uma vez que não existem ciclos, dado que as conexões são unidireccionais (convergentes ou divergentes). A topologia mais simples é composta por uma camada de entrada, cujos valores de saída são fixados externamente, e por uma camada de saída. A camada de entrada não é contabilizada como camada numa RNA devido ao facto de nesta não serem efectuados cálculos.

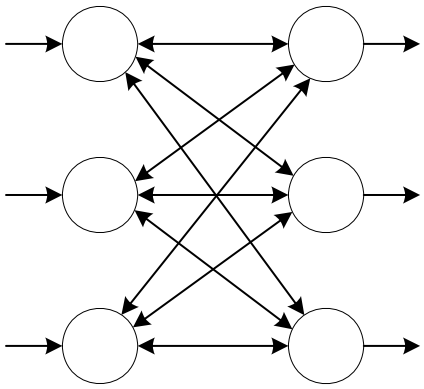


Figura 4.6 Rede de uma só camada.

2. *Redes Feedforward MultiCamada* (Figura 4.7). Esta classe de redes feedforward distingue-se por possuir uma ou mais camadas intermédias, cujos nodos são designados por nodos intermédios, sendo a sua função intervir de forma útil entre a entrada e a saída da rede. O aumento do número de camadas intermédias, eleva a capacidade da rede em modelar funções de maior complexidade. No entanto, este acréscimo implica o aumento de forma exponencial do tempo necessário para a aprendizagem.

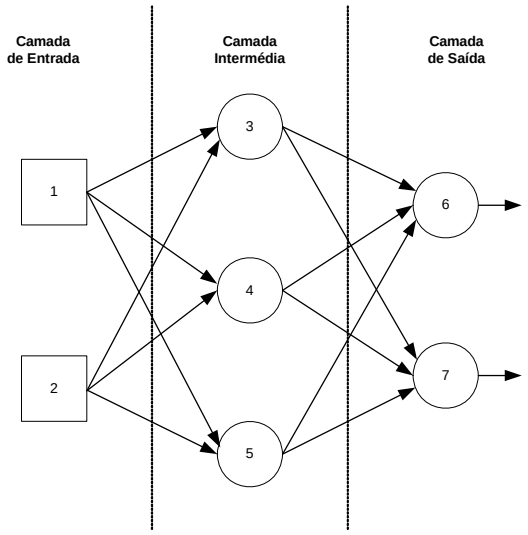


Figura 4.7 Arquitectura de uma *Rede Feedforward MultiCamada*.

3. *Redes Recorrentes* (Figura 4.8). A recorrência existe em sistemas dinâmicos quando uma saída de um elemento influencia de algum modo a entrada para esse mesmo elemento, criando-se assim um ou mais circuitos fechados. Quando se incluem uma ou mais conexões cíclicas numa rede, esta passa a ter um comportamento não linear, de natureza espacial e/ou temporal. Estas redes podem formar topologias arbitrárias.

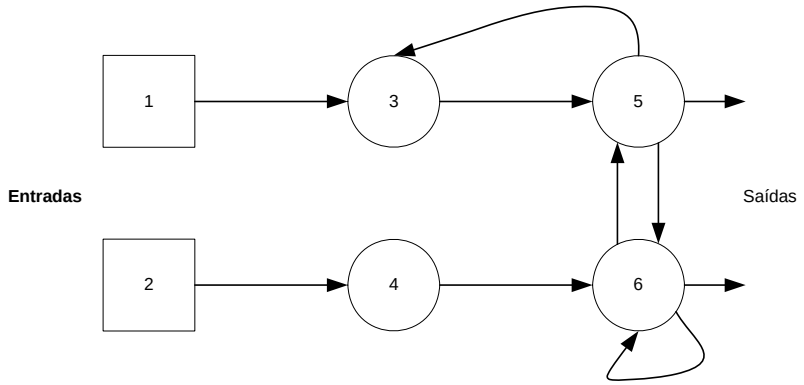


Figura 4.8 Arquitectura de uma *Rede Recorrente*.

Uma das propriedades das RNA é a sua capacidade para aprender a partir do seu ambiente. O processo de ***aprendizagem*** envolve a seguinte sequência de eventos:

- A RNA é estimulada por um dado ambiente;
- Alguns parâmetros livres (e.g., pesos das conexões) são alterados em resultado do estímulo recebido;
- A RNA responde de uma nova forma ao ambiente em virtude das alterações na sua estrutura interna.

A aprendizagem é executada a partir de um algoritmo de aprendizagem. Este consiste num conjunto de regras bem definidas para resolver um problema de aprendizagem. Os algoritmos de aprendizagem relacionam-se com o ambiente, e neste contexto está-se a falar de um paradigma (i.e., o modelo do ambiente em que a rede opera).

Existem três **paradigmas fundamentais de aprendizagem**: **(i)** Supervisionada, **(ii)** De Reforço e, **(iii)** Não Supervisionada (Figura 4.9).

O paradigma de aprendizagem *Supervisionada* **(i)** é bastante popular envolvendo a presença de um *"professor"*, sendo fornecidas respostas correctas à rede. Perante uma configuração que é apresentada a RNA produz uma resposta, que é comparada com a resposta correcta. A rede aprende a partir de um conjunto de padrões (P), onde cada exemplo ou caso de

treino é composto por um vector de entrada e por um vector de resposta ou saída. Durante o processo de aprendizagem é efectuada uma comparação entre o valor desejado com o valor de saída da rede, originando um erro. O erro é utilizado para ajustar os pesos das conexões, de forma a que o erro seja reduzido. Cada iteração do algoritmo de treino é composta por ajustamentos para os casos de treino. A aprendizagem é conseguida quando o erro é minimizado. Idealmente a RNA sabe mais sobre o seu ambiente após cada iteração.

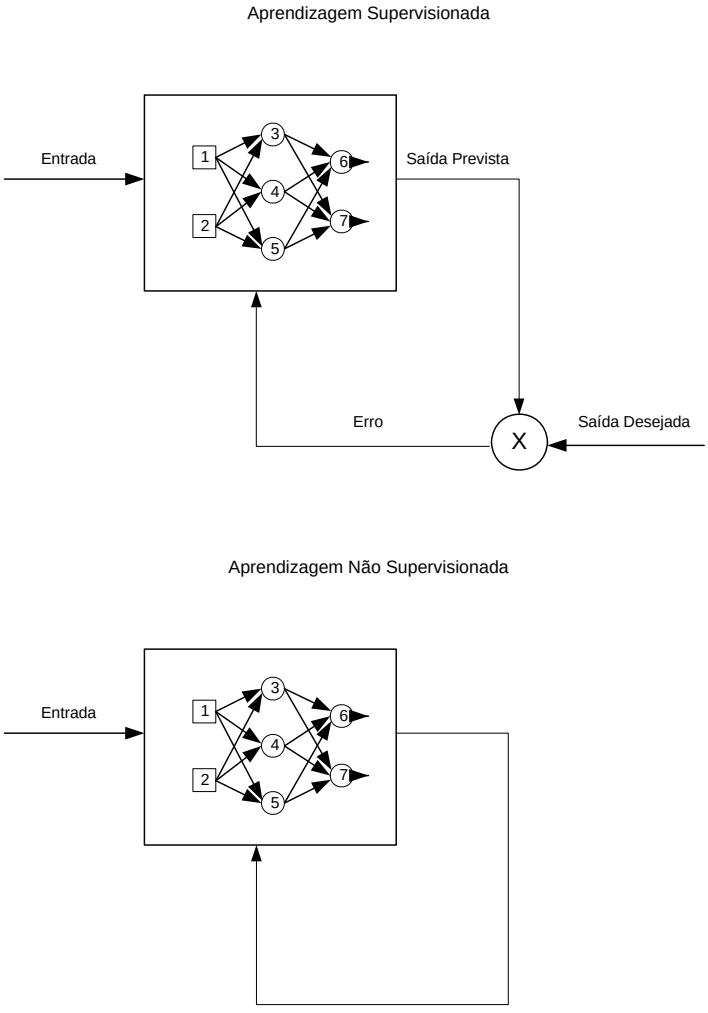


Figura 4.9 Paradigmas de aprendizagem Supervisionada e Não Supervisionada.

O paradigma de aprendizagem *De Reforço* **(ii)**, envolve tal como o anterior a presença de um "professor". No entanto, a resposta correcta não é apresentada à rede. Apenas se fornece uma indicação sobre se a resposta da rede é correcta ou errada. A partir desta informação a rede ajusta-se por forma a melhorar a sua eficácia. Um prémio é dado pelo reforço dos pesos das conexões que dão uma resposta correcta e uma penalidade é dada na situação oposta.

O paradigma **(iii)** *de aprendizagem não supervisionada*, segue uma abordagem diferente, onde não é fornecida ao sistema uma indicação externa acerca da resposta correcta. A aprendizagem é realizada através da identificação de características nos dados de entrada, adaptando-se a regularidades estatísticas ou agrupamentos de padrões dos exemplos de treino (e.g., Redes de Kohonen).

A escolha da arquitectura e do método de aprendizagem é influenciada pela tarefa de aprendizagem a ser desempenhada pela RNA sendo as categorias principais as seguintes: memória associativa, diagnóstico, reconhecimento de padrões, regressão/previsão, controlo, optimização e, filtragem/compressão de dados.

Existem várias **classes de RNA**, tendo as primeiras surgido nos anos 50. As redes do tipo ***Perceptron*** (Figura 4.10) são redes feedforward com apenas uma camada de nodos com várias entradas e saídas. Cada nodo calcula a soma pesada das suas entradas, sendo o valor de saída do tipo binário (0 ou 1) de acordo com determinado limite. A função de activação deste tipo de redes é a função *Step*.

Estas redes *Perceptron* destacam-se pela simplicidade de utilização, derivada de um número reduzido de parâmetros a ajustar, e ao facto do conjunto de padrões (*P*) de entrada não necessitar de um pré-processamento elaborado. Devido a estas características a aplicação resume-se contudo a padrões de complexidade não muito elevada, linearmente separáveis [2].

[2] Duas classes são consideradas linearmente separáveis se puderem ser separadas por uma linha recta.

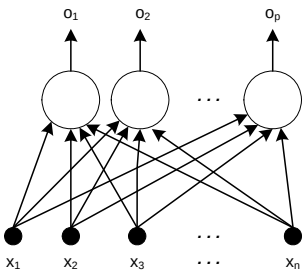


Figura 4.10 Rede *Perceptron*.

No final da década de 60 Minsky e Papert demostraram que uma rede feedforward com duas camadas pode solucionar muitas das restrições até aí encontradas na utilização das redes do tipo *Perceptron*. Contudo não apresentaram nenhuma solução para o problema do ajustamento dos pesos para as camadas escondidas. Só em 1986, Rumelhart, Hinton e Williams apresentaram uma solução para este problema, o algoritmo de ***Back-Propagation (Retropropagação)***.

As *Redes Feedforward Multicamada* (RMFC), ou *Redes Perceptrão Multicamada*, constituem uma das mais importantes e populares classes de RNA, sendo utilizadas em múltiplos domínios de aplicação, em problemas de memória associativa, classificação, reconhecimento de padrões, optimização e regressão. A não linearidade, a existência de nodos intermédios e o alto grau de conectividade tornam esta arquitectura muito poderosa como máquina de aprendizagem. No entanto estas características dificultam uma análise teórica ao processo de aprendizagem.

As RFMC são compostas por:

- um *conjunto de nodos de entrada*, onde surgem os estímulos do ambiente;
- um *conjunto de nodos intermédios*, unidades internas de processamento que aumentam a capacidade de aprendizagem de tarefas complexas, através da extracção progressiva de mais características;
- um *conjunto de conexões pesadas unidireccionais*;

- um *conjunto de funções de activação*, normalmente do tipo não linear e diferenciável sendo a função logística uma das mais utilizadas.

O sinal de entrada propaga-se para a frente através da rede, camada por camada, não existindo ciclos. O primeiro algoritmo de aprendizagem por correcção de erros e aprendizagem supervisionada foi desenvolvido por *Widrow e Hoff*, sendo conhecido por *Delta Rule*, *Least Mean Square* (LMS) ou *Adaptive Linear Neuron*. Trata-se de uma generalização do *Perceptron*, estendendo a técnica para entradas e saídas contínuas, apresentando uma única camada de neurónios. O erro é calculado como a diferença entre a resposta desejada e a resposta produzida pela RNA, ajustando-se o peso de forma a que se torne zero.

O algoritmo mais popular usado na aprendizagem supervisionada é o algoritmo de ***Back-Propagation***, ou os seus derivados, uma variação da regra de Widrow-Hoff. Trata-se de um algoritmo de referênca, já que constitui um método eficiente de computação para o treino de RFMCs, procurando o mínimo da função de erro no espaço de procura dos pesos, baseando-se em métodos de gradiente descendente. A combinação dos pesos que minimiza a função do erro é considerada a solução para o problema de aprendizagem.

O algoritmo de *Back-Propagation* utiliza dois passos [Cortez, 2000]:

1. *Em frente,* o vector de entrada é fornecido aos nodos de entrada, propagando-se em frente, camada por camada, estando neste passo os pesos da rede fixos.
2. *Retropropagação*, onde o erro é propagado para trás, desde a saída até aos nodos de entrada. De seguida, os pesos são ajustados segundo a regra de *Widrow-Hoff*.

Antes de se proceder ao inicio do treino de uma rede procede-se à escolha dos valores iniciais dos pesos associados às conexões entre os nodos, devendo ser pequenos e gerados de forma aleatória. Inicia-se então o treino da rede, seleccionando-se um caso de treino, de forma iterativa ou em lote. Em seguida, calcula-se o gradiente e ajustam-se os pesos. Uma iteração

termina quando todos os casos disponíveis tiverem sido considerados. O processo é terminado por critérios de paragem, por exemplo quando as mudanças nos pesos e na função de erro foram insignificantes. O algoritmo de aprendizagem pode convergir para um mínimo local, porém constata-se que quando se parte de um número elevado de casos de treino esta questão não assume relevância.

seja *A* o n.º de unidades na camada de entrada = tamanho do vector de entrada

C o n.º de unidades na camada de saída

Escolher o n.º de unidades na camada escondida (intermédia) *B*

1.Iniciar aleatoriamente os pesos da rede W ∈ [-0.1, 0.1]

2.Escolher um par de entrada-saída x_i – vector de entrada

x_j – vector de saida

3.Propagar as activações desde a camada de entrada até à camada de saída usando a função de activação Sigmóide (x_i -> h_i -> o_i -> y_i)

4.Calcular os erros das unidades na camada de saída, de seguida para a(s) camada(s) intermédia(s)

$\sigma 2_j = o_j(1 \quad o_j)(y_j \quad o_j)$ j=1,…,C

$\sigma 1_j = h_j(1 \quad h_j)\sum_{i=1}^{c}\sigma 2_j.w2_{ji}$ j=1,…,B

5.Ajustar os pesos desde a última camada escondida até à camada de entrada

$\Delta w2_{ij} = \eta.\sigma 2_j.h_i$ i=0,…,B j=1,…,C

$\Delta w1_{ij} = \eta.\sigma 1_j.h_i$ i=0,…,A j=1,…,B

η - taxa de aprendizagem (0.35 é considerado um bom valor)

6.Voltar ao passo 2. Quando todos os pares de entrada-saída do conjunto de treino estiveram processados passou-se uma época.

7.Repetir os passos 2..6 durante as épocas desejadas.

Algoritmo 4.2 Algoritmo de *Back-Propagation*.

O surgimento do algoritmo de *Back-Propagation*, influenciou de forma decisiva a investigação na área das *Redes Feedforward Multicamada*, o que motivou o aparecimento de novos algoritmos de treino, devido a dois factores: **(i)** o algoritmo de *BP* apresenta uma convergência lenta, e **(ii)** baseia-se no gradiente descendente, pelo que todas as técnicas de optimização não linear do gradiente podem ser aplicadas.

Diversas variantes baseadas no algoritmo de *Back-Propagation* têm sido propostas, tendo como base o uso de uma topologia fixa. No entanto, as melhorias mais signicativas advêm da utilização de algoritmos que adaptam não só os pesos mas também a topologia interna da rede a uma dada tarefa. Estas variantes podem ser classificadas em duas categorias: de adaptação global ou local. A primeira utiliza um conhecimento global do estado completo da rede, como a direcção de todo o vector de actualização dos pesos. Os últimos são baseados na informação específica de um peso, como o comportamento temporal da sua derivada parcial. Esta estratégia é mais próxima ao conceito das RNA sendo mais facilmente paralelizável, e tendendo a ser mais eficazes e robustas, apesar de usarem menos informação.

Uma *Rede Feedforward Multicamada* treinada por *Backpropagation* pode ser vista como uma forma prática para efectuar uma qualquer correspondência não linear, conseguindo com uma camada intermédia computar uma aproximação de uma qualquer função contínua. Com duas camadas intermédias é possível representar até funções descontínuas [Cortez, 2002]. Na utilização de RMFC um dos aspectos mais importantes é também o tempo de aprendizagem. De uma forma geral, a aprendizagem implica a procura dos elementos desconhecidos de uma RNA, normalmente pelo ajuste dos pesos. A aprendizagem numa rede com 100 pesos é bastante mais pesada em termos computacionais do que a de uma rede com 10 pesos, sendo uma relação bem maior que o factor 1:10 poderia sugerir. Seria muito útil que o tempo de aprendizagem fosse limitado por uma função polinomial sobre o número de variáveis, o que não acontece em termos práticos. O problema geral de aprendizagem em RNA não pode ser resolvido eficientemente para todas as instâncias. Não é conhecido um algoritmo que consiga realizar a aprendizagem num tempo polinomial, sendo até muito pouco provável que tal possa vir a existir. Com estes constrangimentos, diz-se que em geral o problema de aprendizagem em RNA

é NP-completo. Uma das possibilidades para ultrapassar a aprendizagem NP-completa das RNA reside no uso de arquitecturas adaptativas [Cortez, 2002].

Na **classe de RNA de aprendizagem não supervisionada**, existem dois algoritmos com grande utilização: as *redes competitivas* e as *redes de Kohonen*.

Nas **redes competitivas** quando um exemplo é processado pela rede, todas as unidades de saída vão concorrer pelo direito à resposta. Aquela que responde mais fortemente é a célula mais activa, assim, os pesos das ligações existentes nesta unidade são ajustados de forma a que a sua resposta seja reforçada, tornando assim mais provável que a identificação dessa qualidade da entrada seja efectuada por esta unidade.

Na década de 80 assistiu-se ao aparecimento de uma nova versão para aprendizagem não-supervisionada, conhecida como **redes auto-organizáveis ou redes Kohonen** (Algoritmo 3.2) – Kohonen self-organizing map (SOM) (Figura 4.11). Este tipo de RNA corresponde frequentemente a redes de camada única, formando uma família com grande plausibilidade biológica que se auto-organizam através do mecanismo de competição [Chester, 1993], por forma a considerar todos os casos da amostra. Nas redes competitivas quando um exemplo é processado pela rede, todas as unidades de saída vão concorrer pelo direito à resposta. Aquela que responde mais fortemente é a célula mais activa, assim, os pesos das ligações existentes nesta unidade são ajustados de forma a que a sua resposta seja reforçada, tornando assim mais provável que a identificação dessa qualidade da entrada seja efectuada por esta unidade.

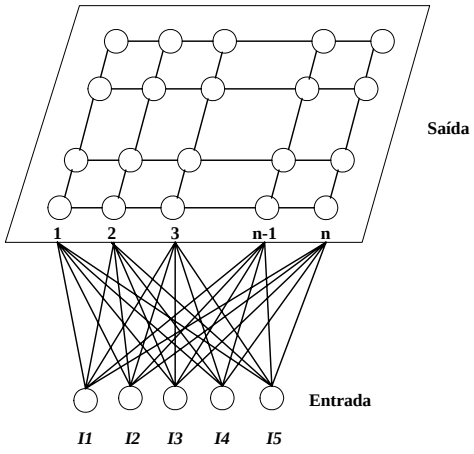


Figura 4.11 Arquitectura de uma Rede de Kohonen.

As redes de Kohonen permitem a identificação de similaridades entre vários sinais, agrupando-os em segmentos, tornando-se eficientes quando utilizadas sobre padrões com alguma relação entre si, podendo desta forma ser segmentados. Por outro lado, este modelo é complexo em comparação com outros, pois **(i)** a variável do raio de vizinhança deve ser ajustada adequadamente e, **(ii)** o número mínimo de iterações necessárias é de 500 vezes o número de neurónios de saída [Chester, 1993].

Estas redes são constituídas por um conjunto de nodos que se encontram directamente ligados a todos os nodos vizinhos, sendo normalmente representadas por estruturas bi-dimensionais, nas quais cada nodo tem associada uma determinada posição física na estrutura. Inicialmente, cada nodo possui uma posição aleatória, a qual é ajustada de forma sucessiva durante a fase de aprendizagem.

As redes de Kohonen não possuem níveis intermédios. O número de nodos no nível de entrada é calculado em função dos atributos de entrada, sendo o número de nodos no nível de saída igual ao número de segmentos obtidos na fase de aprendizagem. Nesta fase, cada nodo de saída compete com os outros nodos para ganhar a classificação de um dado registo. Os pesos das conexões são ajustados em função do sucesso/insucesso de cada nodo. O processo de modelação conduz ao agrupamento dos nodos em vectores, que representam as classes identificadas. Os pesos obtidos para as conexões permitem verificar a influência que cada atributo teve na identificação das classes [Santos, 2001].

1. Iniciar os pesos da rede com valores baixos, escolhidos aleatoriamente;

2. Inserir o padrão de entrada;

3. Calcular as distâncias de cada saída;

4. Seleccionar a menor distância;

5. Actualizar o peso no neurónio com menor distância (neurónio vencedor) e nos vizinhos deste. Ou seja, actualizar os pesos nos neurónios definidos pelo raio de vizinhança, que se determina através da fórmula:

$$w^j(t+1) = w^j(t) + \alpha(t)\left[x(t) \quad w^j(t)\right]$$

6. Voltar ao passo 2.

Algoritmo 4.3 Algoritmo de Kohonen.

Nas **redes de aprendizagem por reforço**, não são fornecidas as saídas correctas, mas sim atribuídos prémios ou castigos de acordo com a saída da rede. As alterações dos pesos das ligações são efectuadas somente nos níveis de actividade entre unidades conectadas entre si. Neste paradigma de aprendizagem, quando uma modificação no peso de uma conexão é efectuada não é conhecido o desempenho global do sistema. A aprendizagem do modelo termina quando a actividade está concluída ou é apresenta níveis baixos.

Nesta classe de RNA encontram-se os algoritmos de aprendizagem de: **Hebb**, **Hopfield**, e **Máquina de Boltzmann**.

