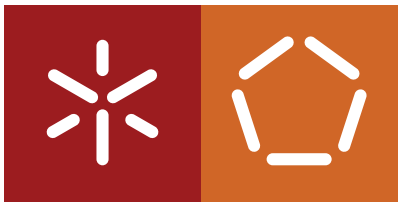
Universidade do Minho
Escola de Engenharia
Departamento de Informática

Nuno Rafael Amorim Gonçalves

Sistema de Suporte à Decisão na Área da Saúde
Predisposição de Acidentes Vasculares Cerebrais

October 2015



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Nuno Rafael Amorim Gonçalves

Sistema de Suporte à Decisão na Área da Saúde
Predisposição de Acidentes Vasculares Cerebrais

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho efetuado sob orientação de
José Manuel Ferreira Machado

October 2015

AGRADECIMENTOS

Foi com o apoio de várias pessoas que este percurso enriquecedor, embora fatigante e árduo, repleto de conhecimento e interação com excelentes detentores de sabedoria, culminou neste trabalho de dissertação de mestrado. A estas devo uma palavra de carinho e agradecimento pelo esforço e dedicação mas também pela paciência e compreensão que tiveram para comigo. No entanto gostaria de exprimir algumas palavras de reconhecimento para aquelas, que de uma forma mais direta, contribuíram para a conclusão deste trabalho.

ao Professor Doutor José Manuel Ferreira Machado pela aceitação e disponibilidade na orientação deste trabalho mas também pelo incansável e incondicional acompanhamento ao logo deste período, pelo espírito crítico, metodologias partilhadas e técnicas de desenvolvimento, confiança depositada desde o início do trabalho mas também pelo suporte em todos os passos na estrutura e literatura facultada que estão na base deste trabalho.

à Tânia pela ternura e apoio, caloroso e inqualificável, e carinho das palavras mas também pela compreensão inestimável necessária nos momentos de maior trabalho, tendo uma enorme capacidade de suavizar e acalmar mesmo quando, à força, viveu todo este projeto. Por outro lado, aos pais da Tânia que me acolheram generosamente no seu lar enumeras vezes, sempre prontos a ajudar.

à minha família, que apesar de estar no fim representa um pilar enorme na minha educação, que em todos os momentos esteve presente e me apoiaram, principalmente aos meus pais e à minha irmã que, ao longo de todo este trabalho conseguiram encorajar, incentivar e nunca desistir de acreditar no meu trabalho e na minha capacidade, pois a eles devo tudo aquilo que sou hoje, porque sem o seu suporte nada do que aqui será desenvolvido seria possível.

ABSTRACT

Stroke diseases have severe consequences for the patients and for the society in general. Despite the recent decreases in mortality rates across the world, cardiovascular diseases remain the leading cause of death among the world, claiming about 5000 lives per day in Europe. In 2010, a study made by The Global Burden of Disease estimated that 29.6% of all deaths worldwide (15 616.1 million deaths) were caused by cardiovascular disease. Also, it was estimated that cardiovascular diseases costs the EU economy almost 196 billion a year (European Society of Cardiology, European Heart Network. European Cardiovascular Disease Statistics, 2012 Edition). In particular, elderly people have a higher risk of stroke, with almost 80% of strokes occurring in individuals over 60 years of age, and at an earlier age than in women, although women are catching up fast (in fact more women than men die from heart incidents). On the other hand these facts reveal that it is extremely important be cautious about this kind of diseases. Although this would seem a really threatening disease, leaving aside this facts and being aware of how critical is the early diagnosis, could save lives, since, if we focus on the most important percentage number, 80% of premature heart diseases and strokes are preventable.

This work will focus on a decision support system development, in terms of its knowledge representation and reasoning procedures, under a formal framework based on Logic Programming which intends to prevent these events to happen with an approach of segmentation methods, which allows to distinguish and aggregate clusters of historical records, classification methods, such as Artificial Neural Networks capable of classifying a new record according with its distribution among the clusters, and a Case Based Reasoning Multi-Agent System that evaluates the record returning a similar case.

Keywords: Data Mining, Artificial Intelligence, Healthcare, Machine Learning, Multi Agent Systems, Stroke Disease, Case Based Reasoning, Data Warehouse, Business Intelligence, Logic Programming.

RESUMO

Doenças como Acidentes Vasculares Cerebrais, de uma forma geral, têm graves consequências para os pacientes e para a sociedade. Apesar da recente diminuição das taxas de mortalidade em todo o mundo, as doenças cardiovasculares continuam a ser a principal causa de morte em todo o mundo, reivindicando cerca de 5000 vidas por dia na Europa. Em 2010, um estudo realizado por The Global Burden of Disease estimou que 29.6% de todas as mortes mundiais (15 616.1 milhões de mortes) foram causadas por doenças cardiovasculares. Além disso, as doenças cardiovasculares custam à economia europeia cerca de 196 biliões (European Society of Cardiology, European Heart Network. European Cardiovascular Disease Statistics, 2012 Edition). Em particular, pessoas idosas têm um maior risco de sofrerem acidentes vasculares cerebrais, com quase 80% destes eventos ocorrendo em indivíduos acima dos 60 anos de idade, e numa idade mais cedo em mulheres, porém estas têm vindo a recuperar rapidamente (na verdade, morrem mais mulheres do que homens em incidentes cardíacos). Por outro lado estes factos revelam que é extremamente importante ter atenção em relação a estes tipos de doença. Apesar desta doença ser realmente alarmante, deixando de lado estes factos, estar ciente de quão critico é o diagnóstico precoce, pode salvar enumeras vidas uma vez que, se nos concentrarmos sobre o número percentual mais importante, 80% das doenças cardíacas e acidentes vasculares cerebrais prematuros são evitáveis.

Este trabalho tem como objetivo o desenvolvimento de um sistema de suporta à decisão, em termos da sua representação do conhecimento dos procedimentos de raciocínio, sob uma ferramenta baseada em Programação em Lógica que pretende prevenir estes eventos de acontecer, seguindo uma abordagem através de métodos de segmentação, que permitem distinguir e agregar grupos de registo históricos, métodos classificativos, tal como Redes Neuronais Artificiais capazes de classificar um registo de acordo com a sua distribuição entre os diferentes grupos, e um sistema multiagente que utiliza métodos de raciocínio baseado em casos para avaliar e retornar casos similares.

Palavras-chave: Data Mining, Inteligência Artificial, Saúde, Machine Learning, Sistemas