A teoria de Hebb corresponde ao mais antigo e mais famoso postulado de aprendizagem e foi proposto por Hebb em 1949: "*Quando um axónio da célula A está demasiado próximo para excitar uma célula B e dispara esta de forma repetitiva ou persistente, então ocorre algum processo metabólico numa ou em ambas as células de forma a aumentar a eficiência da célula A em disparar B*" [Cortez, 2000]. Em termos práticos, se numa sinapse dois neurónios são activados simultaneamente, então a força (i.e., o peso) daquela sinapse deve ser aumentada. Por outro lado, se numa sinapse dois neurónios são activados de forma assíncrona, então aquela sinapse deve ser enfraquecida. A modificação nas sinapses tem relação com a correlação entre actividades dos dois neurónios envolvidos na ligação (i.e., quando a correlação é positiva o valor do peso aumenta e caso contrário diminui). Como a saída é reforçada em cada apresentação do padrão, os padrões mais frequentes assumem maior influência no vector dos pesos do neurónio [Sousa, 2004].

As **redes de Hopfield** são redes recorrentes, em que as saídas dos neurónios são realimentadas para a entrada da rede. Trata-se de uma rede de camada única, com uma função de activação do tipo logística. Todos os neurónios são de entrada e saída e as ligações são bidireccionais. Este tipo de rede possui neurónios dinâmicos.

A **Máquina de Boltzmann** é uma função estatística implementada nas RNA durante a década de 80. Neste tipo de rede são utilizados neurónios probabilísticos, ao contrário dos determinísticos que são usados nas redes Hopfield.

A maior **desvantagem** na utilização de **RNA** está relacionada com o facto de não transmitirem a sua aprendizagem (modelo) num formato perceptível pelo utilizador. São comparadas a "caixas negras" que dão respostas, mas que não transmitem conhecimento acerca

do processo que conduziu à obtenção das mesmas. Esta característica representa uma grande limitação na aplicação de RNA para tarefas de análise inteligente de dados. Em muitas aplicações, um dos requisitos é a necessidade de explicação das decisões tomadas. Por exemplo, num caso de atribuição de crédito, mesmo que a rede acerte sempre na decisão de concessão ou não, torna-se necessário explicar ao cliente o porquê da decisão.

Nos últimos anos, têm sido propostas diversas abordagens [Cramer, 1985][Fu, 1994] [Thrun, 1995] para a extracção de conhecimento de Redes Feedforward Multicamada treinadas através do algoritmo *Back-Propagation*. A maioria delas concentra-se na compilação de regras simbólicas a partir do modelo final, sendo conhecidas como *procedimentos de extracção de regras*.

Estes procedimentos podem ser agrupados em duas categorias [Berthold e Hand, 2002]: a primeira chamada de *global*, em que é gerado um conjunto de regras que caracterizam o comportamento da rede em termos de mapeamento de entradas/saídas; a segunda, chamada de *local*, decompõe a rede original num conjunto de redes mais pequenas, de uma só camada, sub-redes, tornando o mapeamento entre entradas/saídas de modelação mais fácil em termos de regras simbólicas. As regras são depois agrupadas de modo a obter um conjunto de regras de maior abrangência que descrevem a rede original.

Contudo estes métodos para extracção de regras a partir das RNA **(i)** implicam um esforço enorme de computação, e **(ii)** apresentam um baixo grau de aplicabilidade das regras. Estas razões motivaram a que o esforço de investigação fosse redireccionado para diferentes arquitecturas de RNA, que apresentam um grau superior de interpretabilidade.

4.5 Algoritmos Genéticos

Os *algoritmos genéticos* (Algoritmo 4.4), conjugam princípios da biologia e das ciências da computação, configurando processos adaptativos de procura num espaço de soluções, por aplicação de operadores modelados de acordo com o conceito de herança inerente à teoria de Darwin da evolução das espécies [Santos, 1999]. Na origem das espécies, Darwin descreve a teoria da evolução recorrendo à selecção natural. Cada espécie possui um número

elevado de indivíduos, e num instinto de sobrevivência, aqueles que sobrevivem são os que melhor se adaptam ao ambiente. As alterações do ambiente provocam mutações genéticas nas espécies, por forma a estas se adaptarem às alterações do seu meio ambiente.

O algoritmo genético inicia-se com uma população de indivíduos, soluções possíveis para o problema, gerados de forma aleatória. A informação que um indivíduo disponibiliza, e que atende aos valores dos parâmetros do problema em equação, e representada por um cromossoma, de um modo análogo à estrutura vigente no *ADN* [Cortez, 2002]. Um cromossoma, é por sua vez composto por um conjunto de genes (caracteres). Um valor possível para um gene é designado por alelo. A qualidade de cada solução (cromossoma) é medida por uma função chamada de aptidão, sendo os indivíduos avaliados de acordo com esta.

Em cada ciclo, parte-se da população actual (P_t). À semelhança do mundo real, os individuos da população são sujeitos a uma série de operações, tais como: selecção dos progenitores, cruzamentos entre pares de progenitores e mutação dos descendentes. Como resultado é criada uma nova geração de indivíduos. O processo repete-se durante várias gerações até que esteja satisfeita uma dada condição de paragem, definida por exemplo, pelo número máximo de gerações (T).

```
1. Iniciar o tempo (t:=0)
2. Gerar a população inicial, de forma aleatória (P0)
3. Avaliar P0
4. Enquanto t<T Fazer
   4.1 Seleccionar os progenitores a partir da população actual (Pt)
   4.2 Aplicar cruzamento aos progenitores para gerar descendentes
   4.3 Aplicar mutação aos descendentes
   4.4 Avaliar a nova população (Pt+1)
   4.5 Actualizar o tempo (t=t+1)
```

Algoritmo 4.4 Algoritmo Genético.

O processo de representação do fenótipo (solução) em genótipo (cromossoma), é designado de codificação, sendo operação crucial para uma boa resolução dos problemas utilizando *AGs*.

Desde que foi criado, o modelo original de Holland sofreu diversas alterações ao longo do tempo, por forma a aumentar a gama de aplicabilidade e o sucesso. A estrutura deste modelo generalista designado por *Algoritmo Genético e Evolucionário* [Rocha, 1997] tem como principais diferenças:

- aceitam-se outros tipos de representação genética;
- a população inicial pode ser criada por outros métodos que não o aleatório;
- admite-se uma hipótese de condição de paragem mais flexível;
- generaliza-se o conceito de operador genético, que engloba as operações de cruzamento e mutação;
- são definidos novos parâmetros que controlam o processo de renovação da população, i.e., como se passa de uma geração de indivíduos para a seguinte;
- os processos de selecção são generalizados.

4.6 Aproximação de Vizinhanças

A técnica de aproximação de vizinhanças é baseada no princípio de que registos semelhantes estão próximos uns dos outros, quando analisados numa perspectiva espacial. A verificação da localização dos registos, interpretados como pontos no espaço, permite a identificação de regiões, denominadas classes (ou segmentos), que definem características comuns para os registos que representam. A complexidade na utilização desta técnica aumenta à medida que cresce o número de registos a analisar, uma vez que cada registo é comparado com os restantes registos da amostra.

A utilização desta técnica é realizada da seguinte forma: construção de partições dos objectos armazenados na *BD*, num conjunto de k classes, sendo k um parâmetro de entrada.

Cada classe pode ser representada através do seu centro de gravidade (estratégia *k-means*), i.e., pela localização média de todos os membros do segmento, ou por um dos objectos da classe próximo do seu centro (estratégia *k-medoid*). Para determinação das classes, cada registo é transformado num ponto no espaço apresentando este tantas dimensões quantos os atributos em análise. O valor de cada campo é interpretado como a distância da origem até à sua localização num dado eixo.

O processo de obtenção das classes é iniciado com centróides em posições aleatórias, as quais são optimizadas iterativamente através da movimentação dos centros.

4.7 Avaliação de Modelos

Após a geração dos modelos, é necessário avaliar o desempenho dos mesmos. Existem vários métodos de amostragem para estimar a capacidade de generalização de um modelo: *Estatística Simples*, *Validação com Divisão da Amostra*, *Validação Cruzada* e, *Bootstrapping*.

O método mais popular para a estimação do erro de generalização é a *Validação com Divisão da Amostra*, que se baseia numa divisão dos dados do problema em casos de treino, para a aprendizagem do modelo, e casos de validação para estimar o erro de validação. Como pontos fortes temos a sua simplicidade e rapidez, embora produza uma redução efectiva dos dados disponíveis para treino.

A *Validação Cruzada* (Figura 4.12) é um melhoramento do método de *Validação com Divisão da Amostra*, que permite a utilização de todos os casos disponíveis. Na validação cruzada *k*-desdobrável, os dados (*P*) são divididos em *k* subconjuntos mutualmente exclusivos (P_1, P_2,…, P_k) de comprimentos aproximadamente iguais. Os modelos são treinados e testados k vezes. O erro final de generalização é dado pela média dos erros de validação obtidos durante k vezes. Os valores de *k* podem variar entre 2 e *n* embora o valor 10 seja o mais popular (e.g., *Ten Fold Cross Validation*). A *Validação Cruzada* é notavelmente superior à validação com divisão da amostra para pequenos conjuntos de exemplos de treino. Todavia, é conseguida à custa de um considerável esforço computacional [Cortez, 2002].

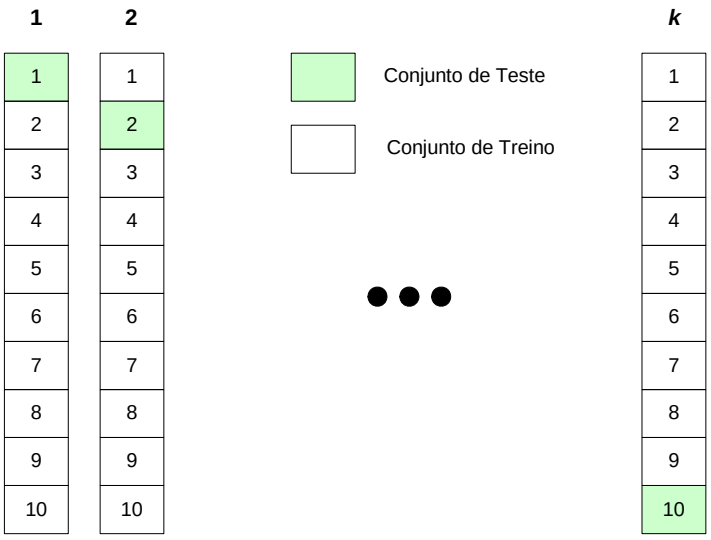


Figura 4.12 Validação Cruzada com *k* iterações.

Existem várias técnicas de avaliação, que devem ser seleccionadas em função do tipo de problema, sendo as mais utilizadas em problemas de classificação, a *Matriz de Confusão*, as *Curvas ROC*, e nos casos de regressão medidas como *Mean Absolute Deviation*, *Sum Squared Error*, *Mean Squared Error* e *Root Mean Squared Error*.

Avaliação de Modelos de Classificação

1. Matriz de Confusão

Quando se trata de problemas de classificação, uma das técnicas mais usadas é a *Matriz de Confusão* [Kohavi e Provost, 1998].

A Matriz de Confusão ou de Erros (Tabela 4.4) é usada para avaliar o resultado de uma classificação, mapeando os valores previstos por um modelo com os valores desejados.

Tabela 4.4 Matriz de Confusão 2 x 2.

↓ Desejado \ Previsto →	**Negativo**	**Positivo**
Negativo	**TN**	FP
Positivo	FN	**TP**

A partir da matriz de confusão é possível calcular as seguintes medidas:

(Fórmula 4.5)

$$\text{Sensibilidade (erro tipo II)} = \frac{TP}{FN + TP} \times 100(\%);$$

(Fórmula 4.6)

$$\text{Especificidade (erro tipo I)} = \frac{TN}{TN + FP} \times 100(\%);$$

(Fórmula 4.7)

$$\text{Precisão (Accuracy)} = \frac{TN + TP}{TN + FP + FN + TP} \times 100(\%);$$

Estas quatro medidas de desempenho são independentes do custo e das probabilidades das classes.

2. Curva ROC – Receiver Operating Characteristic

Outra das medidas de avaliação de classificadores são as curvas ROC (Figura 4.13). A análise ROC [Egan, 1975] surgiu no âmbito da "*Teoria da Detecção de Sinal*", desenvolvida durante a II Guerra Mundial para a análise de imagens de radar. A tarefa consistia em descobrir se um dado sinal era indicador de um barco inimigo, aliado ou simplesmente ruído. A detecção de sinal mede a capacidade do receiver operator do radar em fazer estas importantes distinções. A sua utilidade foi reconhecida mais tarde e a sua utilização estendida para a comparação de exames clínicos para diagnosticos médicos [Ribeiro, 2003]. Esta curva permite a avaliação do desempenho de um classificador, sendo apropriado quando existem dois estados possíveis, i.e., duas classes. A curva ROC estabelece a relação entre a taxa de TP (TP/(TP+FN)) e a taxa de FP (FP/(FP+TN)), variando num determinado *treshold*.

A curva ROC permite visualizar o compromisso entre a sensibilidade (taxa de TP) e a selectividade (1 – taxa de FP) do algoritmo. Na situação ideal, o algoritmo deveria possuir indicadores máximos de sensibilidade e selectividade, ambos iguais a um.

A fim de fornecer uma comparação do desempenho dos diferentes classificadores, independente da distribuição de classes ou de custos de erros, existem duas técnicas frequentemente utlizadas: AUC [Metz, 1978] e ROCCH [Provost e Fawcett, 1997].

A AUC (Area Under Curve) consiste numa métrica de desempenho do classificador obtida através do cálculo da área por baixo da curva ROC do classificador, assumindo valores entre 0 e 1 e pode ser interpretada como a probabilidade de um exemplo positivo, escolhido aleatoriamente, ser classificado como tal.

A análise ROCCH (ROC Convex Hull) permite declarar um subconjunto de classificadores como potencialmente óptimos. Incluídos todos os pontos que constituem as curvas ROC dos diferentes classificadores, e formada a convex hull correspondente, é realizada uma análise dos pontos que estão sobre a linha de convex hull. Se um ponto está sobre a convex hull, existe então uma linha tangente a esse ponto que tem uma taxa de TP superior, sendo o classificador representado por esse ponto considerado como óptimo sob a distribuição assumida correspondente a essa inclinação.

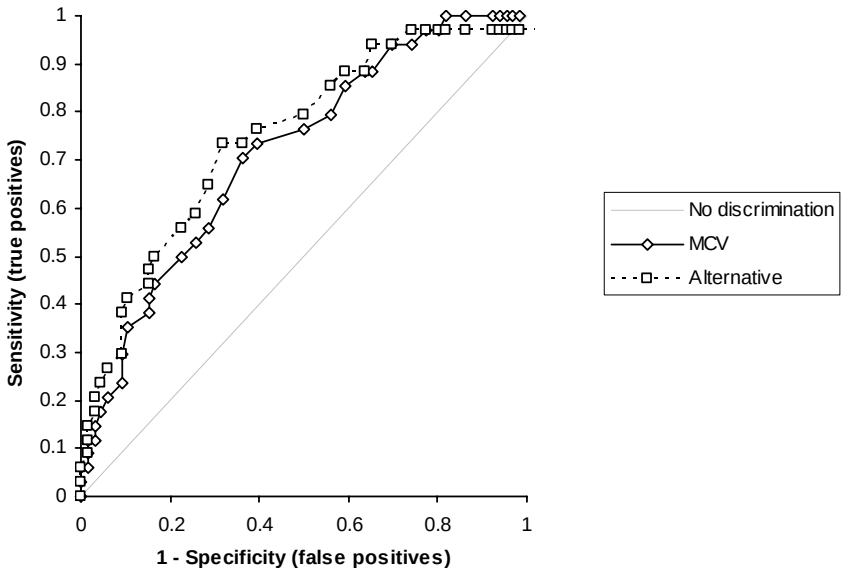


Figura 4.13 Exemplo de Curva ROC para comparação de dois classificadores.

Avaliação de Modelos de Previsão

Um modelo de regressão é um estimador, que tenta estimar o valor associado a cada um dos exemplo. O principal objectivo de um método de regressão é conceber o "melhor" modelo segundo uma medida da estimativa do erro.

Quando se trata de problemas de regressão, o erro (ou resíduo) *e* é medido por:

(Fórmula 4.8)

$$e=d-d'$$

onde *d* representa o valor desejado e *d'* o valor estimado pelo modelo.

Para o conjunto de dados: $x_1 \rightarrow d_1, \ldots, x_n \rightarrow d_n$, pode ser calculado um erro global:

(Fórmula 4.9)

Mean Absolute Deviation (MAD): $MAD = \frac{\sum_{i=1}^{N} |e_i|}{N}$

(Fórmula 4.10)

Sum Squared Error (SSE): $SSE = \sum_{i=1}^{N} e_i^2$

(Fórmula 4.11)

Mean Squared Error (MSE): $MSE = \frac{SSE}{N}$

(Fórmula 4.12)

Root Mean Squared Error (RMSE): $RMSE = \sqrt{MSE}$

As medidas MAD e MSE são duas das medidas de avaliação mais usadas no âmbito da regressão. Ambas têm as suas origens na estatística, tendo como intuito definir quão melhor é

um estimador para um conjunto de valores. Se ele for uma boa medida de centralidade representa a distribuição dos valores, melhor que qualquer outro valor. A melhor medida de centralidade será aquela que miniza as medidas MAD ou MSE [Ribeiro, 2003].

Segundo Torgo [Torgo, 1999], estas medidas quando aplicadas à avaliação de modelos de regressão servem diferentes propósitos. Se na previsão de valores for aceitável cometer alguns erros extremos, desde que nos aproximemos na maior parte das vezes do valor real, a medida MAD é a mais adequada. Quando se está perante situações em que é crucial não cometer erros extremos, mesmo que possam existir erros pequenos e frequentes, então a medida *MSE* deverá ser utilizada. Isto porque, esta medida utiliza o quadrado das distâncias das previsões aos valores reais. Desta forma, grandes distâncias serão amplificadas relativamente às pequenas distâncias.

Capítulo 5

SCAE

Sistema de **C**onhecimento baseado em **D**ata **M**ining para **A**nálise da **E**stabilidade de **E**struturas **M**etálicas

Neste capítulo é apresentado o protótipo de um Sistema de Conhecimento para a Análise de Estabilidade de Estruturas Metálicas de Engenharia Civil (SCAE), baseado em técnicas de Data Mining, e uma abordagem para o desenvolvimento de modelos para a previsão da Carga Crítica em vigas de aço sujeitas a cargas concentradas e da Tensão Crítica de vigas em I de inércia variável.

5.1 Introdução

Uma das dificuldades nas fases de concepção e projecto de estruturas metálicas de engenharia civil reside na determinação da Carga Crítica de vigas sujeitas a cargas concentradas

e da Tensão Crítica de vigas em I de inércia variável, pelas limitações actuais das fórmulas de cálculo, e pelas dificuldades de simulação de diferentes cenários em laboratório. Em face destas limitações os objectivos práticos deste trabalho consistem no (i) desenvolvimento de um protótipo do sistema SCAE que permita simular diferentes cenários e que (ii) incorpore modelos de DM, para previsão, com acuidades superiores às fórmulas de cálculo actualmente existentes.

A revisão teórica realizada nos capítulos anteriores, permitiu a identificação dos requisitos a cumprir no desenvolvimento dos modelos de previsão e justificar as opções estruturais em que a abordagem proposta se baseia.

Este capítulo encontra-se organizado da seguinte forma: nas secções seguintes são enunciados os requisitos do sistema SCAE, tecnologias de desenvolvimento adoptadas, arquitectura do sistema, e o processo de aquisição de conhecimento para o SCAE, através de um processo de Descoberta de Conhecimento em Bases de Dados para geração de modelos de previsão da Carga e da Tensão Crítica.

5.2 Sistema SCAE

5.2.1 Requisitos

Para o Sistema SCAE foram definidos os seguintes requisitos que orientaram o seu desenvolvimento:

- permitir a análise da estabilidade de estruturas metálicas de Engenharia Civil através da utilização dos modelos de previsão da carga crítica e da tensão crítica criados a partir de um processo de Descoberta de Conhecimento de Base de Dados.

- implementar um interface com o utilizador constituído por um painel de controlo, no qual o especialista de Engenharia Civil tem a possibilidade de ajustar os diferentes parâmetros que influenciam o comportamento de uma

estrutura metálica, obtendo como resultado a previsão do valor da carga crítica e da tensão crítica.

Tendo em conta os requisitos enunciados, foi desenvolvido um protótipo utilizando a plataforma JAVA, com incorporação da biblioteca Xelopes – versão 1.15, para a previsão da tensão crítica de vigas com perfil em I de inércia variável.

5.2.2 Tecnologias de Data Mining

Para o desenvolvimento do sistema SCAE foram seleccionadas as seguintes tecnologias: Xelopes Library para implementação do Núcleo do Sistema; PMML (Predictive Markup Language) para especificação do modelo de previsão; JAVA para programação e integração dos diferentes componentes do sistema.

5.2.2.1 Xelopes Library

O projecto Xelopes tem por objectivo proporcionar a integração de componentes de Data Mining em aplicações informáticas das mais diversas áreas de actuação, permitindo o desenvolvimento de Sistemas de Conhecimento com incorporação de técnicas e algoritmos de DM.

Trata-se de um projecto que segue uma filosofia de concepção assente em três pilares:

1. Diferentes níveis de abstracção;
2. Independência de plataformas e fontes de dados;
3. Automatização do processo de geração de modelos.

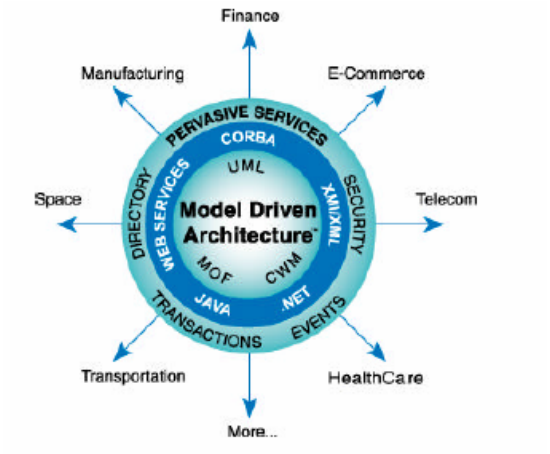


Figura 5.1 Camadas da Arquitectura Orientada aos Modelos [Prudsys, 2003].

A biblioteca Xelopes, especificada em UML, fornece uma framework para utilização de algoritmos de Data Mining, segundo uma Arquitectura Orientada aos Modelos (MAD) (Figura 5.1). A ideia base da MDA é o desenvolvimento de modelos independentes da plataforma (PIM – Platform-Independent Model), que são mapeados para uma ou várias Plataformas Específicas de Modelos (PSMs – Platform-Specific Models).

Além da UML estão dois meta-modelos que são centrais na Arquitectura Orientada aos Modelos: o Meta-Object Facility (MOF), que define um meta-modelo comum para todas as especificações de modelação do Object Management Group (OMG), e a Common Warehouse Metamodel (CWM), que define um meta-modelo que permite a troca de dados entre bases de dados e mesmo entre data warehouses de diferentes organizações. A CWM está para a modelação de dados como a UML está para a modelação de aplicações, assentando num subconjunto da UML.

A biblioteca Xelopes pode ser caracterizada pela:

- independência da plataforma, uma vez que o PIM se encontra modelado em UML, e os interfaces podem ser derivados de forma expedita para diferentes plataformas;
- pela independência da origem de dados, uma vez que uma das ideias de base da biblioteca Xelopes é a abstracção da matriz de dados. Uma vez que os algoritmos de Data Mining trabalham no espaço Cartesiano dos atributos de origem é necessária uma matriz de dados (fornecida de forma implícita);
- pelo suporte aos principais standards de DM como: CWM Data Mining, PMML, OLE DB, bem como para o Weka.

A Platform Independent Model (PIM) da biblioteca Xelopes assenta nos pacotes de Data Mining e Transformação da especificação CWM, extendendo o pacote CWM Core.

Os diagramas de classes da PIM representam oito áreas conceptuais: Modelação (Model), Definições (Settings), Atributos (Attributes), Algoritmos (Algorithms), Acesso aos Dados (Data Acces), Transformações (Transformations), Automação (Automation) e Predictive Model Markup Language (PMML).

A área conceptual do Modelo CWM (Figura 5.2) representa de forma genérica o modelo de Data Mining, que é constituído pelas classes: *MiningModel* (representação do modelo), *MiningSettings* (definições necessárias para o desenvolvimento do modelo), *ApplicationInputSpecification* (especificação os atributos de entrada do modelo) e a *MiningModelResult* (representação do resultado obtido na fase de teste/aplicação do modelo criado). A classe *SupervisedMiningModel* extende a classe *MiningModel* para aprendizagem supervisionada.

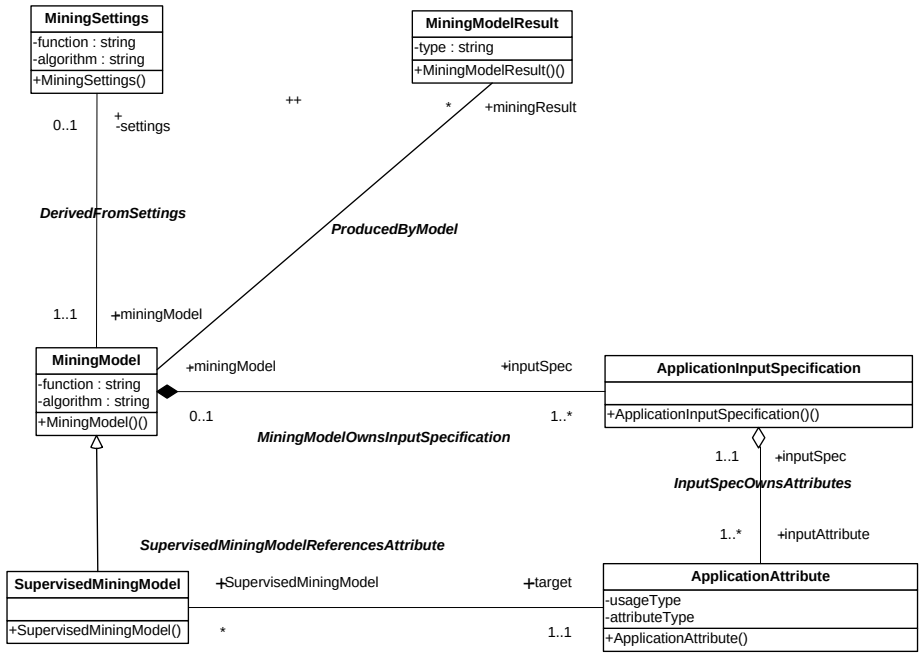


Figura 5.2 Diagrama do Modelo da especificação CWM**.**

O atributo *function* da classe *MiningModel* descreve a classe do objectivo de Data Mining (e.g., Associação - *AsssociationRules*), e o atributo algorithms é utilizado para especificar o algoritmo de aprendizagem automática (e.g., Árvores de Decisão - *DecisionTrees*).

Os modelos criados e manipulados com recurso à biblioteca Xelopes são especificados segundo o standard PMML, uma vez que esta biblioteca dispõe do interface *Pmmlable*, que indica que a classe pode ser convertida para um objecto do tipo PMML (i.e., elemento ou documento) ou construída a partir de um objecto PMML. Os interfaces principais de um algoritmo de Data Mining na Xelopes encontram-se representados na figura 5.3.

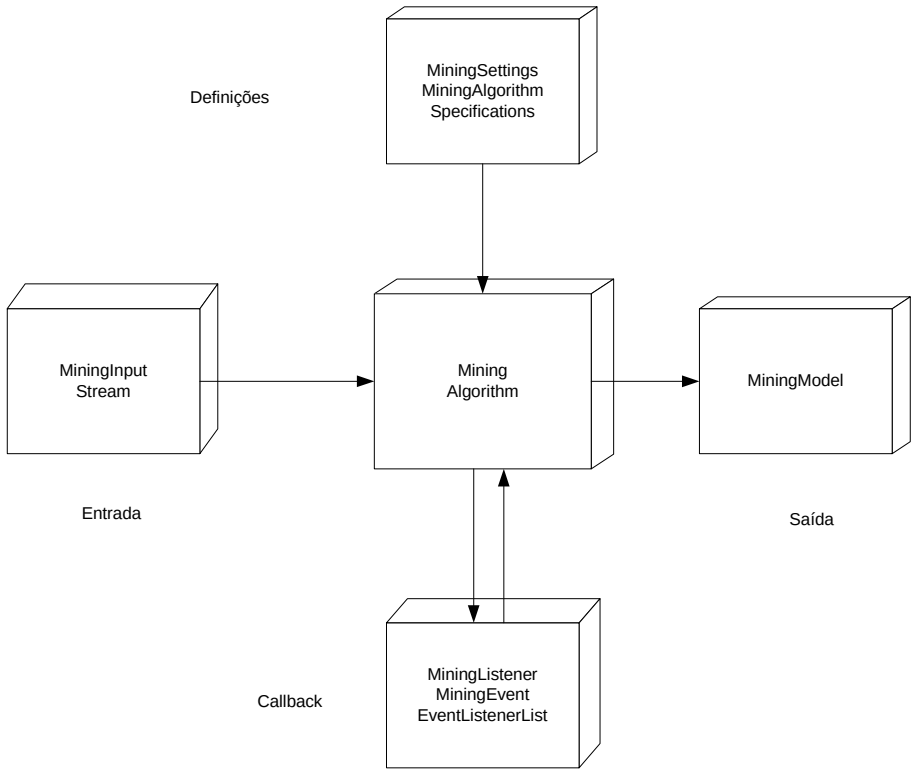


Figura 5.3 Interfaces principais de um algoritmo de Data Mining na Xelopes [Prudsys, 2003].

5.2.2.2 PMML

A especificação PMML (Predictive Model Markup Language) (Figura 5.4) tem a sua origem no esforço conjunto de organizações produtoras de ferramentas de Data Mining e de um grupo de investigadores. Este consórcio, denominado de Grupo de Data Mining (Data Mining Group) [DMG, 2004] desenvolveu uma especificação independente das ferramentas de Data Mining, o que permite a especificação formal de modelos de DM independente da plataforma

em que foram criados, facilitando a eliminação de barreiras e/ou incompatibilidades entre modelos e/ou ferramentas.

```
<!ENTITY % A-PMML-MODEL '(TreeModel |
NeuralNetwork |
ClusteringModel |
RegressionModel |
GeneralRegressionModel |
NaiveBayesModels |
AssociationModel |
SequenceModel )' >

<!ELEMENT PMML ( Header,
MiningBuildTask?,
DataDictionary,
TransformationDictionary?,
(%A-PMML-MODEL;)*,
Extension* )>
<!ATTLIST PMML\
version    CDATA    #REQUIRED
>

<!ELEMENT MiningBuildTask (Extension*) >
```

Figura 5.4 Cabeçalho de um documento PMML.

Esta especificação de modelos de DM assenta na linguagem XML (Extensible Markup Language) – projecto do W3C, sendo o desenvolvimento da especificação PMML supervisionado pelo consórcio XML Working Group. O XML tornou-se atractivo como base para o PMML: pela utilização generalizada na Internet, pelo suporte a *tags* e definição de linguagens de *markup* para diferentes classes de documentos, o que permite que para determinado domínio específico de utilização seja possível definir as descrições de um modelo. O modelo é editado como um documento XML.

5.2.3 Arquitectura

A arquitectura do sistema SCAE assenta na estrutura típica de um Sistema de Conhecimento (Figura 5.5), exceptuando o componente Memória de Trabalho, uma vez que o

Sistema SCAE não incorpora um processo de raciocínio com interacção durante o processo com o utilizador. O utilizador especifica num primeiro momento as dimensões de uma estrutura (i.e., viga metálica) e o sistema submete essa configuração a um modelo de previsão do tipo Rede Neuronal Artificial, que apresenta como *output* a carga crítica prevista para o elemento submetido.

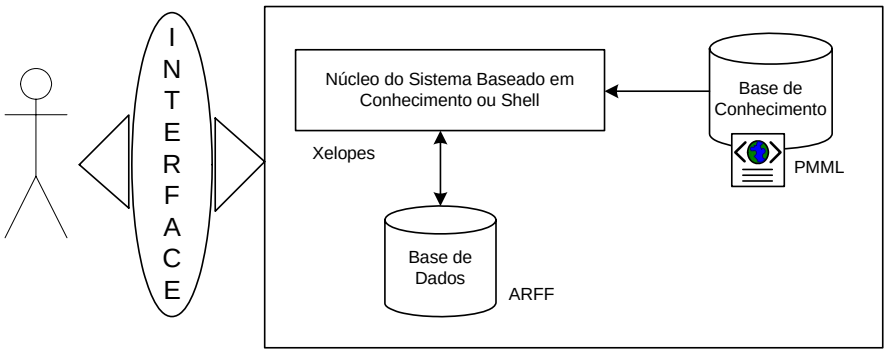


Figura 5.5 Arquitectura do Sistema SCAE.

Para o desenvolvimento do Núcleo do Sistema SCAE foi embebida a biblioteca Xelopes, sendo este núcleo responsável pelas seguintes macro tarefas:

- validação dos dados de entrada e do documento PMML com a especificação do modelo de DM de previsão da tensão crítica;
- aplicação do modelo aos dados;
- transformação da saída do modelo de previsão.

A Base de Conhecimento do Sistema SCAE encontra-se especificada em PMML, que representa o conhecimento sobre o problema, usando um formalismo de representação do tipo sub-simbólico, uma vez que se trata de um modelo de previsão do tipo Rede Neuronal Artificial.

A Base de Dados do sistema proposto para armazenamento das configurações da estrutura de Engenharia Civil para a qual se pretende obter o valor da tensão crítica, definidas a partir do interface, encontra-se especificada em formato ARFF (Attribute-Relation File Format) [ARFF, 2004]. Trata-se de um formato de dados do tipo texto que descreve uma lista de instâncias com o mesmo conjunto de atributos, generalizado para utilização em problemas de Data Mining, desenvolvido pelo Machine Learning Project no Departamento de Ciências da Computação da Universidade de Waikato para utilização pelo sistema de aprendizagem automática Weka [Weka, 2004]. Os pontos fortes deste formato de dados residem na sua versatilidade e facilidade de utilização, por exemplo no que diz respeito à utilização em múltiplas plataformas.

O Interface com o utilizador do sistema SCAE é um painel de controlo (Figura 5.7) para ajustamento dos diversos parâmetros que influenciam a estabilidade/comportamento de vigas metálicas de inércia variável. O ajustamento dos parâmetros é realizado através de barras de deslocamento horizontais, o que facilita a criação e teste de múltiplos cenários com a precisão inerente ao modelo de previsão do tipo Rede Neuronal utilizado, permitindo o teste à tensão crítica de estruturas com parâmetros diferentes, conduzindo assim para uma tomada de decisão pelo especialista da área de Engenharia Civil mais segura, evitando a utilização por exemplo de coeficientes de segurança inadequados, que levam a uma de duas situações: a utilização excessiva de materiais, ou a criação de elementos estruturais mais vulneráveis.

Os intervalos de valores dos parâmetros do painel de controlo do Sistema SCAE estão parametrizados no intervalo de valores utilizados para o desenvolvimento do modelo de previsão, no entanto não foram fornecidas para desenvolvimento do modelo de previsão todas as combinações possíveis dos parâmetros representados, uma vez que a geração de casos de teste é muito complexa e dispendiosa. Desta forma assegura-se também a possibilidade de teste de diferentes cenários impossibilitados de serem realizados de outra forma por constrangimentos temporais, económicos, e de meios.

Definida a arquitectura do sistema SCAE, e adoptadas as tecnologias para o desenvolvimento do protótipo do sistema, a primeira tarefa, consiste na criação da Base de Conhecimento. Apesar do protótipo do sistema SCAE integrar apenas a funcionalidade de previsão da Tensão Crítica de vigas em I de inércia variável, atendendo ao facto de que um dos objectivos principais deste trabalho consiste na criação de modelos de previsão da Carga Crítica e da Tensão Crítica, na próxima secção encontra-se detalhado o processo e abordagem adoptada para a criação dos modelos de previsão enunciados anteriormente.

5.3 Modelos para a Previsão da Carga Crítica e da Tensão Crítica

Tal como referenciado anteriormente (Capítulo 2) uma Estrutura de Engenharia Civil deve ser projectada para suportar a carga a que vai ser submetida da forma mais segura, utilizando a menor quantidade de material. As estruturas de aço têm-se revelado cada vez mais eficientes e económicas, o que contribuiu para a sua utilização crescente. No entanto a procura da melhor solução requer a utilização de perfis mais leves e esbeltos e para alcançar esse objectivo é necessário conseguir uma boa calibração das fórmulas de previsão da carga crítica e da tensão crítica, por forma a evitar a utilização de coeficientes de segurança inadequados que podem ser responsáveis por desperdícios de material ou a rotura dos elementos estruturais.

As formulações propostas para a previsão da carga crítica de vigas sujeitas a cargas concentradas apresentam erros consideráveis, devido ao grande número de parâmetros que influenciam o comportamento de uma viga deste tipo e ao número reduzido de dados experimentais que permitam efectuar uma análise paramétrica completa. Quanto ao cálculo da tensão crítica de vigas em I de inércia variável, têm sido propostos vários modelos no entanto estes apresentam algumas lacunas que fazem com que se tornem bastante conservadores para o dimensionamento deste tipo de estruturas.

Como alternativa às formulações analíticas surgem os modelos de previsão baseados em algoritmos de Aprendizagem Automática, capazes de lidarem e resolverem problemas de complexidade elevada de difícil resolução através de métodos clássicos.

5.3.1 Materiais e Métodos

As Base de Dados de suporte ao desenvolvimento dos dois casos práticos, foram construídas a partir de dados experimentais, i.e., de experiências laboratoriais desenvolvidas por diferentes equipas de investigação no primeiro caso prático [Fonseca, 1999], e por Zárate no segundo [Zárate, 2001].

O processo de DCBD, para a geração dos modelos de previsão foi desenvolvido segundo a metodologia CRISP-DM (apresentada no capítulo 3). A ferramenta de DM escolhida para o desenvolvimento dos casos de estudo foi o *Clementine Data Mining System* da *SPSS Inc* (Anexo A). Este software está naturalmente alinhado com a metodologia aplicada, apesar de esta ser independente da plataforma tecnológica em que os processos de DCBD são desenvolvidos. Refira-se o facto do software pertencer a uma das organizações envolvidas no desenvolvimento da metodologia CRISP-DM.

5.3.2 Abordagem

Para a indução dos modelos de previsão foi seguida uma abordagem (Figura 5.6) tri-etápica: (i) análise dos dados (segundo a metodologia CRISP-DM), (ii) segmentação exploratória do conjunto de dados, e (iii) criação dos modelos de previsão. As fases que integram o processo permitem a identificação de padrões ou outros relacionamentos implícitos, existentes na BD analisada.

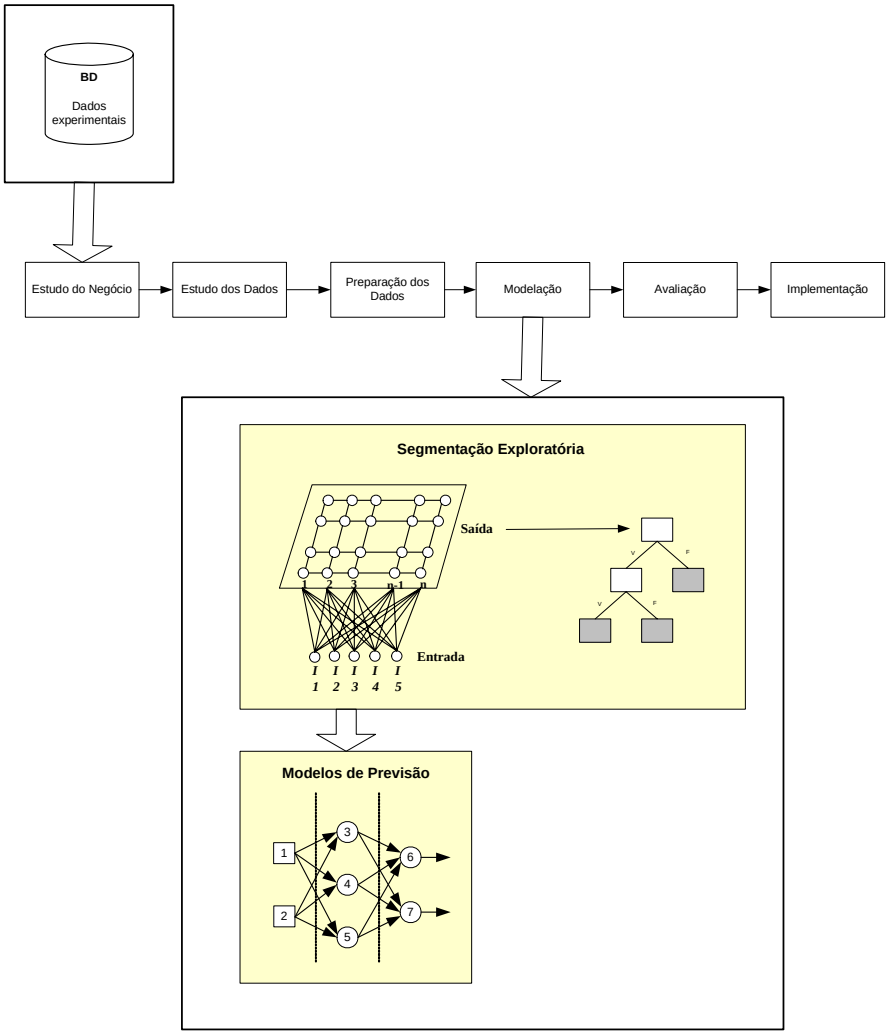


Figura 5.6 Abordagem para o processo de geração dos modelos de previsão.

Na abordagem proposta, a etapa de modelação comporta uma fase de exploração dos dados baseada em algoritmos de segmentação, que permitiram a identificação de segmentos homogéneos, para cada um dos quais foram gerados modelos de previsão. Isto permitiu que na maior parte dos segmentos, os modelos apresentassem um desempenho superior quando comparado com o desempenho de um modelo genérico. Por exemplo para o caso de estudo 1 - Modelos de Previsão da Carga Crítica em Vigas de Aço Sujeitas a Cargas Concentradas -, o erro máximo atingido pelo modelo genérico foi de 39.32% com um desvio padrão de 8.45%, o que

comparando com os resultados alcançados após a segmentação (Tabela 5.10) comprova a validade da abordagem proposta.

Este tipo de abordagem através de segmentação tem conhecido um interesse crescente em áreas como: medicina, Customer-Relationship Management (CRM). Por exemplo, em projectos de CRM, em que se torna imperioso seguir estratégias de marketing que foquem o esforço numa relação com o consumidor do tipo 1-para-1, uma abordagem de segmentação permite agrupar os clientes através do seu perfil.

A etapa de Implementação foi realizada através da integração do modelo de previsão da tensão crítica de vigas em I de inércia variável num Sistema de Conhecimento baseado em Data Mining para a Análise da Estabilidade de Estruturas de Engenharia Civil.

5.3.3 Modelos de Previsão

O primeiro caso de estudo teve por objectivo a indução de Modelos de Previsão da Carga Crítica em Vigas de Aço sujeitas a cargas concentradas, por forma a constituírem uma alternativa às formulações existentes, uma vez que estas apresentam ainda erros de precisão consideráveis (Capítulo 2). No segundo caso de estudo, o objectivo foi a indução de Modelos de Previsão da Tensão Crítica de Vigas com Perfil em I de Inércia Variável.

As técnicas de DM surgem como alternativa a explorar, uma vez que apresentam características que permitem o estudo de problemas complexos de difícil resolução através de abordagens mais convencionais, sendo por isso cada vez mais utilizadas nas diferentes áreas da engenharia.

Os casos de estudo foram desenvolvidos segundo a metodologia apresentada, estando a mesma reflectida na forma como é apresentada a descrição do desenvolvimento de cada um deles. Os resultados alcançados permitiram confirmar a viabilidade do desenvolvimento do

processo de DCBD na área da Engenharia Civil segundo a abordagem proposta, e validar a utilização de Redes Neuronais Artificiais na previsão da Carga Crítica e da Tensão Crítica.

1 Modelos de Previsão da Carga Crítica em Vigas de Aço Sujeitas a Cargas Concentradas

Compreensão dos Dados

A BD de suporte ao processo de geração dos modelos de previsão foi construída a partir de dados experimentais, recolhidos em estudos anteriores de previsão da carga crítica em vigas de aço sujeitas a cargas concentradas [Fonseca, 1999]. Deve realçar-se que a criação de novos dados em laboratório é um processo moroso e dispendioso, o que dificulta a obtenção de uma fórmula de projecto com um grau de precisão mais elevado, uma vez que a quantidade de dados não é suficiente para a validação de uma análise paramétrica completa.

Contudo, utilizando técnicas de Aprendizagem Automática (e.g., *Redes Neuronais*) mesmo com um número relativamente pequeno de casos, desde que esteja de alguma forma garantido que a amostra é representativa do universo a estudar, é possível induzir um padrão genérico. No entanto quanto mais significativa for a amostra melhor, embora em muitas situações, tal como este caso, a obtenção de resultados experimentais, torna-se difícil.

O conjunto de dados é composto por 161 registos, referentes a estudos anteriores, apresentando uma distribuição da *carga crítica experimental* $\in \{0\ldots4010\}$ *kN*, caracterizados pelos atributos constantes da Tabela 5.1.

Tabela 5.1 Atributos da BD.

Atributo	Descrição	Tipo
a	Distância entre os reforços	Inteiro
h	Altura da alma	Inteiro
t_w	Espessura da alma	Real
b_f	Largura do banzo	Inteiro
t_f	Espessura do Banzo	Real
c	Largura da carga – momento de inércia da secção	Inteiro
f_f	Tensão de cedência do banzo	Inteiro
f_w	Tensão de cedência da alma	Inteiro
pe	Carga Crítica Experimental	Real

Da análise deste conjunto de dados e distribuição das variáveis, foi possível verificar que 127 dos casos da amostra (Anexo B) eram referentes a vigas com carga crítica situada no intervalo, c*arga crítica* $\in \{0\ldots200\}$ *kN* (Figura 5.7), permitindo ainda identificar 6 casos da amostra que apresentavam valores anormais da carga crítica experimental (Tabela 5.2), o que motivou a sua eliminação.

Tabela 5.2 Casos eliminados da amostra.

a	h	t_w	b_f	t_f	c	f_w	f_f	pe
600	250	0.99	149	3.05	50	193	221	9.02
600	250	0.99	149	6.75	50	193	279	11.50
600	250	0.99	149	11.75	50	193	305	27.84
600	500	0.99	149	3.05	50	192	221	8.45
600	500	0.99	149	6.75	50	192	279	10.80
600	500	0.99	149	11.75	50	192	305	28.80

Recorreu-se ao gráfico *box-and-whisker plot*, mais frequentemente denominado por *box-plot*, para análise do atributo *carga crítica experimental*, uma vez que este sumaria de uma forma mais ou menos simples o conjunto de dados, expressando a distribuição de valores de uma variável. Este gráfico é também denominado de "caixa de bigodes", por se apresentar sob a forma de uma caixa cujos limites são o 1.º e o 3.º quartis.

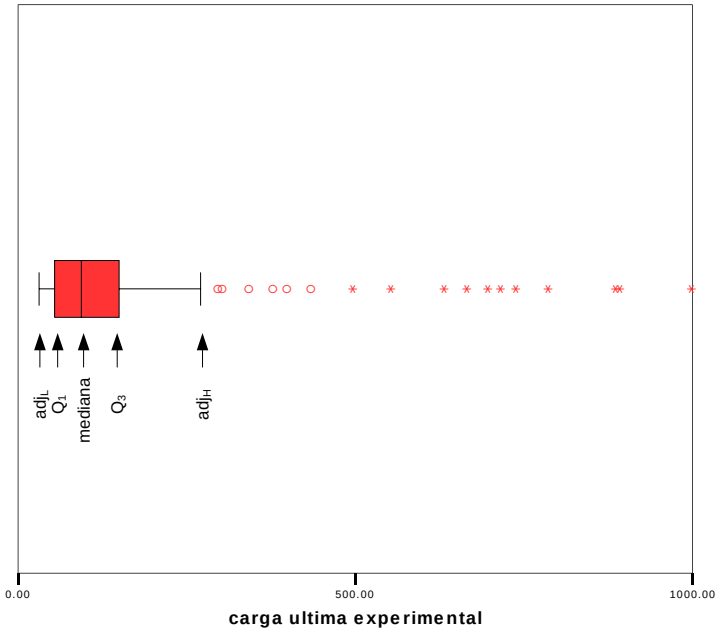


Figura 5.7 Gráfico *Box-Plot* do atributo carga crítica experimental.

Esta caixa é dividida por uma linha que representa a mediana, existindo também dois eixos, ou "bigodes", ligados à caixa que se estendem até ao valor mínimo e máximo dos dados, excluindo os valores discrepantes (*outliers*), e mostrando a forma como o conjunto de dados se espalha em torno da mediana, através de cinco conceitos estatísticos importantes:

- Q_1 (1.º Quartil): valor abaixo do qual estão 25% dos valores da variável;
- Q_3 (3.º Quartil): valor abaixo do qual estão 75% dos valores da variável;
- *IQ* (intervalo inter-quartil): intervalo onde estão 50% dos dados, i.e., $Q_3 - Q_1$:

- adj_L (adjacent Low): ponto adjacente inferior, isto é, o menor valor superior ou igual a $Q_1 – 1.5 . IQ$;
- adj_H (adjacent High): ponto adjacente superior, isto é, o maior valor menor ou igual a $Q_3 + 1.5 . Q1$.

Pela análise da Figura 5.7 é possível verificar que 79% dos dados apresentam *carga crítica <= 200 kN*. Tendo em consideração o reduzido número de casos disponíveis para um intervalo de valores tão grande (*carga crítica* $\in$ *{201,…,4010} kN*) foi tomada a decisão de centrar o desenvolvimento de modelos de previsão para vigas esbeltas (com *carga crítica* $\in$ *{0,…,100} kN*) e para vigas intermédias (com carga crítica situada no intervalo , *carga crítica* $\in$ *{100,…,200} kN*).

Foram desta forma considerados 127 registos com carga crítica máxima de 200 kN (Anexo B), com a estrutura apresentada na Tabela 5.3.

Tabela 5.3 Estrutura da Base de Dados.

teste	*A*	*h*	t_w	t_w^2	b_f	t_f	*C*	f_w	f_f	*a/h*	t_w/h	t_f/t_w	*c/h*	$c.t_w/a.h$	*prob*	*pe*
1	2400	700	3.26	10.63	150	6.10	0	326	347	3.4286	0.0047	1.8712	0	0	116.39	95.16
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
127	1100	800	3	9	250	12	120	215	268	1.3750	0.0038	4	0.150	261.82	108.45	91.50

Preparação dos Dados

De forma a incrementar a aprendizagem na geração dos modelos de previsão, uma vez que os modelos a utilizar na fase de modelação (RNA) não têm capacidade de correlação das entradas, foram introduzidos na BD (Tabela 5.3) os atributos resultantes da combinação de parâmetros geométricos (i.e., rácios) apresentados na Tabela 5.1: a/h, t_w/h, t_f/t_w, c/h, $c.t_w/a.h$, t_w^2, e o resultado da aplicação da Fórmula de Roberts (Fórmula 2.1), em utilização corrente na área da Engenharia Civil. Na Tabela 5.3, o atributo *prob* corresponde à carga crítica obtida pela

aplicação da Fórmula de Roberts, correspondendo o atributo *pe* à carga crítica experimental dos casos do conjunto de dados.

As estatísticas da Base de Dados após pré-processamento e transformação dos dados estão apresentadas na Tabela 5.4.

Tabela 5.4 Estatísticas dos atributos da BD.

Atributo	**Minimo**	**Máximo**	**Média**	**Desvio Padrão**
a	300	9800	1437.35	1529.83
h	250	1000	619.72	219,85
tw	1.96	4.01	2.67	0,68
bf	46	300	150.33	74,59
tf	3.05	30.50	10.89	5,50
c	0	200	65.67	48,75
fw	178.00	354.00	262.56	42,9
ff	221.00	347.00	275.36	24,91
a/h	0.75	14.00	2.40	2.37
tw/h	0.00	0.01	0,004	0,005
tf/tw	1.00	12.40	4.24	2.33
c/h	0.00	0.60	0.12	0.09
c tw / a h	0.00	300.00	108.52	78.40
Fórmula de Roberts (2.1)	38.34	193.53	89.80	42.98

Modelação

Na fase de modelação, a abordagem seguida contempla duas macro-tarefas: **(i)** segmentação e **(ii)** geração dos modelos de previsão.

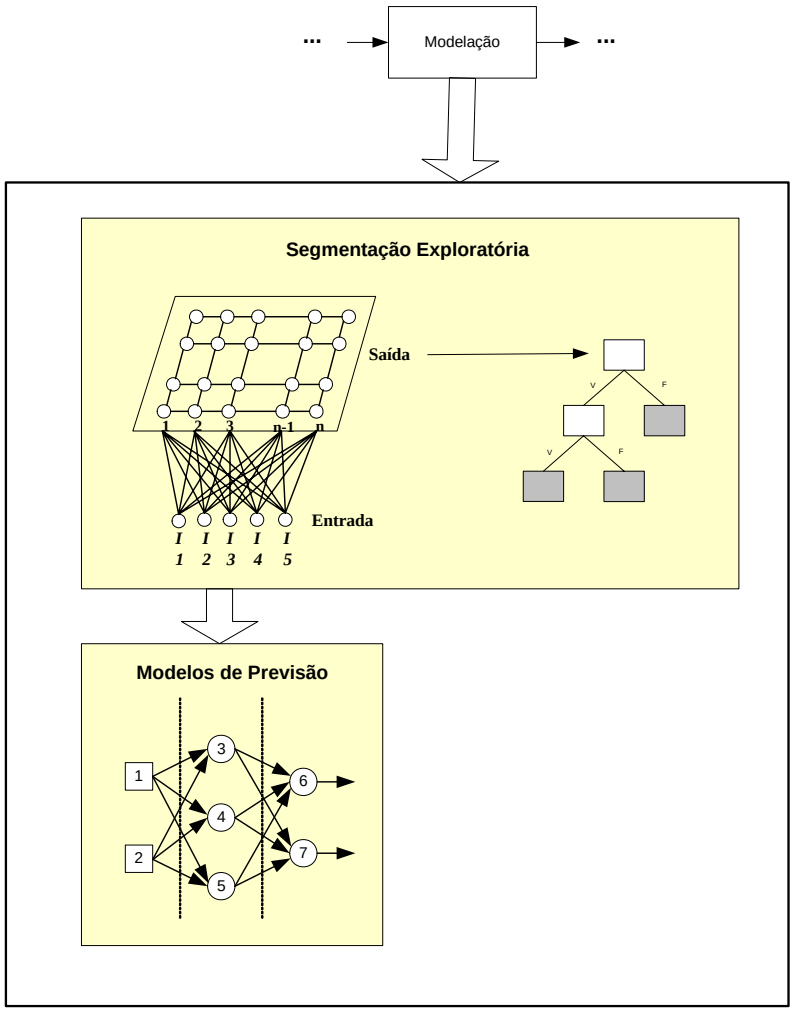


Figura 5.8 Procedimentos da Fase de Modelação.

Seguindo a abordagem especificada na secção 5.3.2, foram seleccionadas as seguintes técnicas de modelação (Figura 5.8):

1. Para o procedimento de segmentação foi utilizada uma ***rede de Kohonen*** que gera uma mapa bidimensional, no qual cada nodo tem associada uma determinada posição física na estrutura;
2. Para compreensão do modelo de classificação, foi realizada uma análise de grande granularidade aos dados classificados através da rede de Kohonen, utilizando o algoritmo **C5.0**, no sentido de gerar um modelo explicativo da segmentação;
3. Para geração dos modelos de previsão foi utilizada uma ***Rede Neuronal Artificial***, tendo a escolha por esta técnica sido determinada pelo facto de, quando a acuidade preditiva é um factor crítico, estes modelos apresentarem desempenhos superiores relativamente às Árvores de Regressão, por exemplo.

Segmentação

Tal como apresentado na Figura 5.6, o primeiro procedimento da fase de Modelação consistiu na realização de um trabalho de segmentação (objectivo intermédio) através da utilização de duas técnicas de DM, com métodos de aprendizagem diferentes. O primeiro, uma rede de *Kohonen*, baseado em aprendizagem não supervisionada para identificação de similaridades nos dados agrupando-os em grupos. O segundo, o algoritmo C5.0 para indução de uma árvore de decisão, que permitisse interpretar o modo como foi realizada a classificação em segmentos.

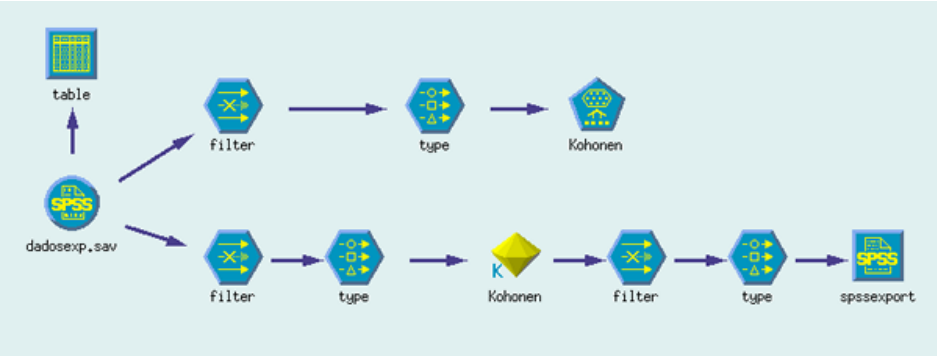


Figura 5.9 *Stream* de Segmentação utilizando uma *rede de Kohonen*.

A saída do modelo de segmentação Kohonen, consiste em dois atributos (*kx* e *ky*), correspondentes às coordenadas do mapa bidimensional. A concatenação destes atributos permite a definição de segmentos (também denominados *clusters*), e foi incluída na BD através do produto cartesiano.

A *stream* construída para a segmentação dos dados (Figura 5.9) apresenta um nodo *Filter* para selecção dos atributos a considerar para a geração do modelo de segmentação (rede de Kohonen); um nodo *Type* para configuração do tipo dos atributos de entrada; e o nodo *Kohonen* para geração do modelo com base no algoritmo de Kohonen. Após a geração do modelo, o mesmo é utilizado na parte inferior da *stream* (diamante) sendo o nodo *spssexport* utilizado para exportação do resultado (em formato SPSS) da segmentação (Tabela 5.5).

Tabela 5.5 Saída da rede de Kohonen.

tw	**…**	***kx***	***ky***	***cluster***
3.26	…	0	2	2
3.0	…	0	2	2
2.0	…	0	1	1
…	…	…	…	…
2.12	…	0	1	1

Na definição da topologia final da rede de Kohonen (Figura 5.10) foram experimentadas diversas configurações (i.e., número máximo de segmentos), no entanto em todas elas o modelo final resultante, identificou dois segmentos homogéneos com a distribuição de casos apresentada na Tabela 5.6.

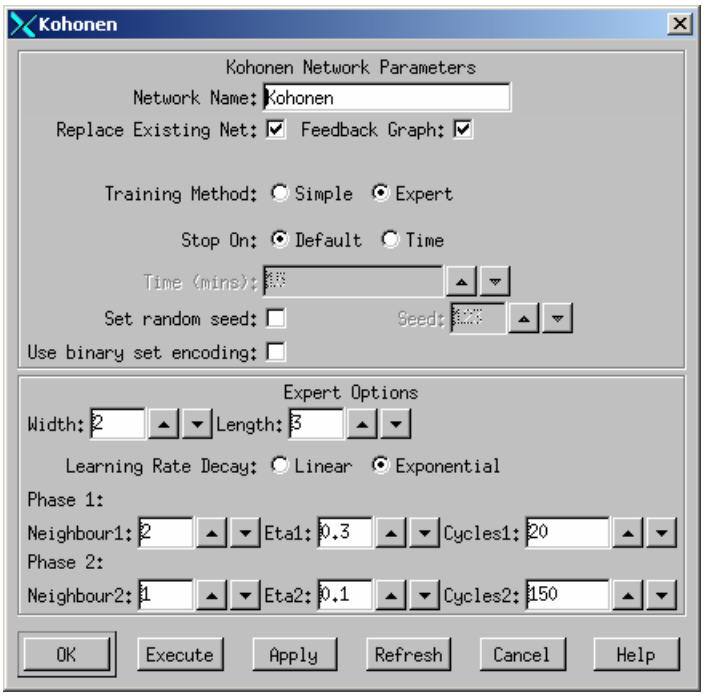


Figura 5.10 Parâmetros de configuração do nodo *Kohonen*.

O treino da rede de Kohonen é realizado em duas fases: tipicamente na primeira o valor da vizinhança é relativamente elevado tal como a taxa de aprendizagem, para aprendizagem global dos dados, enquanto na segunda é usada uma vizinhança mais pequena e um valor da taxa de aprendizagem mais baixa para ajustamento dos centros dos segmentos.

As distâncias na rede de Kohonen são calculadas como distâncias Euclidianas entre o vector de entrada e o centro do segmento para a unidade de saída.

Na rede de Kohonen, em contraste com outros tipos de RNA, a unidade de saída com o valor de activação mais baixo é a unidade vencedora.

A função de vizinhança baseia-se na distância de Chebychev, que considera a distância máxima em qualquer dimensão.

Tabela 5.6 Frequência de casos em cada segmento.

Cluster	N.º de casos	%
1	59	46.5%
2	68	53.5%
N	*127*	*100%*

Para a compreensão do modelo de classificação foi efectuada uma análise aos dados classificados, utilizando o algoritmo C5.0, através de uma *stream* (Figura 5.11) que tinha como *input* a saída da rede de Kohonen (Tabela 5.5) armazenada num ficheiro para persistência de dados, verificando-se a seguinte regra de segmentação:

Se t_w*<=2.12*
Então ***Cluster 1***
Senão ***Cluster 2****.*

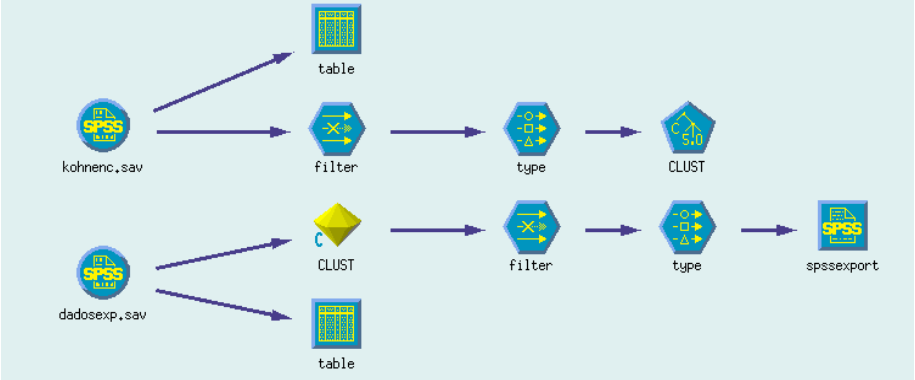


Figura 5.11 *Stream* regras de segmentação.

A segmentação permitiu identificar que um dos parâmetros decisivos na classificação das vigas em estudo é a espessura da alma (t_w). Estudos anteriores [Roberts, 1981] já haviam identificado que a carga crítica era proporcional ao quadrado da espessura da alma, e com menor intensidade, proporcional à espessura da mesa.

A avaliação do modelo de classificação foi realizada através da matriz de confusão (Tabela 5.7), verificando-se que todos os casos da amostra foram correctamente classificados.

Tabela 5.7 Matriz de confusão 2 x 2.

↓ Desejado\ Previsto →	Cluster 1	Cluster 2
Cluster 1	*TN = 59*	*FP = 0*
Cluster 2	*FN = 0*	*TP = 68*

Completada a fase de segmentação foram gerados dois modelos de previsão (um para cada segmento), tendo sido utilizados 60% dos dados para treino e 40% para validação. Dos dados utilizados para treino das Redes Neuronais de previsão, foram distribuídos 75% para treino e 25% para teste, de forma a prevenir o sobreajustamento (*overfitting*), uma vez que uma rede treinada em demasia perde capacidade de generalização.

O sobreajustamento acontece quando um modelo tem uma elevada capacidade de aprendizagem (devido ao número excessivo de unidades de processamento ou neurónios artificiais nas camadas intermédias) relativamente à complexidade inerente ao problema e/ou ao número de casos de treino. Quando tal se verifica, o modelo fixa em demasia os casos de treino e perde capacidade de generalização. Neste caso de estudo, com uma amostra de 127 registos, justifica-se utilizar casos de validação para impedir que a RNA perca capacidade de generalização [Sarle, 1995].

Antes da apresentação do conjunto de dados para treino da RNA foi necessário efectuar o escalonamento dos dados de forma a que os atributos com domínios de valores mais elevados

não assumissem preponderância relativamente a atributos com domínios de valores mais baixos. Os valores de todos os atributos foram escalonados no intervalo [0, 1].

Foram testadas diversas configurações para o treino da RNA combinando o método de treino com o número de épocas e com a topologia, sendo os parâmetros de treino os apresentados na Tabela 5.8. Para treino da RNA os pesos das conexões da rede foram inicializados de forma aleatória com valores no intervalo [-0.5, 0.5].

Tabela 5.8 Parâmetros do nodo *Train Net*.

Método de Treino	**% Dados de Treino**	**N.º de ciclos (acuidade)**
Multiple	75%	100%

O nodo *Train Net* do *Clementine* dispõe de 6 métodos de aprendizagem: *quick*, *dynamic*, *multiple*, *prune*, *exhaustive prune* e *RBF*. O método de treino seleccionado foi o *multiple* (Figura 5.12), uma vez que são criadas várias redes com diferentes topologias (o número exacto depende do conjunto de dados de treino), treinadas de forma paralela, sendo no final do treino, seleccionado o modelo que apresenta o erro RMSE mais baixo. Este método foi seleccionado uma vez que nos casos em estudo é crucial não cometer erros extremos, mesmo que possam existir erros pequenos e frequentes.

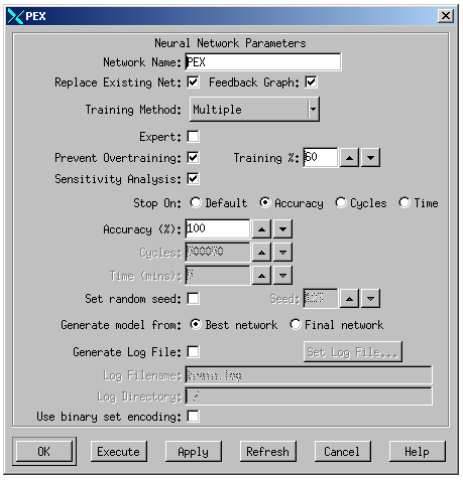


Figura 5.12 Parâmetros do nodo *Train Net*.

O método de treino *quick* consiste no treino de uma única rede neuronal, com uma camada escondida e que contendo $max(3,(n_i+n_o)/20)$ neurónios, onde n_i corresponde ao número de neurónios na camada de entrada e n_o ao número de neurónios na camada de saída.

Quando é seleccionado o modo *dynamic* a topologia da rede varia durante o processo de treino, através da introdução de novos neurónios até que a rede atinja a acuidade desejada. Este método envolve duas fases: selecção da topologia e treino na da rede segundo a topologia final.

O método de treino prune é conceptualmente o oposto do método *dynamic*. Em vez do processo se iniciar com uma rede com poucos neurónios e acrescentar durante o processo de treino, o método prune inicia com uma rede de grandes dimensões e efectua um processo de redução de neurónios nas camadas de entrada e nas escondidas. O método *exaustive prune* é um refinamento do método de treino *prune*.

As redes RBF (*radial basis function network*) são um caso especial de redes neuronais, que consistem numa arquitectura de três camadas: camada de entrada, camada escondida (também conhecida por camada de recepção) e camada de saída. As camadas de entrada e de saída são similares às redes do tipo Perceptrão Multicamada, contudo a camada escondida consiste em neurónios que representam segmentos de padrões de entrada.

Modelos de Previsão

Os modelos de previsão gerados são do tipo Rede Neuronal, tendo sido usado o nodo *Train Net* do Clementine. Tratam-se de redes do tipo *feedforward multicamada*, treinadas por *Back-Propagation*.

Para o *Cluster 1* foi gerada uma rede neuronal (modelo A) para previsão da carga crítica de vigas com espessura da alma inferior ou igual a 2.12 mm, com a seguinte arquitectura: 14 neurónios na camada de entrada, 19 na 1.ª camada escondida, 14 na 2.ª camada escondida e 1 na camada de saída. A acuidade máxima do modelo gerado é de 98.91% (Tabela 5.9).

Tabela 5.9 Modelos de Previsão.

Modelo	**Variáveis de entrada e importância relativa**		**Arquitectura**	**Acuidade Máxima**
A	f_w	0,04243	14-19-17-1	98,91%
	t_f	0,03922		
espessura	*a*/*h*	0,03126		
da alma	t_f / t_w	0,02321		
<=	f_f	0,01794		
	c/*h*	0,01686		
2.12	*a*	0,01406		
mm	*bf*	0,01096		
	t_w	0,00673		
	c	0,00555		
	h	0,00479		
	t_w /*h*	0,00424		
	t_w^2	0,00360		
	c * t_w / *a* * *h*	0,00040		
B	*Roberts*	0,12116	10-4-1	99,19%
	t_f	0,08078		
	f_w	0,08020		
espessura	t_w	0,06410		
da alma	*h*	0,04839		
>	f_f	0,03447		
	c	0,03260		
2.12	*bf*	0,01337		
mm	t_w^2	0,01156		
	a	0,01085		

O modelo gerado para o *Cluster 2* (modelo B), ajustado para vigas com espessura da alma superior a 2.12, é composto por 3 camadas, com 10 neurónios na camada de entrada, 4 na

camada escondida e 1 na camada de saída, atingindo uma acuidade máxima de 99.19% (Tabela 5.9).

A importância relativa das variáveis de entrada é resultado da Análise de Sensibilidade, sendo apresentadas por ordem decrescente de importância (0 indica um atributo que não influencia a previsão e 1 indica um atributo que influência de forma decisiva a previsão da saída da RNA). A análise de sensibilidade envolve a variação dos pesos das conexões e a observação do impacto dessa variação nos resultados de modo a determinar quais os atributos mais importantes, e aqueles que apresentam uma menor relevância para o resultado final.

Na tabela 5.9 é possível observar que a arquitectura do modelo de previsão A é mais complexa relativamente ao modelo B, facto que indicia uma maior dificuldade para a resolução do problema em estudo. Outro facto que aponta nesse sentido é a análise da importância relativa dos atributos de entrada, em que nenhum deles assume uma importância determinante. Esta observação é válida também para o modelo B, o que comprova que a Carga Crítica de vigas sujeitas a cargas concentradas depende de múltiplos factores, o que também está origem da dificuldade de obtenção de uma fórmula analítica mais precisa. No entanto existem dois atributos que assumem maior relevância nos dois modelos: a tensão de cedência da alma (f_w) e a espessura do banzo (t_f). No modelo B, a análise da importância relativa dos atributos de entrada do modelo, permite aferir que a inclusão do resultado da aplicação da Fórmula de Roberts na BD permitiu uma convergência nos resultados de previsão do modelo proposto, apesar do erro que esta fórmula de cálculo apresenta, verificando-se que as RNA apresentam características como robustez e degradação suave, pois foram capazes de processar informação com ruído de forma eficiente.

Avaliação dos modelos

De forma a validar os resultados alcançados e avaliar o grau de robustez dos modelos gerados, recorreu-se primeiro a uma validação na amostra de 40% dos dados usados, e posteriormente à validação cruzada (cross validation), método que se baseia na utilização de

todos os dados disponíveis. Este tipo de validação é mais adequado relativamente à validação baseada na divisão da amostra em conjunto de treino e de teste para pequenos conjuntos de exemplos de treino como é o caso. Os dados disponíveis considerados foram então divididos em 10 subconjuntos mutualmente exclusivos de comprimentos iguais, e a rede gerada foi testada 10 vezes, sendo aplicada em cada iteração a um dos subconjuntos de dados.

Para aferição dos resultados alcançados foram utilizadas as métricas MAD, SSE, MSE e RMSE de uso comum em problemas de previsão (ver capítulo 4).

Os resultados alcançados (Tabela 5.10) permitem concluir que a realização do trabalho prévio de segmentação dos dados, aumentou o grau de homogeneidade nos conjuntos de dados, verificando-se um incremento na acuidade preditiva e na generalização dos modelos de previsão.

Tabela 5.10 Avaliação de Resultados.

Cluster 1			Cluster 2		
Error	ANN	*Roberts*	Error	ANN	*Roberts*
MAD	0.81	1.63	MAD	1.17	1.70
SSE	372.50	704.84	SSE	884.61	3390.95
MSE	64.49	124.13	MSE	147.44	565.15
RMSE	6.78	10.21	RMSE	10.25	22.08

Quer a comparação seja efectuada através da medida MAD (i.e., na previsão é aceitável cometer alguns erros extremos desde que nos aproximemos na maior parte das vezes do valor real) ou da medida MSE (i.e., quando é crucial não cometer erros extremos, mesmo que possam existir erros pequenos e frequentes), os modelos induzidos apresentam melhores resultados na previsão da carga crítica de vigas esbeltas e intermédias relativamente à fórmula de cálculo (Fórmula 2.1).

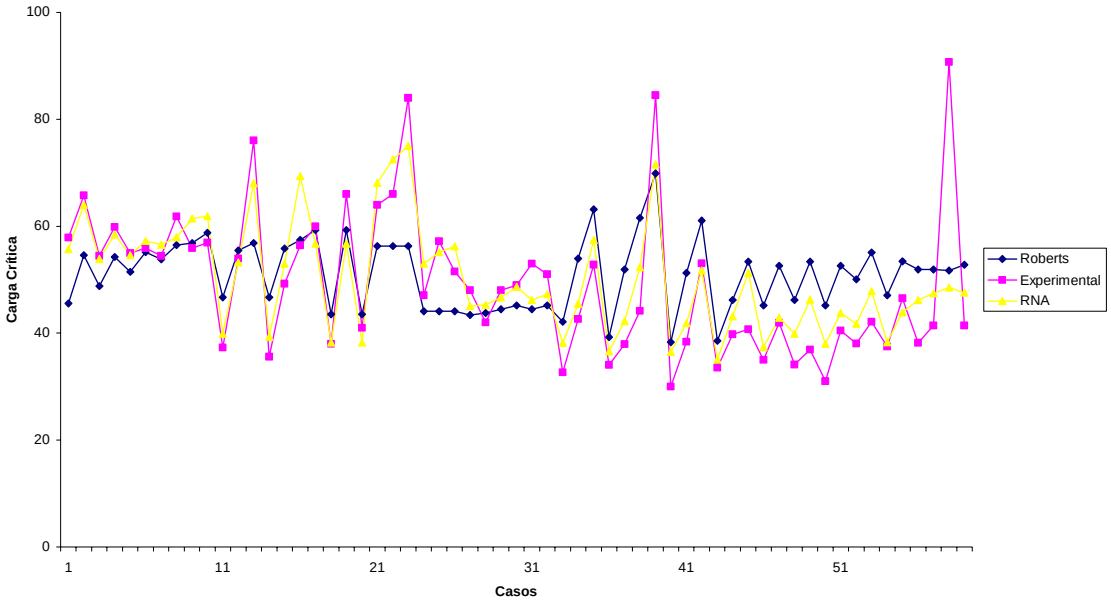


Figura 5.13 Comparação entre os valores da Carga Crítica Experimental, e a Prevista através da aplicação da Fórmula de Roberts e da aplicação do modelo de previsão A.

O modelo de previsão A (Figura 5.13), apresenta um bom ajustamento da variável *Carga Crítica* verificando-se contudo uma diminuição da capacidade preditiva nas vigas com carga crítica superior a 80 kN. Nestes casos o valor previsto pelo modelo para o parâmetro de saída aproxima-se do que resulta da aplicação da fórmula de cálculo existente (Fórmula 2.1).

Relativamente ao modelo de previsão B (Figura 5.14), este apresenta, para todos os casos da amostra, uma boa capacidade preditiva relativamente ao valor experimental do parâmetro de saída.

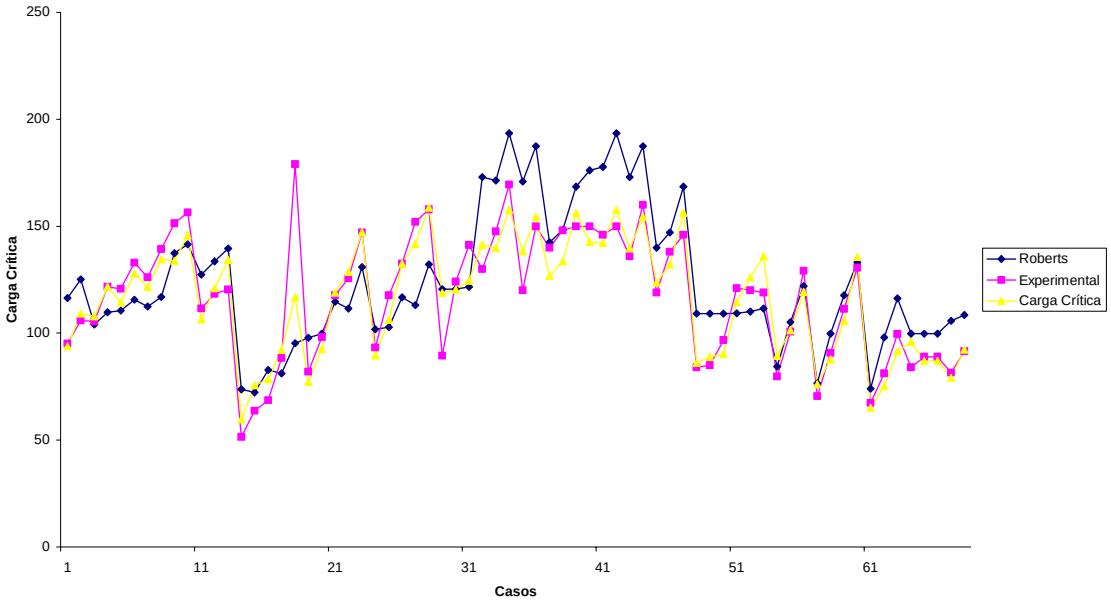


Figura 5.14 Comparação entre os valores da Carga Crítica Experimental e a Prevista através da aplicação da Fórmula de Roberts e da aplicação do modelo de previsão B.

2 Modelos de Previsão da Tensão Crítica de Vigas com Perfil em I de Inércia Variável

Estudo dos Dados

Para o desenvolvimento do processo de DCBD, foram utilizados 179 exemplos de vigas esbeltas de inércia variável (Anexo C), estudadas por Zárate. Estes exemplos cobrem uma ampla gama de coeficientes de forma (relação entre a distância dos reforços transversais e a altura da alma), da esbeltez da alma e da pendente do banzo inferior.

Os 179 registos considerados apresentam uma distribuição da tensão crítica que varia entre 24.97 e 130.34 (Tabela 5.12), com os atributos constantes da tabela 5.11, referentes a parâmetros geométricos e a parâmetros geométricos combinados (α e λ_f). Na Figura 5.15 podem ser visualizadas as distribuições de alguns dos atributos da BD.

Tabela 5.11 Atributos da BD.

Variável	Descrição	Tipo
α	Relação entre a distância dos reforços transversais e a altura da alma (a/h_1)	Real
λ_f	Esbeltez	Real
η	Relação largura do banzo / altura da alma (b_f/h_1)	Real
$\mathrm{tg}\,\phi$	Pendente do banzo inferior	Real
t_w	Espessura da alma	Real
h_0	Altura menor da alma	Real
k_{mod}	Coeficiente de enfunamento de vigas metálicas de inércia variável obtidos através do Modelo de Galván	Real
k_{mef}	Coeficiente de enfunamento de vigas metálicas de inércia variável obtidos através do Método de Elementos Finitos	Real
$\tau_{cr,\mathrm{mod}}$	Tensão crítica Modelo de *Galván*	Real
$\tau_{cr,\mathrm{mef}}$	Tensão crítica Modelo Elementos Finitos	Real

Tabela 5.12 Estatísticas dos atributos da BD.

Variável	Mínimo	Máximo	Média	Desvio Padrão
α	0.50	4.00	1.17	0.58
λ_f	12.50	60.00	33.33	13.35
η	0.20	0.45	0.3184	0.09
$\mathrm{tg}\,\phi$	0.00	0.60	0.31	0.17
t_w	7.15	8.00	7.99	0.06
h_0	551	1976	1334.94	377.78
k_{mod}	2.71	25.32	8.37	4.36
k_{mef}	2.43	25.52	8.41	4.33
$\tau_{cr,\mathrm{mod}}$	24.57	145.17	60.37	22.38
$\tau_{cr,\mathrm{mef}}$	24.97	130.34	60.54	21.80

Preparação dos Dados

Neste caso de estudo não se verificou a necessidade de inclusão de novos atributos, nem se registaram casos com valores estranhos em nenhum dos atributos considerados. A principal razão para a qualidade patente nos dados utilizados advém do facto de terem sido gerados por uma equipa de investigação em ambiente laboratorial, sendo neste caso a taxa de erros nos dados nula ou muito reduzida.

Na Figura 5.15 estão representados os histogramas de quatro dos atributos considerados no estudo, podendo verificar-se que a relação entre a distância dos reforços transversais e a altura da alma (α) se apresenta enviesada enquanto que por exemplo o pendente do banzo inferior ($\mathrm{tg}\,\phi$) apresenta uma distribuição aproximadamente normal. Este tipo de análise é de grande relevância uma vez que a existência de atributos que apresentem distribuições com tendências tem impacto nos resultados dos modelos gerados [Santos et al, 2004].

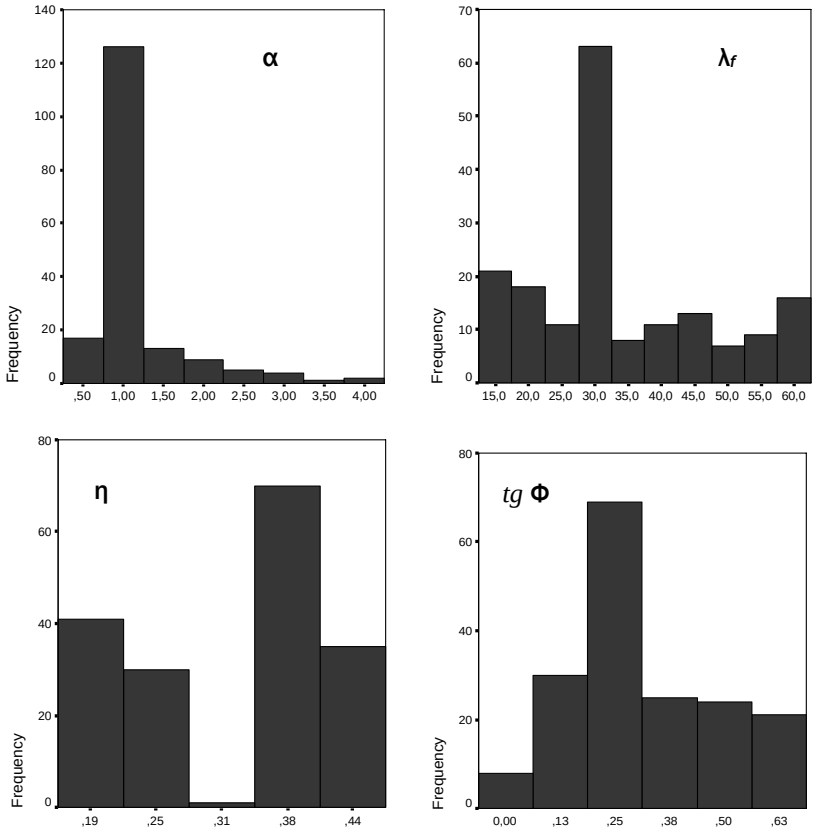


Figura 5.15 Estatísticas dos atributos da BD.

Modelação

Seguindo a abordagem especificada na secção 5.3.2, foram identificados dois segmentos homogéneos nos dados utilizados, segundo a distribuição de casos apresentada na tabela 5.13.

Tabela 5.13 Frequência de casos em cada segmento.

Cluster	N.º de casos	%
1	105	59.7%
2	71	40.3%
N	*176*	*100%*

Na definição da topologia final da rede de Kohonen foram experimentadas diversas configurações (i.e., número máximo de segmentos), no entanto em todas elas o modelo final resultante, identificou dois segmentos homogéneos com a distribuição de casos apresentada na Tabela 5.13.

Para a compreensão do modelo de classificação foi efectuada uma micro-análise aos dados classificados, utilizando o algoritmo C5.0, verificando-se a seguinte regra de segmentação:

Se h_0*<=1400*
Então ***Cluster 1***
Senão ***Cluster 2****.*

A segmentação permitiu identificar que o parâmetro decisivo na classificação das vigas em estudo (i.e., vigas em I de inércia variável) é a altura menor da alma (h_0).

A avaliação do modelo de classificação foi realizada através da matriz de confusão (Tabela 5.14), tendo-se verificado que foram incluídas duas vigas com valor de h_0 >1400 mm no Cluster 1 e uma viga foi classificada no segmento 2 com um valor de h_0 <= 1400 mm.

Tabela 5.14 Matriz de confusão 2 x 2.

↓ Desejado \ Previsto →	Cluster 1	Cluster 2
Cluster 1	*TN = 103*	*FP = 1*
Cluster 2	*FN = 2*	*TP = 70*

A partir da matriz de confusão foram calculadas as métricas de avaliação (Capítulo 4) que são apresentadas na Tabela 5.15.

Tabela 5.15 Métricas de Avaliação.

Sensibilidade	**Especificidade**	**Precisão**
97.22	99.04	98.30

O erro de tipo I (especificidade) ocorre quando o modelo deveria rejeitar um caso e não o rejeita, e o erro de tipo II (sensibilidade) ocorre quando o modelo rejeita um caso que não deveria rejeitar. Em diversas situações (e.g., previsão de cancro) o custo de ter um falso negativo é superior a ter um falso positivo.

Neste caso de estudo verifica-se que três dos casos considerados não foram correctamente classificados, apresentando a sensibilidade um valor menor relativamente à especificidade e previsão, uma vez que o modelo obtém uma melhor performance para os casos do Cluster 1.

Modelos de Previsão

Tal como no caso de estudo anterior os modelos de previsão são do tipo Redes Neuronais *Feedforward Multicamada*, treinadas por *Back-Propagation*.

Foram gerados dois modelos de previsão (i.e., um para cada segmento), tendo sido utilizados 75% dos dados para treino e 25% para validação. Dos dados utilizados para treino das Redes Neuronais de previsão, foram distribuídos 70% para treino e 30% para teste, de forma a prevenir o sobreajustamento (*overfitting*).

A *stream* (Figura 5.16) construída para a geração dos modelos de previsão da Tensão Crítica com base em RNA apresenta os seguintes nodos: um nodo *Select* para selecção dos dados do cluster 1 e seguidamente do cluster 2; um nodo *Filter* para selecção dos atributos relevantes; um nodo *Type* para selecção dos tipos dos atributos; um nodo *Sample* para divisão do conjunto de dados, em conjunto de teste e conjunto de treino; um nodo *Train Net* para geração da RNA (com os parâmetros apresentados na Tabela 5.16).

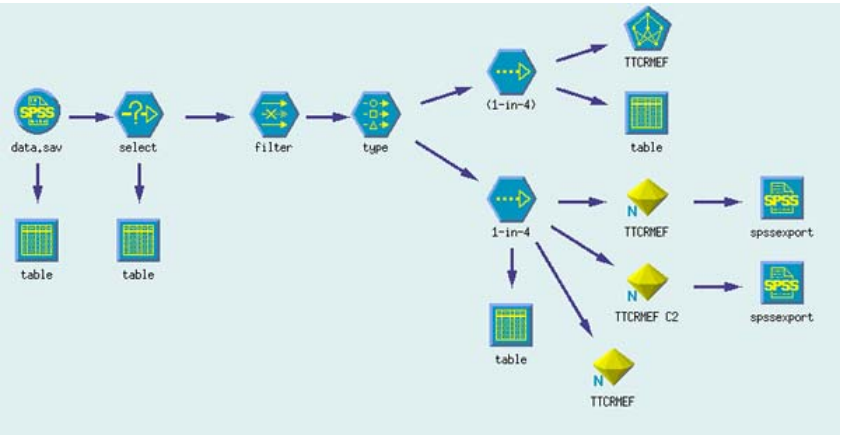


Figura 5.16 *Stream* de Geração dos Modelos de Previsão da Tensão Crítica.

Tabela 5.16 Parâmetros de treino do Modelo de Previsão da Tensão Crítica.

Método de Treino	% Dados de Treino	N.º de ciclos	Camadas Escondidas
Exhaustive Prune	70%	Default	2

Os dois modelos apresentam características comuns: 7 neurónios na camada de entrada, 4 na camada escondida e 1 na camada de saída (Tabela 5.17). Para a maior parte dos problemas, uma arquitectura de 3 camadas é suficiente para a sua resolução [Cortez, 2002].

Tabela 5.17 Modelos de Previsão.

Modelo	Variáveis de entrada e importância relativa		Arquitectura	Acuidade Máxima
A	h_0	0.93957	7-4-1	98.98%
altura	k_{MEF}	0.67376		
menor	t_w	0.14969		
da alma	$tg\ \Phi$	0.08220		
	α	0.06613		
<=	λ_f	0.04282		
1 400	η	0.04149		
B	k_{MEF}	0.78930	7-4-1	99.58%
altura	h_0	0.67887		
menor	$tg\ \Phi$	0.14969		
da alma	α	0.11338		
	t_w	0.11335		
>	η	0.01122		
1 400	λ_f	0.00983		

Da análise de sensibilidade realizada é possível verificar que os atributos: altura menor da alma, e coeficiente de enfunamento são os mais relevantes na determinação da tensão crítica de vigas em I de inércia variável, verificando-se que os atributos considerados relevantes para a previsão nos dois segmentos de vigas são os mesmos, verificando apenas diferenças no posicionamento da importância relativa.

Avaliação

Para avaliação dos modelos gerados foram calculados valores correspondentes às métricas típicas de casos de regressão apresentadas no capítulo 4.

Tabela 5.18 Avaliação de Resultados.

Cluster 1			Cluster 2		
Error	ANN	Zárate	Error	ANN	Zárate
MAD	0,658846	0,87	MAD	0,223529	0,566471
SSE	38,0461	22,7292	SSE	1,9988	13,8585
MSE	1,415938	0,8742	MSE	0,11757647	0,815206
RMSE	1,189932	0,934987	RMSE	0,34289426	0,902888

Os resultados alcançados (Tabela 5.18) permitem concluir que o modelo de previsão para o Cluster 2 tem um desempenho superior ao modelo analítico proposto por Zárate, enquanto que para o Cluster 1 os resultados são similares.

Se efectuarmos a comparação entre as duas abordagens através da medida MAD (i.e. na previsão é aceitável cometer alguns erros extremos desde que nos aproximemos na maior parte das vezes do valor real) os modelos de previsão com base em RNA apresentam melhores resultados, quer para o Cluster 1, quer para o Cluster 2.

Quando se considera a medida MSE (i.e., quando é crucial não cometer erros extremos, mesmo que possam existir erros pequenos e frequentes) para o Cluster 1 o modelo proposto por Zárate apresenta um desempenho superior, enquanto para o Cluster 2 o modelo de previsão gerado neste trabalho apresenta uma melhor acuidade preditiva.

5.4 Interface

O interface do sistema SCAE com o utilizador (Figura 5.17) consiste num painel de controlo para criação de estruturas com diferentes configurações, limitada neste protótipo a vigas em I de inércia variável, para previsão da Tensão Crítica.

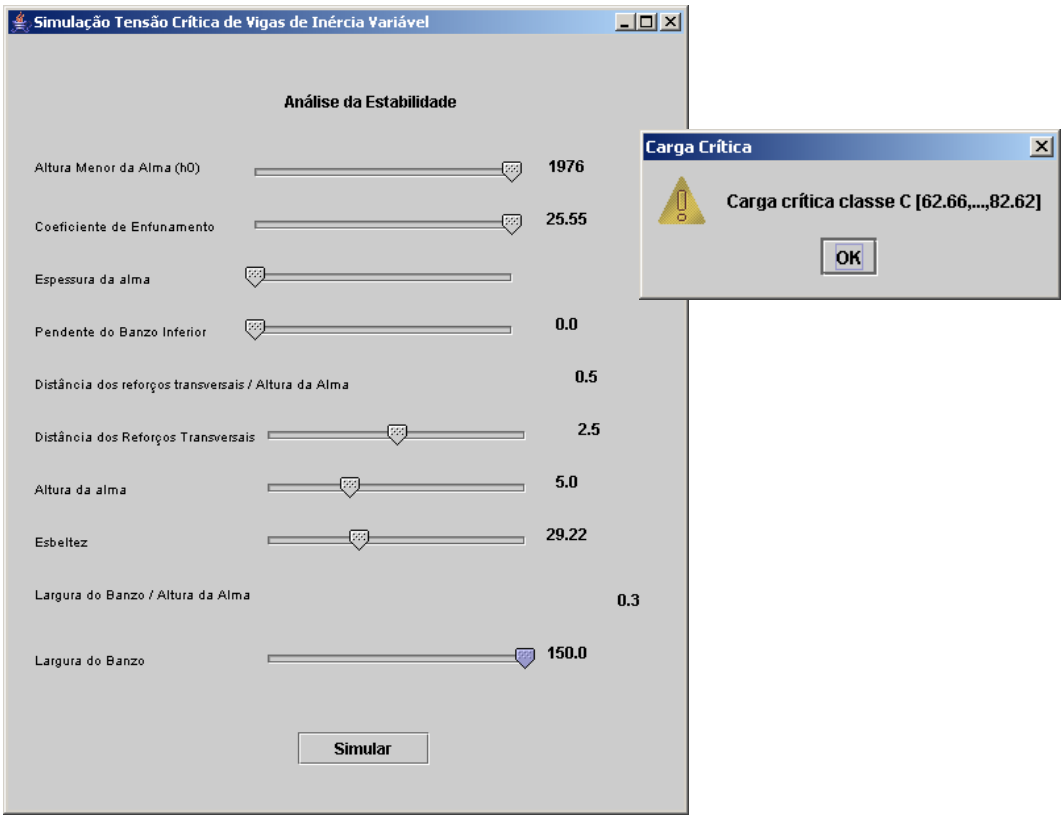


Figura 5.17 Interface do Sistema SCAE.

5.5 Integração de Componentes

Na fase de integração dos componentes do sistema SCAE verificou-se que a especificação PMML dos modelos gerados no Clementine Data Mining System, e a especificação PMML processados pelo Xelopes apresentavam diferenças, exigindo um esforço de uniformização.

Em virtude deste facto optou-se por incorporar neste primeiro protótipo do sistema SCAE um modelo de previsão gerado no próprio sistema através das funcionalidades da biblioteca Xelopes. Foi contudo necessário transformar a saída em classes, criando intervalos de tensão crítica (Tabela 5.19), em função de uma restrição da biblioteca que não dispõe ainda das funcionalidades que permitam efectuar a transformação das saídas de um modelo do tipo Rede Neuronal para previsão de um valor contínuo. Em face desta limitação, o atributo tensão crítica foi discretizado em 4 classes (A, B, C, D, E).

Tabela 5.19 Classes de Tensão Crítica de Vigas de Inércia Variável.

Classe	Tensão Crítica
A	[24.97,…,42.85[
B	[42.85,…,62.66[
C	[62.66,…,82.62[
D	[82.62,…,117.71]

O modelo de classificação gerado no sistema com base numa RNA permitiu a classificação tensão crítica com uma acuidade de 100%. Todos os casos do conjunto de teste foram correctamente classificados na classe de tensão crítica correspondente.

5.6 Resultados e Discussão

A criação do protótipo do sistema SCAE seguiu o seguinte método de desenvolvimento: definição dos requisitos do sistema; definição da arquitectura; desenvolvimento da base de conhecimento (i.e., geração dos modelos de previsão); criação do interface com o utilizador; integração dos componentes do sistema. A arquitectura do sistema assenta na estrutura típica de um Sistema de Conhecimento: Núcleo do Sistema Baseado em Conhecimento ou Shell (desenvolvido com recurso à biblioteca Xelopes); Base de Conhecimento (modelo de previsão do tipo RNA especificado em PMML); Base de Dados (em formato ARFF); Interface com o utilizador (painel de controlo para criação de estruturas com diferentes configurações).

O desenvolvimento deste protótipo de um Sistema de Conhecimento permitiu testar a incorporação de funcionalidades de DM num Sistema Inteligente para Análise da Estabilidade de Estruturas de Engenharia Civil. Esta incorporação foi possibilitada pela utilização da biblioteca Xelopes que proporciona um modelo que permite a integração de componentes de Data Mining em aplicações de software, independente de plataformas e fontes de dados, o que proporciona a utilização de modelos de DM desenvolvidos numa outra plataforma (e.g., Clementine) através da especificação dos modelos em PMML.

Os resultados alcançados permitiram validar: (i) a abordagem proposta, e (ii) a utilização de modelos de Aprendizagem Automática para a resolução de problemas complexos na área de Engenharia Civil.

A abordagem proposta integra na fase de modelação duas tarefas: (i) segmentação dos dados, e (ii) geração dos modelos de previsão. Através desta abordagem é possível incrementar o grau de homogeneidade dos casos de cada segmento, o que se traduz na geração de modelos de previsão com um grau de ajuste mais elevado. Tendo em consideração que na área em estudo a acuidade preditiva é um factor crítico, este procedimento revelou-se extremamente útil.

Nos dois casos de estudos a geração dos modelos de previsão baseados em Redes Neuronais Artificiais através de um processo de Descoberta de Conhecimento em Base de Dados, envolveu a criação de um projecto com 3 streams no Clementine Data Mining System com os seguintes objectivos: (i) segmentação dos dados com utilização de rede de Kohonen; (2) compreensão do modelo de classificação com utilização do algoritmo C5.0 para geração das

regras de classificação seguidas no ponto (i); e (iii) geração dos modelos de previsão (um para cada segmento identificado) com base em RNA.

Para a previsão da carga crítica de vigas de aço sujeitas a cargas concentradas os modelos desenvolvidos apresentam índices de confiança superiores às actuais formulações analíticas. A realização de uma etapa intermédia de segmentação, permitiu identificar dois segmentos homogéneos no conjunto de dados disponível, com uma regra de decisão com apenas uma condição, sendo o atributo da condição a espessura da alma, o que está de acordo com o estudos de Roberts que confirmaram que a carga última é aproximadamente proporcional ao quadrado da espessura da alma.

Relativamente à previsão da Tensão Crítica de vigas em I de inércia variável, os modelos propostos apresentam para um dos segmentos um desempenho superior à formulação analítica proposta por Zárate, enquanto que no outro segmento os resultados dos dois modelos (RNA e Zárate) são semelhantes. No entanto se os resultados forem analisados através da medida Mean Absolute Deviation, privilegiando uma maior quantidade de acertos relativamente à não existência de erros extremos, os resultados obtidos pelos modelos propostos neste trabalho (Cluster 1 = 0.658846, Cluster 2 = 0.223529) são melhores do que os resultantes da aplicação do modelo analítico de Zárate (Cluster 1 = 0.223529, Cluster 2 = 0.566471).

Estes resultados abrem portas ao desenvolvimento de Sistemas Inteligentes para Análise da Estabilidade de Estruturas de Engenharia Civil para monitorização e planeamento, embebendo os modelos desenvolvidos para previsão da carga crítica de vigas sujeitas a cargas concentradas e da tensão crítica de vigas em I de inércia variável.

Capítulo 6

Conclusões e **T**rabalho **F**uturo

São apresentas as conclusões do trabalho desenvolvido, identificando-se as principais contribuições para as áreas de Tecnologias e Sistemas de Informação e de Engenharia Civil, sendo lançadas algumas linhas orientadoras para trabalho a desenvolver futuramente.

6.1 Conclusões

As técnicas de DM foram aplicadas neste trabalho para a previsão da Carga Crítica de vigas em aço sujeitas a cargas concentradas e a Tensão Crítica de vigas em I de inércia variável, mais ajustados comparativamente às formulações analíticas mais clássicas em utilização quotidiana.

Os resultados alcançados, nos dois casos de estudo desenvolvidos, comprovam a viabilidade do desenvolvimento do processo de DCBD na área de Engenharia Civil, seguindo uma abordagem que priviligia uma etapa intermédia de Segmentação com utilização de um mapa de características auto organizáveis (i.e., Kohonen), para identificação de segmentos homogéneos de dados, e a validade da utilização de Redes Neuronais Artificiais na previsão da Carga Crítica de vigas de aço sujeitas a cargas concentradas (com carga crítica inferior) e da

Tensão Crítica de vigas em I de inércia variável, uma vez que a acuidade dos modelos gerados é superior à conseguida pelas actuais fórmulas de cálculo.

Em alternativa aos métodos convencionais surgem os modelos de DM. Neste trabalho foram considerados dois tipos de modelos (descritos em pormenor no Capítulo 4): RNA e Árvores de Decisão, tendo relativamente às RNA sido utilizadas dois tipos: redes de Kohonen, e redes feedforward multicamada (também conhecidas como Multilayer Perceptron).

No decorrer do processo de DCBD, foram encontrados alguns obstáculos, relacionados com o grande número de parâmetros que influenciam o comportamento das estruturas estudadas e o número reduzido de casos.

Contudo, mesmo com um número relativamente pequeno de casos, desde que esteja de alguma forma garantida que a amostra é representativa do universo a estudar, é possível que uma RNA consiga induzir um padrão genérico.

Relativamente a esta problemática e relacionado com o segundo caso de estudo, futuramente poderá ser possível a geração de novos casos experimentais através da elaboração de um trabalho integrado de geração de casos com especialistas da área de Engenharia Civil e da área dos Sistemas de Informação.

Na fase de Descoberta de Conhecimento em Base de Dados, e dada a dificuldade em criar um modelo de previsão único para cada um dos problemas em estudo, foi elaborada e implementada uma abordagem que incorpora duas tarefas na fase de Modelação: (i) segmentação e (ii) geração dos modelos de previsão para cada um dos segmentos identificados em (i).

A redução conseguida relativamente ao erro de previsão foi obtida através de:

- etapa intermédia de segmentação na abordagem apresentada, utilizando um modelo inspirado no algoritmo de Kohonen;
- redução do erro global e ajustamento do período de aprendizagem;
- introdução de parâmetros combinados;
- método de treino.

Os modelos gerados com base em redes neuronais, provaram ser mais eficazes que os métodos convencionais (apenas no segundo caso de estudo, para um dos segmentos atingiram um desempenho semelhante), comprovando-se que com implementações cuidadas, seguindo a metodologia e abordagem apresentada, as técnicas de DM e AA podem contribuir para a resolução de problemas complexos na área da Engenharia Civil.

Face aos resultados alcançados foi desenvolvida uma primeira versão de um Sistema de Conhecimento baseado em Data Mining para Análise da Estabilidade de Estruturas de Engenharia Civil (SCAE), embebendo no mesmo modelos de previsão baseados em RNA.

Os resultados obtidos nos testes pelo protótipo desenvolvido permitiram estabelecer uma base de trabalho para o desenvolvimento futuro de um Sistema Inteligente de monitorização em tempo real de Estruturas de Engenharia Civil, recorrendo às tecnologias utilizadas. Este trabalho permitiu também identificar pontos críticos tais como: necessidade de uniformização dos documentos PMML gerados pelo Clementine e processados pela biblioteca Xelopes e, necessidade de desenvolvimento de uma classe JAVA responsável pelo processamento da saída de um modelo do tipo Rede Neuronal num problema de regressão para integração na biblioteca Xelopes.

6.2 Contribuições

Os contributos deste trabalho apresentados anteriormente, podem ser agrupados em duas áreas: as Tecnologias e Sistemas de Informação (TSI) e a Engenharia Civil.

Relativamente à área de TSI podem elencar-se os seguintes contributos: (i) aplicação de técnicas de DM a uma nova área (Análise da Estabilidade de Estruturas Metálicas de Engenharia Civil); (ii) definição de uma abordagem para o processo de Descoberta de Conhecimento em Bases de Dados, que se traduz na incorporação de duas tarefas na fase de Modelação; (iii) concepção do SCAE – Sistema de Conhecimento baseado em Data Mining para a Análise da Estabilidade de Estruturas Metálicas, com incorporação de técnicas automáticas para aquisição de conhecimento.

Para a área de aplicação da Engenharia Civil os contributos podem ser enumerados da seguinte forma: (i) geração de modelos de previsão da Carga Crítica e da Tensão Crítica com acuidades superiores às formulações actualmente existentes; e (ii) protótipo do sistema SCAE.

O sistema SCAE coloca à disposição dos especialistas de Engenharia Civil um Ambiente Virtual de Simulação para a parametrização de uma estrutura metálica, permitindo obter resultados de forma mais rápida e com acuidade superior relativamente às formulações actuais. Através deste ambiente de simulação é possível gerar dados e apoiar o estudo na obtenção de fórmulas de cálculo mais ajustadas que permitam reduzir a utilização coeficientes de segurança inadequados.

O protótipo do Sistema SCAE desenvolvido neste trabalho abre portas ao desenvolvimento de um sistema de monitorização da segurança de estruturas em tempo real.

6.3 Trabalho Futuro

A área da manutenção de estruturas e obras de arte de Engenharia Civil assume aspectos cada vez mais importantes, permitindo evitar a sua deterioração e potenciando uma intervenção pró-activa, para prevenir eventuais desastres como o ocorrido com a Ponte de Entre-os-Rios em Castelo de Paiva.

O estudo do ciclo de vida útil de qualquer estrutura de Engenharia Civil é um tema de crescente actualidade. Com efeito, o número de estruturas é tão grande que as administrações necessitam de investir cada vez mais dinheiro na inspecção, manutenção, reparação ou reabilitação das mesmas.

Os custos estimados de reparação ou substituição são cada vez mais importantes e serão um pesado fardo para as economias das futuras gerações dos distintos países. Por exemplo, a Federal Highway Administration dos Estados unidos, estima que esse custo atinja 50.000 milhões de dólares. Apesar da grandeza do problema, muitas das decisões relacionadas com este tema são, ainda, tomadas tendo por base o dia a dia, sob uma enorme pressão de colocar as estruturas em serviço o mais rapidamente possível e pelo menor custo. Esta situação impede

uma análise a mais longo prazo e que os escassos recursos disponíveis sejam investidos onde são mais necessários.

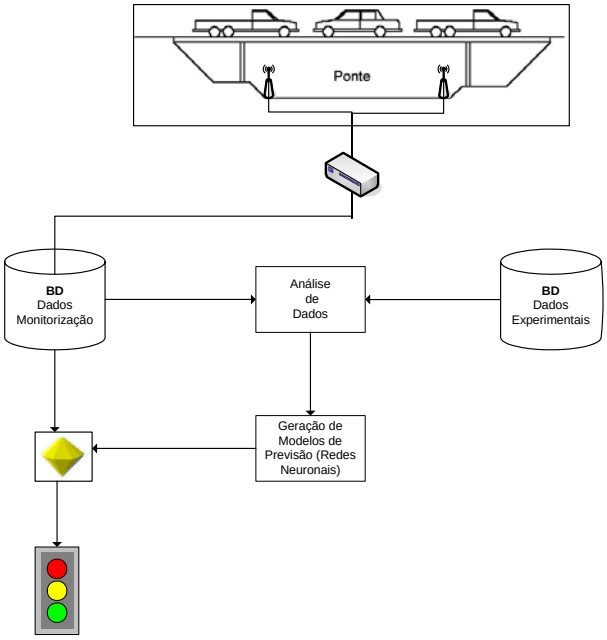


Figura 6.1 Arquitectura Protótipo para um sistema de monitorização em tempo real e detecção de necessidade de manutenção.

Verificada a acuidade preditiva dos modelos com base em técnicas de DM e a possibilidade de desenvolvimento de um Sistema de Conhecimento para Análise da Estabilidade de Estrutura de Engenharia Civil, pretende-se o desenvolvimento de um sistema de monitorização em tempo real das estruturas, que permita detectar sinais de alarme, através do aproveitamento dos sensores de recolha de dados instalados, implementando um sistema que inclua a análise e tomada de decisão para manutenção (Figura 6.1).

Na próxima versão do sistema SCAE deverão ser implementadas funcionalidade tais como: submissão de um conjunto de cenários criados individualmente a partir da interface do

sistema mas não submetidos de forma individual ao modelo de previsão, devendo ser armazenados os cenários e os resultados num ficheiro que permita um estudo posterior mais detalhado; importação de modelos de previsão desenvolvidos em outras plataformas; utilização do Sistema SCAE para análise em tempo real de estruturas em funcionamento através da captura de dados por sensores de medição remota; disponibilização do Sistema SCAE sob a forma de Web Service.

O Sistema SCAE deverá permitir no futuro a criação, de forma automática, de diferentes configurações para as estruturas a partir da definição do nível de segurança pretendido em termos de resistência.

O desenvolvimento/evolução do Sistema de Conhecimento proposto permitirá o prolongamento da vida útil das estruturas de Engenharia Civil, e um correcto planeamento da manutenção, assegurando níveis adequados de segurança.

No decurso deste trabalho foram efectuadas as seguintes publicações:

Quintela, H., Santos, M., Cruz, P., Descoberta de Conhecimento em Bases de Dados - Modelos de Previsão da Carga Crítica em Estruturas de Engenharia Civil, Actas da 4.ª Conferência da Associação Portuguesa de Sistemas de Informação, Porto, Portugal, Outubro 2003, APSI.

Santos, M., **Quintela, H.**, Cruz, P., Forecasting of the ultimate resistance of steel beams subjected to concentrated loads using data mining techniques, Proceedings of the Fourth International Conference on Data Mining, Rio de Janeiro, Brazil, December 2003, Wessex Institute.

Quintela, H., Santos, M., Cruz, P., Modelos de Previsão da Carga Crítica e Tensão Última de Estruturas de Engenharia Civil, Utilizando Técnicas de Inteligência Artificial, Actas Sessão de Estudantes (EPIA'03), Beja, Portugal, Dezembro 2003, APPIA.

Quintela, H., Cruz, P.J.S., Santos, M., Previsão da Tensão Última de Vigas em I de Inércia Variável Utilizando Técnicas de Data Mining, Actas do IV Congresso de Construção Metálica e Mista, Lisboa, Portugal, Dezembro 2003, CMM.

Cruz, P.J.S., **Quintela, H.**, Santos, M.F., Prediction of ultimate shear resistance of non-prismatic tapered plate girders using data mining techniques, IABMAS'04 – Second International Conference on Bridge Maintenance, Safety and Management, Kyoto, Japan, October 2004, IABMAS.

Anexo A

Clementine **D**ata **M**ining **S**ystem

Neste anexo é apresentada de forma resumida a aplicação de Data Mining, Clementine Data Mining System.

1 Introdução

O Clementine Data Mining System é um sistema de DCBD baseado em programação visual, que inclui diversas técnicas de aprendizagem automática, como Indução de Regras, Redes Neuronais, Árvores de Decisão. Disponibiliza ainda ferramentas que permitem manipular, explorar, visualizar e construir modelos sobre os dados (Figura A.1).

Figura A.1 Interface do Clementine Data Mining System v.6.5.

A filosofia de trabalho do Clementine assenta na construção de *streams* (Figura A.2) nas quais cada operação sobre os dados é representado por um nodo. Nodos com funções similares encontram-se agrupados em paletas (Tabela A.1), permitindo ao utilizador escolher o nodo mais apropriado para a execução de determinada tarefa.

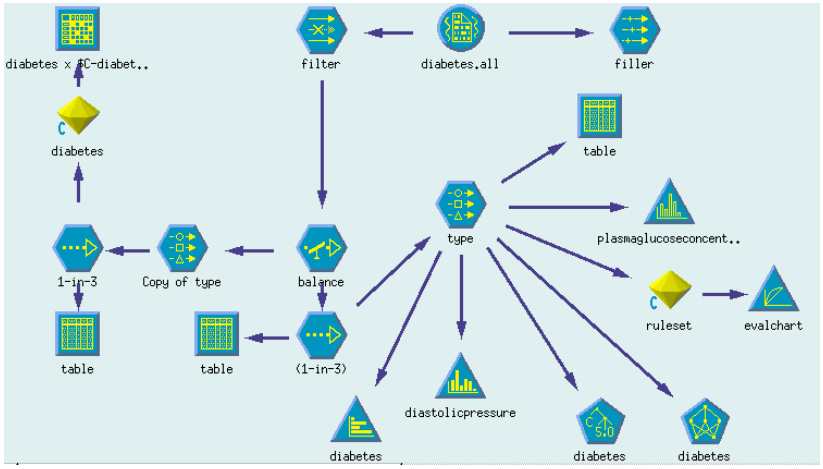


Figura A.2 Exemplo de uma *Stream*.

Tabela A.1 Paletas do Clementine.

PALETA	DESCRIÇÃO	FUNCIONALIDADES MAIS IMPORTANTES
SOURCES (Accesso aos Dados)	São disponibilizados diversos mecanismos de acesso aos dados.	Importação/leitura de dados de ficheiros com tamanhos fixos ou variáveis; Importação/leitura de dados via remota (ODBC); Importação/leitura de dados através de ligação ao software SPSS (incluindo *value labels* e *variable labels*);
RECORD OPS (Operações sobre registos)	Permite efectuar diversas operações sobre registos.	Criação de amostras; Selecção de registos; Ordenação de registos; Fusão de ficheiros num só; Balanceamento dos dados;
FIELD OPS (Operações sobre atributos)	Permite efectuar diversas operações sobre os campos. dos registos dos dados.	Criação de novos atributos; Instanciação de atributos; Filtragem; Tratamento de valores nulos;
GRAPHICS (Gráficos)	Permite a exploração dos dados.	Histogramas; Distribuição; Análise de Correlações;
MODELLING (Modelação)	Disponibiliza diversos algoritmos de Data Mining.	Árvores de Decisão; Algoritmos de Segmentação; Algoritmos de Associação; Algoritmos de Redes Neuronais; Regressão Linear; Regressão Logística
OUTPUT (Resultados)	Permite a visualização, análise e exportação dos resultados associados à realização de uma determinada tarefa.	Análise dos resultados obtidos; Análises Estatísticas; Gravação de dados; Exportação de dados via ODBC; Exportação de dados no formato SPSS; Visualização dos dados em vários formatos (e.g., tabelas, ficheiros HTML, ficheiros de texto);

Anexo B

Base de **D**ados de **V**igas de **A**ço

Sujeitas a **C**argas **C**oncentradas

Neste anexo é apresentada a Base de Dados utilizada para a geração de modelos de previsão da Carga Crítica de vigas de aço sujeitas a cargas concentradas.

Tabela B.1 Base de Dados de vigas de aço sujeitas a cargas concentradas.

a	***h***	***tw***	***bf***	***tf***	***c***	***fw***	***ff***	***pe***
2400	700	3.26	150	6.10	0	326	347	95.16
2400	700	3.26	150	6.10	100	326	347	105.95
2400	700	3.26	200	8.50	0	326	235	105.46
2400	700	3.26	200	8.50	100	326	235	121.64
2400	700	3.26	250	10.10	0	326	243	120.66
2400	700	3.26	250	10.10	100	326	243	132.93
2400	700	3.26	250	11.90	0	326	232	126.06
2400	700	3.26	250	11.90	100	326	232	139.30
2400	700	3.26	300	15.30	0	326	305	151.36
2400	700	3.26	300	15.30	100	326	305	156.47
2400	300	2.00	100	6.00	0	294	294	57.88

a	*h*	*tw*	*bf*	*tf*	*c*	*fw*	*ff*	*pe*
2400	300	2.00	100	6.00	180	294	294	65.73
2400	400	2.00	100	8.00	0	294	294	54.44
2400	400	2.00	100	8.00	180	294	294	59.84
2400	500	2.00	100	10.00	0	294	294	54.94
2400	500	2.00	100	10.00	180	294	294	55.92
2400	600	2.00	100	12.00	0	294	294	54.44
2400	600	2.00	100	12.00	180	294	294	61.80
2400	700	2.00	100	15.00	0	294	294	55.92
2400	700	2.00	100	15.00	180	294	294	56.90
9800	700	3.40	250	10.00	0	280	280	111.46
9800	700	3.40	250	10.00	100	280	280	118.36
9800	700	3.40	250	10.00	200	280	280	120.47
1000	1000	2.50	160	5.50	100	298	342	51.50
1000	1000	2.50	200	10.09	100	299	253	63.76
1000	1000	2.50	200	16.24	100	251	266	68.67
1000	1000	2.50	200	20.17	100	254	231	88.29
1000	1000	2.50	250	30.50	100	289	261	179.00
2000	1000	3.00	160	6.39	100	290	294	81.90
2000	1000	3.00	200	10.00	100	297	253	98.10
2000	1000	3.00	200	16.55	100	308	266	117.72
2000	1000	3.00	200	19.78	100	300	231	125.57
2000	1000	3.00	250	30.00	100	299	261	147.15
2000	1000	3.00	160	6.29	200	290	294	93.19
2000	1000	3.00	200	10.00	200	297	253	117.72
2000	1000	3.00	200	16.55	200	308	266	132.43
2000	1000	3.00	200	19.78	200	300	231	152.05
2000	1000	3.00	250	30.00	200	299	261	157.94
500	500	2.00	50	5.95	50	243	294	37.28
500	500	2.00	45	16.25	50	243	261	53.95
500	500	2.00	50	24.60	50	243	225	76.03
1000	500	2.00	50	4.97	100	243	294	35.56
1000	500	2.00	45	15.88	100	243	261	49.26
1000	500	2.00	60	24.80	100	243	225	56.40
1270	635	3.25	152	12.70	50	250	250	89.40
864	635	3.25	152	12.70	50	250	250	124.00
660	635	3.25	152	12.70	75	250	250	141.20
300	300	3.97	49	10.00	30	285	269	130.00
300	300	4.00	51	9.90	30	270	258	147.50
300	300	4.01	49	15.90	30	281	265	169.50
450	450	3.97	49	10.00	45	257	267	120.00
450	450	3.96	50	15.80	45	249	265	150.00
600	600	3.57	51	10.00	60	257	274	140.00
600	600	3.63	50	10.10	60	282	279	148.00
600	600	3.67	49	16.00	60	306	282	150.00

a	*h*	*tw*	*bf*	*tf*	*c*	*fw*	*ff*	*pe*
300	300	3.97	49	10.00	45	285	269	150.00
300	300	4.00	51	9.90	60	270	258	146.00
300	300	4.01	49	15.90	30	281	265	150.00
450	450	3.97	49	10.00	60	257	267	136.00
450	450	3.96	50	15.80	45	249	265	160.00
600	600	3.57	51	10.00	30	257	274	119.00
600	600	3.63	50	10.10	45	282	279	138.00
600	600	3.67	49	16.00	60	306	282	146.00
800	800	2.05	300	15.50	40	210	285	60.00
800	800	2.00	120	5.10	40	210	285	38.00
800	800	2.05	300	15.50	40	210	285	66.00
800	800	2.00	120	5.10	40	210	285	41.00
2500	800	2.00	300	15.00	40	308	285	64.00
1200	800	2.00	300	15.00	40	308	285	66.00
600	800	2.00	300	15.00	40	308	285	84.00
2500	800	3.00	250	12.00	40	245	285	84.00
1200	800	3.00	250	12.00	40	245	285	85.00
600	800	3.00	250	12.00	40	245	285	96.80
2200	680	2.00	120	5.00	40	354	292	47.10
1020	680	2.00	120	5.00	40	354	292	57.17
510	680	2.00	120	5.00	40	354	292	51.50
800	800	2.00	120	5.00	40	285	286	48.00
800	600	2.00	120	5.00	40	285	286	42.00
800	400	2.00	120	5.00	40	285	286	48.00
800	300	2.00	120	5.00	40	285	286	49.00
400	400	2.00	120	5.00	40	285	286	53.00
400	300	2.00	120	5.00	40	285	286	51.00
800	800	3.00	250	12.00	40	328	286	121.00
800	600	3.00	250	12.00	40	328	286	120.00
800	400	3.00	250	12.00	40	328	286	119.00
600	250	2.12	149	3.05	50	224	221	32.64
600	250	2.12	149	6.75	50	224	279	42.64
600	250	2.12	149	11.75	50	224	305	52.80
600	250	3.05	149	3.05	50	221	221	79.68
600	250	3.05	149	6.75	50	221	279	100.70
600	250	3.05	149	11.75	50	221	305	129.10
600	500	2.12	149	3.05	50	224	221	34.08
600	500	2.12	149	6.75	50	224	279	37.92
600	500	2.12	149	11.75	50	224	305	44.16
600	500	2.12	149	20.06	50	224	305	84.48
600	500	3.05	149	3.05	50	221	221	70.56
600	500	3.05	149	6.75	50	221	279	90.72
600	500	3.05	149	11.75	50	221	305	111.36
600	500	3.05	149	20.06	50	221	305	130.60

a	*h*	*tw*	*bf*	*tf*	*c*	*fw*	*ff*	*pe*
600	750	2.12	149	3.05	50	224	221	30.00
600	750	2.12	149	6.75	50	224	279	38.40
600	750	2.12	149	11.75	50	224	305	53.04
600	750	3.05	149	3.05	50	221	221	67.39
600	750	3.05	149	6.75	50	221	279	81.12
600	750	3.05	149	11.75	50	221	305	99.55
760	380	1.96	80	3.05	50	178	272	33.55
760	380	2.99	80	6.25	50	245	298	84.10
500	500	3.01	150	5.94	50	242	308	89.00
500	500	3.01	150	5.94	50	242	308	89.00
2400	300	2.00	100	6.00	40	207	277	39.80
3000	400	2.00	100	12.00	40	205	278	40.70
2400	600	2.00	100	6.00	40	206	280	35.00
3000	800	2.00	100	12.00	40	205	277	41.90
900	300	2.00	100	6.00	40	207	277	34.10
1100	400	2.00	100	12.00	40	205	278	36.90
900	600	2.00	100	6.00	40	206	280	31.00
1100	800	2.00	100	12.00	40	205	277	40.50
900	300	2.00	100	6.00	120	207	277	38.40
1100	400	2.00	100	12.00	120	205	278	42.10
900	600	2.00	100	6.00	120	206	280	37.50
1100	800	2.00	100	12.00	120	205	277	46.50
3000	800	2.00	250	12.00	40	206	270	38.20
3000	800	3.00	250	12.00	40	215	268	81.50
1100	800	2.00	250	12.00	40	206	270	41.40
1100	800	2.00	250	12.00	40	215	268	90.70
1100	800	2.00	250	12.00	120	206	270	41.40
1100	800	3.00	250	12.00	120	215	268	91.50

Anexo C

Base de **D**ados de **V**igas em **I** de **I**nércia **V**ariável

Neste anexo é apresentada a Base de Dados utilizada para a geração de modelos de previsão da Tensão Crítica de vigas com perfil em I de inércia variável.

Tabela C.1 Base de Dados de vigas em I de inércia variável.

α	λ_f	η	$\mathbf{tg}\,\phi$	t_w	h_0	k_{mod}	k_{mef}	$\tau_{cr,\mathrm{mod}}$	$\tau_{cr,\mathrm{mef}}$
2.38	29.16	.35	.30	8.00	551.00	2.72	2.66	108.69	106.43
1.75	29.16	.35	.40	8.00	576.00	3.00	2.97	109.82	108.86
1.38	29.16	.35	.50	8.00	601.00	3.40	3.50	114.50	117.71
2.25	29.16	.35	.30	8.00	626.00	3.04	2.96	94.11	91.80
1.25	29.16	.35	.50	8.00	726.00	4.21	4.17	96.98	96.03
2.00	29.16	.35	.30	8.00	776.00	3.72	3.66	70.60	69.53
3.00	29.16	.35	.20	8.00	776.00	3.46	3.46	69.75	69.75
3.00	56.25	.45	.20	8.00	784.00	3.25	3.23	64.32	63.75
1.19	29.16	.35	.50	8.00	788.50	4.66	4.69	91.12	91.55
1.00	12.50	.20	.60	8.00	800.00	5.11	5.04	97.02	95.60
1.00	16.67	.20	.60	8.00	800.00	4.89	4.87	92.86	92.48

α	λ_f	η	$\mathbf{tg}\,\phi$	t_w	h_0	k_{mod}	k_{mef}	$\tau_{cr,\mathrm{mod}}$	$\tau_{cr,\mathrm{mef}}$
1.00	20.00	.20	.60	8.00	800.00	4.72	4.74	89.68	90.23
1.00	25.00	.20	.60	8.00	800.00	4.47	4.52	84.80	85.74
1.00	40.00	.20	.60	8.00	800.00	3.70	3.69	70.30	70.11
1.00	50.00	.20	.60	8.00	800.00	3.18	3.21	60.35	60.94
1.00	15.63	.25	.60	8.00	800.00	5.23	5.14	99.31	97.56
1.00	20.00	.25	.60	8.00	800.00	5.08	5.04	96.48	95.70
1.00	30.00	.25	.60	8.00	800.00	4.74	4.74	90.01	89.94
1.00	40.00	.25	.60	8.00	800.00	4.40	4.35	83.53	82.62
1.00	50.00	.25	.60	8.00	800.00	4.06	3.88	77.05	73.62
1.00	21.88	.35	.60	8.00	800.00	5.38	5.27	102.12	100.00
1.00	29.17	.35	.60	8.00	800.00	5.25	5.20	99.58	98.63
1.00	43.75	.35	.60	8.00	800.00	4.98	4.94	94.50	93.76
1.00	60.00	.35	.60	8.00	800.00	4.68	4.42	88.85	83.98
1.00	16.67	.45	.60	8.00	800.00	5.60	5.40	106.26	102.46
1.00	28.13	.45	.60	8.00	800.00	5.47	5.36	103.79	101.66
1.00	37.50	.45	.60	8.00	800.00	5.36	5.35	101.74	101.55
1.00	40.00	.45	.60	8.00	800.00	5.33	5.33	101.25	101.25
1.00	56.25	.45	.60	8.00	800.00	5.14	5.12	97.64	97.26
1.00	60.00	.45	.60	8.00	800.00	5.10	5.04	96.82	95.60
1.13	29.16	.35	.50	8.00	851.00	5.16	5.21	86.50	87.51
1.38	29.16	.35	.40	8.00	876.00	4.71	4.75	74.55	75.13
1.75	29.16	.35	.30	8.00	926.00	4.51	4.73	63.84	67.06
1.25	29.16	.35	.40	8.00	976.00	5.45	5.57	69.55	71.07
2.50	29.16	.35	.20	8.00	976.00	4.34	4.39	55.39	56.06
1.00	12.50	.20	.50	8.00	1000.00	6.22	6.26	75.55	76.02
1.00	16.67	.20	.50	8.00	1000.00	5.97	6.06	72.54	73.59
1.00	20.00	.20	.50	8.00	1000.00	5.77	5.92	70.12	72.03
1.00	25.00	.20	.50	8.00	1000.00	5.48	5.63	66.51	68.36
1.00	35.00	.20	.50	8.00	1000.00	4.88	5.02	59.27	60.92
1.00	50.00	.20	.50	8.00	1000.00	3.99	4.15	48.42	50.42
1.00	15.63	.25	.50	8.00	1000.00	6.36	6.41	77.21	77.81
1.00	20.83	.25	.50	8.00	1000.00	6.15	6.21	74.67	75.47
1.00	31.25	.25	.50	8.00	1000.00	5.73	5.81	69.59	70.55
1.00	45.00	.25	.50	8.00	1000.00	5.18	5.18	62.89	62.97
1.00	60.00	.25	.50	8.00	1000.00	4.58	4.40	55.57	53.44
1.00	21.88	.35	.50	8.00	1000.00	6.53	6.56	79.29	79.69
1.00	29.17	.35	.50	8.00	1000.00	6.37	6.43	77.33	78.16
1.00	43.75	.35	.50	8.00	1000.00	6.04	6.12	73.41	74.37
1.00	60.00	.35	.50	8.00	1000.00	5.68	5.60	69.04	67.97
1.00	28.13	.45	.50	8.00	1000.00	6.63	6.67	80.58	81.02
1.00	37.50	.45	.50	8.00	1000.00	6.50	6.57	78.96	79.79
1.00	45.00	.45	.50	8.00	1000.00	6.39	6.49	77.67	78.83
1.00	60.00	.45	.50	8.00	1000.00	6.18	6.24	75.09	75.81

α	λ_f	η	$\mathbf{tg}\,\phi$	t_w	h_0	k_{mod}	k_{mef}	$\tau_{cr,\mathrm{mod}}$	$\tau_{cr,\mathrm{mef}}$
1.50	29.16	.35	.30	8.00	1076.00	5.45	5.51	57.23	57.79
2.00	29.16	.35	.20	8.00	1176.00	5.39	5.48	47.37	48.13
4.00	29.16	.35	.10	8.00	1176.00	4.79	4.97	42.08	43.66
2.00	25.00	.20	.20	8.00	1184.00	4.36	4.73	37.77	41.01
2.00	56.25	.45	.20	8.00	1184.00	5.19	5.21	44.97	45.13
1.00	12.50	.20	.40	8.00	1200.00	7.34	7.46	61.95	62.96
1.00	16.67	.20	.40	8.00	1200.00	7.07	7.23	59.66	60.97
1.00	20.00	.20	.40	8.00	1200.00	6.85	7.06	57.83	59.63
1.00	25.00	.20	.40	8.00	1200.00	6.53	6.68	55.09	56.38
1.00	35.00	.20	.40	8.00	1200.00	5.88	6.02	49.60	50.79
1.00	50.00	.20	.40	8.00	1200.00	4.90	5.14	41.36	43.36
1.00	15.63	.25	.40	8.00	1200.00	7.49	7.61	63.18	64.19
1.00	20.00	.25	.40	8.00	1200.00	7.29	7.43	61.53	62.66
1.00	30.00	.25	.40	8.00	1200.00	6.85	6.98	57.77	58.89
1.00	40.00	.25	.40	8.00	1200.00	6.40	6.48	54.01	54.69
1.00	60.00	.25	.40	8.00	1200.00	5.51	5.43	46.49	45.79
1.00	21.88	.35	.40	8.00	1200.00	7.68	7.81	64.78	65.88
1.00	29.17	.35	.40	8.00	1200.00	7.49	7.65	63.22	64.52
1.00	43.75	.35	.40	8.00	1200.00	7.13	7.25	60.12	61.20
1.00	60.00	.35	.40	8.00	1200.00	6.72	6.71	56.66	56.58
1.00	16.67	.45	.40	8.00	1200.00	7.99	8.01	67.40	67.57
1.00	28.13	.45	.40	8.00	1200.00	7.80	7.92	65.80	66.80
1.00	35.00	.45	.40	8.00	1200.00	7.68	7.88	64.75	66.44
1.00	37.50	.45	.40	8.00	1200.00	7.65	7.79	64.50	65.69
1.00	56.25	.45	.40	8.00	1200.00	7.34	7.49	61.89	63.15
1.25	29.16	.35	.30	8.00	1226.00	6.71	6.86	54.26	55.45
3.50	29.16	.35	.10	8.00	1276.00	5.26	5.47	39.27	40.83
1.50	29.16	.35	.20	8.00	1376.00	6.86	7.00	44.02	44.91
1.50	29.16	.35	.40	8.00	1376.00	4.07	4.10	26.12	26.31
3.00	29.16	.35	.10	8.00	1376.00	5.77	5.90	37.00	37.84
1.50	25.00	.20	.20	8.00	1384.00	5.83	6.04	36.95	38.29
1.50	56.25	.45	.20	8.00	1384.00	6.66	6.76	42.22	42.85
1.00	12.50	.20	.30	8.00	1400.00	8.48	8.60	52.54	53.32
1.00	16.67	.20	.30	8.00	1400.00	8.19	8.31	50.77	51.51
1.00	20.00	.20	.30	8.00	1400.00	7.96	8.13	49.35	50.44
1.00	25.00	.20	.30	8.00	1400.00	7.62	7.76	47.23	48.08
1.00	35.00	.20	.30	8.00	1400.00	6.94	7.02	42.98	43.53
1.00	50.00	.20	.30	8.00	1400.00	5.91	6.09	36.61	37.72
1.00	15.63	.25	.30	8.00	1400.00	8.63	8.80	53.48	54.54
1.00	20.83	.25	.30	8.00	1400.00	8.38	8.51	51.94	52.73
1.00	31.25	.25	.30	8.00	1400.00	7.88	8.02	48.86	49.69
1.00	45.00	.25	.30	8.00	1400.00	7.23	7.27	44.79	45.03
1.00	60.00	.25	.30	8.00	1400.00	6.51	6.47	40.35	40.12

α	λ_f	η	$\mathbf{tg}\,\phi$	t_w	h_0	k_{mod}	k_{mef}	$\tau_{cr,\mathrm{mod}}$	$\tau_{cr,\mathrm{mef}}$
1.00	21.88	.35	.30	8.00	1400.00	8.83	8.98	54.75	55.64
1.00	29.17	.35	.30	8.00	1400.00	8.63	8.79	53.49	54.46
1.00	43.75	.35	.30	8.00	1400.00	8.23	8.36	50.99	51.81
1.00	60.00	.35	.30	8.00	1400.00	7.78	7.78	48.20	48.21
1.00	28.13	.45	.30	8.00	1400.00	8.97	9.09	55.58	56.30
1.00	37.50	.45	.30	8.00	1400.00	8.80	8.95	54.51	55.45
1.00	45.00	.45	.30	8.00	1400.00	8.66	8.81	53.65	54.58
1.00	60.00	.45	.30	8.00	1400.00	8.38	8.49	51.94	52.62
2.00	36.00	.30	.15	8.00	1400.00	5.71	5.81	24.57	25.00
1.50	29.16	.35	.50	8.00	1476.00	2.71	2.43	145.17	130.34
.50	29.16	.35	.50	8.00	1476.00	19.04	19.17	106.14	106.87
1.25	29.16	.35	.20	8.00	1476.00	7.99	7.87	44.54	43.89
2.50	29.16	.35	.10	8.00	1476.00	6.37	6.58	35.51	36.66
.60	29.16	.35	.40	8.00	1496.00	15.16	15.66	82.30	84.97
.80	29.16	.35	.30	8.00	1496.00	11.12	11.50	60.35	62.39
.50	29.16	.35	.40	8.00	1576.00	20.26	20.31	99.13	99.37
2.00	29.16	.35	.10	8.00	1576.00	7.13	7.30	33.81	34.62
1.00	12.50	.20	.20	8.00	1600.00	9.62	9.73	45.63	46.19
1.00	16.67	.20	.20	8.00	1600.00	9.32	9.45	44.24	44.82
1.00	20.00	.20	.20	8.00	1600.00	9.10	9.25	43.13	43.84
1.00	25.00	.20	.20	8.00	1600.00	8.74	8.83	41.46	41.89
1.00	35.00	.20	.20	8.00	1600.00	8.03	8.06	38.13	38.27
1.00	50.00	.20	.20	8.00	1600.00	6.98	7.06	33.12	33.49
1.00	15.63	.25	.20	8.00	1600.00	9.77	9.88	46.38	46.88
1.00	20.00	.25	.20	8.00	1600.00	9.56	9.68	45.34	45.92
1.00	30.00	.25	.20	8.00	1600.00	9.06	9.16	42.98	43.46
1.00	40.00	.25	.20	8.00	1600.00	8.56	8.62	40.61	40.92
1.00	60.00	.25	.20	8.00	1600.00	7.56	7.47	35.88	35.45
1.00	21.88	.35	.20	8.00	1600.00	9.99	10.09	47.40	47.85
1.00	29.17	.35	.20	8.00	1600.00	9.77	9.88	46.37	46.89
1.00	43.75	.35	.20	8.00	1600.00	9.34	9.46	44.32	44.89
1.00	60.00	.35	.20	8.00	1600.00	8.86	8.86	42.03	42.04
1.00	16.67	.45	.20	8.00	1600.00	10.36	10.31	49.14	48.91
1.00	28.13	.45	.20	8.00	1600.00	10.14	10.19	48.10	48.34
1.00	35.00	.45	.20	8.00	1600.00	10.00	10.08	47.44	47.82
1.00	37.50	.45	.20	8.00	1600.00	9.95	10.05	47.20	47.70
1.00	56.25	.45	.20	8.00	1600.00	9.57	9.66	45.41	45.83
1.00	60.00	.45	.20	8.00	1600.00	9.49	9.59	45.05	45.50
.60	29.16	.35	.30	8.00	1616.00	16.38	16.54	76.19	76.93
.80	29.16	.35	.20	8.00	1656.00	12.30	12.32	54.49	54.56
.80	25.00	.20	.20	8.00	1664.00	11.27	11.27	49.43	49.44
.80	56.25	.45	.20	8.00	1664.00	12.10	12.15	53.08	53.29
.50	29.16	.35	.30	8.00	1676.00	21.50	21.47	93.00	92.86

α	λ_f	η	$\mathbf{tg}\,\phi$	t_w	h_0	k_{mod}	k_{mef}	$\tau_{cr,\mathrm{mod}}$	$\tau_{cr,\mathrm{mef}}$
1.50	29.16	.35	.10	8.00	1676.00	8.29	8.41	35.85	36.36
.50	44.80	.40	.30	7.15	1700.00	21.33	21.15	71.50	70.90
1.25	29.16	.35	.10	8.00	1726.00	9.28	9.39	37.84	38.29
.60	29.16	.35	.20	8.00	1736.00	17.60	17.40	70.96	70.13
.60	25.00	.20	.20	8.00	1744.00	16.57	16.18	66.18	64.62
.60	56.25	.45	.20	8.00	1744.00	17.40	17.18	69.50	68.61
.50	29.16	.35	.20	8.00	1776.00	22.75	22.67	87.62	87.32
.50	25.00	.20	.20	8.00	1784.00	21.72	21.26	82.89	81.16
.50	56.25	.45	.20	8.00	1784.00	22.55	22.34	86.06	85.27
.50	20.00	.20	.10	8.00	1800.00	10.22	10.29	38.34	38.62
1.00	12.50	.20	.10	8.00	1800.00	10.76	10.79	40.33	40.45
1.00	16.67	.20	.10	8.00	1800.00	10.46	10.52	39.23	39.45
1.00	25.00	.20	.10	8.00	1800.00	9.87	9.93	37.02	37.24
1.00	35.00	.20	.10	8.00	1800.00	9.16	9.13	34.36	34.24
1.00	50.00	.20	.10	8.00	1800.00	8.10	8.13	30.38	30.47
1.00	16.63	.25	.10	8.00	1800.00	10.87	10.90	40.74	40.88
1.00	20.83	.25	.10	8.00	1800.00	10.65	10.66	39.94	39.97
1.00	31.25	.25	.10	8.00	1800.00	10.12	10.19	37.94	38.19
1.00	45.00	.25	.10	8.00	1800.00	9.42	9.41	35.31	35.29
1.00	60.00	.25	.10	8.00	1800.00	8.65	8.56	32.44	32.07
1.00	21.88	.35	.10	8.00	1800.00	11.15	11.09	41.78	41.58
1.00	29.16	.35	.10	8.00	1800.00	10.92	10.89	40.93	40.84
1.00	43.75	.35	.10	8.00	1800.00	10.46	10.47	39.23	39.23
1.00	60.00	.35	.10	8.00	1800.00	9.96	9.94	37.33	37.26
1.00	28.13	.45	.10	8.00	1800.00	11.30	11.15	42.38	41.80
1.00	37.50	.45	.10	8.00	1800.00	11.10	11.00	41.62	41.23
1.00	45.00	.45	.10	8.00	1800.00	10.94	10.87	41.02	40.76
1.00	60.00	.45	.10	8.00	1800.00	10.62	10.60	39.80	39.76
.80	29.16	.35	.10	8.00	1816.00	13.33	13.06	49.09	48.12
.60	29.16	.35	.10	8.00	1856.00	18.35	18.26	64.69	64.38
.50	29.16	.35	.10	8.00	1876.00	24.01	24.00	82.87	82.84
.50	29.16	.35	.00	8.00	1976.00	25.32	25.52	78.77	79.39
.60	29.16	.35	.00	8.00	1976.00	20.12	19.33	62.58	60.15
1.00	29.16	.35	.00	8.00	1976.00	12.06	11.91	37.53	37.04
1.50	29.16	.35	.00	8.00	1976.00	9.84	9.89	30.62	30.77
2.00	29.16	.35	.00	8.00	1976.00	9.06	9.11	28.20	28.33
2.50	29.16	.35	.00	8.00	1976.00	8.70	8.74	27.08	27.20
3.00	29.16	.35	.00	8.00	1976.00	8.51	8.46	26.47	26.31
4.00	29.16	.35	.00	8.00	1976.00	8.31	8.03	25.87	24.97

Glossário de Termos

Termo	Descrição
Alma de uma Viga	Parte de uma viga que resiste principalmente aos esforços transversais.
Apoio	Sistema que realiza uma ligação exterior.
Bambeamento	Fenómeno de instabilidade que se caracteriza pela ocorrência de grandes deformações transversais ao plano em que actuam os esforços de flexão.
Banzo	Parte de uma viga que resiste principalmente aos momentos flectores.
Carga	Força exterior devido à acção da gravidade.
Cedência	Fenómeno de deformação rápida e não recuperável de um corpo, sem aumento apreciável da tensão.
Coeficiente de Esbelteza	Quociente do comprimento de uma escora de secção constante pelo raio de giração mínimo da sua secção transversal.

Termo	Descrição
Coeficiente de Poisson	Constante elástica de um corpo homogéneo e isótropo em estado de tensão simples, que define a relação entre a extensão principal mínima e a extensão principal máxima. Esta constante relaciona o módulo de elasticidade com o módulo de distorção do corpo.
Coluna	Peça linear, de eixo rectilíneo vertical, sujeita principalmente a esforços de compressão.
Deformação	Transformação que se traduz por variação da distância entre pontos de um corpo.
Elasticidade	Propriedade de um corpo recuperar a sua forma primitiva quando deixa de actuar a solicitação que produziu a sua deformação.
Encurvadura	Fenómeno de instabilidade que se caracteriza pela ocorrência de grandes deformações transversais.
Enfunamento	Encurvadura de uma placa.
Estrutura	Corpo ou conjunto de corpos adequado a resistir a solicitações.
Fadiga	Diminuição da resistência de um corpo por efeito denuma solicitação periódica.
Força de Massa	Força exterior aplicada a um elemento de volume de um corpo.
Força de Superfície	Força exterior que se considera aplicada num ponto.
Força Exterior	Força, proveniente do meio exterior, aplicada a um corpo.
Força Interior	Força interior de reacção às solicitações aplicadas a um corpo.
Ligação Exterior	Sujeição imposta pelo meio exterior à liberdade de deslocamento de pontos de um corpo.
Módulo de Distorção	Constante elástica de um corpo homogéneo e isótropo em estado de tensão tangencial simples, que define a relação entre a tensão tangencial máxima e a distorção correspondente.

Termo	**Descrição**
Módulo de Elasticidade	Constante elástica de um corpo homógeneo e isótropo em estado de tensão simples, que define a relação entre a tensão principal num ponto e a extensão principal correspondente.
Plasticidade	Propriedade de um corpo não recuperar a sua forma primitiva quando deixa de actuar a solicitação que produziu a sua deformação.
Solicitação	Causa exterior capaz de produzir ou de alterar o estado de tensão ou de deformação dum corpo (forças de massa ou de superfície; variações de temperatura ou acções equivalentes;cedências de apoios).
Tensão de Cedência	Tensão correspondente ao início da cedência.
Viga	Peça linear sujeita principalmente a esforços de flexão.

Bibliografia

A

[Adriaans e Zantinge, 1996] Adriaans P, Zantinge, D. (1996), Data Mining, Addison Wesley Longman, Edimburgo.

[Akhras et al., 1994] Akhras, G., Foo, HC (1994), A knowledge-based system for selecting proportions of normal concrete, Ex Sys Appl, 7, pp. 323-335.

[ARFF, 2004] Attribute-Relation File Format, http://www.cs.waikato.ac.nz/l, Dezembro, 2004.

B

[Bento et al., 1997] Bento, J., Ndumu D., Dias, J. (1997), Application of Neural Networks to the Seismic Analysis of Structures.

[Berry Linoff, 2000] Berry, M. J. A., Linoff, G. (2000), Mastering Data Mining: The Art and Science of Customer Relationships Management, John Wiley & Sons, Inc., USA.

[Berson et al., 2000] Berson, A., Smith, S.. Thearling, K. (2000), Building Data Mining Applications for CRM, McGraw-Hill, USA.

[Berthold e Hand, 2002] Berthold, M., Hand, D. J. (Eds.) (2002), Intelligent Data Analysis – An Introduction (Second Edition), Springer.

[Bien, 2004] Bien, J. (2004), Identification of Bridge Damages During Monitoring, International Bridge Engineering School (BRIDMO), Portugal.

[Buchanan et al., 1983] Buchanan, B., Barstow, D., Bechtel, R. (1983), Building Expert Systems, Constructing an Expert System, Addison-Wesley, pp.127-169.

[Brown, 2003] Brown, M.L., Kros, J. F. (2003), Data Mining and the impact of missing data, Industrial Management and Data Systems, 108, 200-219.

C

[Cielen et al., 2004] Cielen, A., Peeters, L., Vanhoof, K. (2004), Bankruptcy Prediction using a data development analysis (2004), European Journal of Operational Research, Volume 154, Issue 2, 526-532.

[Chapman et al, 2000] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000), Crisp-DM 1.0: Step-by-step data mining methods, http://www.crisp-dm.org - Agosto, 2004.

[Chern & Ostapenko, 1969] Chern, C., Ostapenko, A. (1996), Ultimate Strength of Plate Girders Under Shear, Fritz Eng. Lab. Report N.º 38.7, University of Lehigh, Bethlehem, Pa., Aug.

[Chester, 1993] Chester, M. (1993), Neural Network – A tutorial, PTR Prentice-Hall Inc., USA.

[Cortez e Neves, 2000] Cortez, P., Neves, J. (2000), Redes Neuronais Artificiais, Unidade de Ensino, Departamento de Informática, Escola de Engenharia, Universidade do Minho, Braga, Portugal.

[Cortez, 2002] Cortez. P. (2002), Modelos Inspirados na Natureza para a Previsão de Séries Temporais, Tese de Doutoramento, Departamento de Informática, Universidade do Minho, Portugal.

[Cortez, 2004] Cortez, P. (2004), Aprendizagem e Avaliação de Modelos, Apontamentos Pedagógicos, Departamento de Sistemas de Informação, Universidade do Minho, Portugal.

[Cramer, 1985] Cramer, M.L. (1985), A Representation for the adaptive generation of simple sequential programs, in Proceedings of the 1st. International Conference on Genetic Algorithms and their Applications, pp. 183-187, Lawrence Erlbaum Associates.

[Cruz e Guimarães, 2003] Cruz, P. S. J., Guimarães, L. C. B. (2003), Análise Numérica do Efeito da Redução de Espessura da Alma na Estabilidade de Vigas I de Inércia Variável, em Lamas, A., Calado, L., Ferreira, J. e Real, P.V. (Eds.), IV Congresso de Construção Metálica e Mista, CMM, 2003.

D

[DMG, 2004] Data Mining Group, http://www.dmg.org, Dezembro,2004.

[Davies e Mandal, 1979] Davies, G., Mandal, S.N. (1979), The collapse behaviour of tapered plate girders loaded within tip, Proceedings of the Institution of Civil Engineers, Part 2, pp. 65 – 80.

E

[Egan, 1975] Egan, J. (1975), Signal detection theory and ROC analysis, Academic Press, New York.

F

[Falby e Lee, 1976] Falbym W.E., Lee, G.C. (1976), Tension Field design of tapered webs, Engineering Journal, AISC, pp. 11-17.

[Fayyad et al., 1996] U.M. Fayyad, G. Piatetsky-Shapiro, e P.Smyth. From Data Mining to Knowledge Discovery: An Overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P.Smyth, e R. Uthurusamy (Eds.) (1996), *Advances in Knowledge Discovery and Data Mining*, 1-34. The MIT Press, Massachussets.

[Fayyad e Stolorz, 1997] Fayyad, U., Stolorz, P. (1997), Data Mining and KDD: Promise and challenges, Future Generation Computer Systems, Vol. 13, pp. 99-115.

[Feelders, 2002] Feeelders, A. J. (2002), Statistical Concepts. In Berthold, M., Hand, D. J. (Eds), *Intelligent Data Analysis: An Introduction, Second Edition*, 17-68. Springer-Verlag.

[Fonseca, 1999] Fonseca, E. (1999), Avaliação do Efeito de Cargas Concentradas em vigas de aço através de algoritmos de redes neuronais, Dissertação de Mestrado, Pontíficia Universidade Católica do Rio de Janeiro, Brasil.

[Foo et al., 1993] Foo, HC, Akhras, G. (1993), Expert systems and design of concrete mixtures, Conc. Int., 15, pp. 42-46.

[Fu, 1994] Fu, L.M. (1994), Rule Generation from Neural Networks, IEEE Transactions on Systems, Man and Cybernetics, N.º 24, pp. 183-187.

G

[Gallant, 1995] Gallant, S. (1995), Neural Network Learning and Expert Systems, MIT Press, USA.

[Gelman et al., 1995] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.R. (1995), Bayesian Data Analysis, Chapman & Hall, London.

[Groth, 2000] Groth, Robert (2000), Data Mining: Building Competitive Advantage, Prentice Hall PTR, USA.

H

[Hadi, 2003] Hadi, M. N. S. (2003), Neural Networks applications in concrete structures, Computers and Structures, 81, 373-381.

[Hagan et al., 1996] Hagan, M. T., Demuth, H. B., Beale, M. (1996), Neural network design, PWS Publishing Company.

[Harrison, 1998] Harrison, T. H. (1998), Intranet Data Warehouse, São Paulo, Berkeley Brasil

J

[Jadid et al, 1994] Jadid, N.M., Fairbairn, D.R. (1994), The application of neural network techniques to structural analysis by implementing and adaptive finite element mesh generation, Artificial Intelligence for Engineering Design, Analysis and Manufacturing.

K

[Kohavi e Provost, 1998] Kohavi, R., Provost, F. (1998), Glossary of Terms, Machine Learning, 20(2/3):271-274.

L

[Langley e Simon, 1995] Langley, P., Simon, H.A. (1995), Applications of Machine Learning and Rule Induction, Communications of the ACM, 38.

[Laudon et al., 2002] Laudon, K.C., Laudon, J.P. (2002), Essential of management information systems, Englewood cliffs, Prentice Hall.

[Lamb, 1997] Lamb, C. (1997), Introducing Business Miner, Technical White Paper, Business Objects.

[Liao, 2005] Liao, S.H. (2005), Expert Systems methodologies and applicatios – a decade review from 1995 to 2004, Expert Systems with Applications, 28, pp. 93-103.

[Little, 1992] Little, R. (1992), Regression with missing X's: a review, Journal of the American Statistical Association, Vol. 87, pp. 1227-1237.

[Lyse, 1935] Lyse, Y., Godfrey, H.J. (1935), Investigation of Web Buckling in Steel Beams, ASCE Transactions, Vol. 100, paper 1907, 675-695.

M

[Metz, 1978] Metz, C.E. (1978), Basic Principles of ROC analysis, volume Semin Nucl Med 8.

[Microsoft, 2004] SQL Server Home, http://www.microsoft.com/sql/default.mspx, Novembro, 2004.

[Motta, 1998] Motta, E. (1998), Reusable Componentes for Knowledge Models, PhD Thesis, Knowledge Media University – Open University – UK, 1998.

O

[Oracle,2004] Oracle Business Intelligence Solutions, http://www.oracle.com/solutions, Novembro, 2004.

[OMG, 2004] UML – Unified Modeling Language Resource Page, http://www.uml.org, Dezembro, 2004.

P

[Pinto e al, 2004] Pinto, F., Santos, M.F., Cortez, P., Quintela, H. (2004), Data Pre-processing for Database Marketing, Data Gadgets 2004, Málaga Spain, pp 76-84.

[Provost e Fawcett, 1997] Provost, F., Fawcett, T. (1997), Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions, in Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, pages 43-48, AAAI Press.

[Prudsys, 2003] Thess, M. (2003), XELOPES Library Documentation – Version 1.1.5, Prudsys AG.

Q

[Quinlan, 1993] Quinlan, J.R. (1993), C4.5: Programs for machine learning, Morgan Kaufmann, San Francisco.

[Quinlan, 1996] Quinlan, J.R. (1996), Bagging Boosting and C4.5, Proceedings of Fourteenth National Conference on Artifical Intelligence.

[Quinlan, 1997] Quinlan, J.R. (1997), C5.0 Data Mining Tool, http://www.rulequest.com.

R

[Rezende, 2003] Rezende, S.O., Sistemas Inteligentes – Fundamentos e Aplicações (2003), Editora Manole, Brasil.

[Ribeiro, 2003] Ribeiro, P.A. (2003), Modelos de Previsão de Fenómenos Raros, Tese de Mestrado, Faculdade de Economia da Universidade do Porto, Portugal.

[Roberts et al., 1978] Roberts, T.M., Rockey, K.C. (1978), Methode Pour Predire la Charge de Ruine d'une Poutre a Ame Mince Soumis a une Charge Semi-Repartie dans le Plan de L'ame, Construction Metallique, N.º 3, 3-13.

[Roberts, 1981] Roberts, T. M. (1981), Slender Plate Girders Subjected to Edge Loading, Proc. Inst. of Civil Eng., Part. 2, Vol. 71, 805-819.

[Roberts et al., 1988] Roberts, T. M., Coric, B. (1988), Collapse of Plate Girders Subjected to Patch Loading, Miscellany Dedicated to the 65th Birthday of Academician Professor Dr. Nicola Hajdin, University of Belgrade, 203-209.

[Roberts et al., 1997] Roberts, T. M., Newark, A.C.B. (1997) Strength of Webs Subjected to Compressive Edge Loading, Journal of Structural Engineering, American Society of Civil Engineers, 123, N.º 2, 176-183.

[Rocha, 1997] Rocha, M. (1997), Uma aproximação à resolução do caixeiro viajante via programação genética, Dissertação de Mestrado, Departamento de Informática, Escola de Engenharia, Universidade do Minho, Portugal.

[Rubin, 1996] Rubin, D. (1996), Multiple Imputation after 18 years (with discussion), Journal of the American Statistical Association, Vol. 91, pp. 473-489.

[Rud, 2001] Rud, Olivia Parr (2001), Data Mining Cookbook, John Wiley & Sons, Inc., USA.

S

[Santos, 1999] Santos, M. F. (1999), Sistemas de Classificação em Ambientes Distribuídos, Tese Doutoramento, Universidade do Minho.

[Santos, 2001] Santos, M. (2001). Padrão: Um Sistema de Descoberta de Conhecimento em Bases de Dados Geo-referenciadas, Tese de Doutoramento, Universidade do Minho, 2001.

[Santos et al, 2004] Santos, M.F., Cortez, P., Quintela, H., Pinto, F., A Clustering Approach for Knowledge Discovery in Database Marketing, Data Mining 2004 – Accepted for Presentation, Malaga, Spain.

[Sarle, 1995] Sarle, W. (1995), Stopped Training and Other Remedies for Overfitting, Proceedings of the 27th. Symposium on the Interface of Computer Science and Statistics.

[SAS, 2004] SAS, http://www.sas.com, Agosto, 2004.

[Schafer, 1997] Schafer, J. (1997), Analysis of Incomplete Multivariate Data, Chapman & Hall, London.

[Schapire, 2002] Schapire, E.R. (2002), The boosting approach to machine learning: An overview, MSRI Workshop on Nonlinear Estimation and Classification.

[Shim et. al., 2002] Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R., Carlsson, C. (2002), Past, presente, and future of decision support technology, Decision Support Systems, 33, pp. 111 – 126.

[Silva et al., 2003a] Silva, A., Pereira, J., Santos, M., Gomes, L., Neves, J. (2003), Organ Failure Prediction Based on Clinical Adverse Events: A Cluster Model Approach, Artificial Intelligence and Applications (AIA'2003), 695-700.

[Silva et al, 2003b] Silva, A., Cortez, P., Santos, M., Gomes, L., Neves, J., (2003), Organ Failure Diagnosis by Artificial Neural Networks.

[Silva et al., 2004] Silva A., Cortez, P., Santos, M. F., Gomes, L., Neves, J. (2294) Multiple Organ Failure Diagnosis Using Adverse Events and Neural Networks (2004), In Seruca et al. Eds., Proceedings of 6th International Conference on Enterprise Information Systems – ICEIS 2004, Vol. 2, 401-408.

[Sousa, 2004] Sousa, C.F. (2004), Data Mining: Metodologias, Tecnologias, Modelos e Aplicações, Dissertação de Mestrado, Departamento de Sistemas de Informação, Escola de Engenharia, Universidade do Minho, Portugal.

[Spillers, 1966] Spillers W.R. (1966), Artificial Intelligence and Structural Design, J. Structural Division, Proc. ASCE 92, N.º ST6, 591-497.

[SPSS, 2004] SPSS Inc., http://www.spss.com, Outubro, 2004.

[Sun, 2005] Java Technology, http://www.sun.com/java/, Janeiro, 2005.

T

[Takeda e Mikami, 1972] Takeda, H. e Mikami, J. (1972), Ultimate strength of plate girder with varying depth loaded in shear, Journal of Structural Engineering, Vol. 33, pp. 115-126.

[Thrun, 1995] Thrun, S. (1995), Extracting rules from artificial neural networks with distributed representations, Advances in Neural Information Processing Systems, N.º 7, pp. 505-512.

[Torgo, 1998] Torgo, L. (1998), Árvores de Regressão – Métodos e Aplicações, Apontamentos Pedagógicos, LIACC, FEP, Porto.

[Torgo, 1999] Torgo, L. (1999), Inductive Learning of Tree-based Regression Models, PhD thesis, Universidade do Porto, Portugal.

Z

[Zain et al., 2005] Zain, M.F.M., Islam, M., Basri, I.R. (2005), Na expert system for mix design of high performance concrete, Adavances in Engineering Software, 36, pp. 325-337.

[Zárate, 2002] Zárate, A.V. (2002), Un modelo para el dimensionamento de vigas armadas de inercia variable de alma esbelta, Tesis Doctoral, Departamento de Ingeniería de la Construcción, Universidad Politécnica de Cataluña, Barcelona, Espanã.

W

[W3Ca, 2004] HyperText Markup Language (HTML) - W3C HTML Home Page, http://www.w3.org/MarkUp/, Novembro, 2004.

[W3Cb, 2004] Extensible Markup Language (XML) – W3C, http://www.w3.org/XML/, Novembro, 2004.

[W3Cc] World Wide Web Consortium, http://www.w3.org, Novembro, 2004.

[Weiss e Provost, 2001] Weiss, G., Provost, F. (2001), The effect of class distribution on classifier learning: na empirical study, Technical Report ML-TR-44, Department of Computer Science, Rutgers University.

[Weka, 2004] Weka 3 – Data Mining with Open Source Machine Learning Software in Java, http://www.cs.waikato.ac.nz/ml/weka/, Dezembro, 2004.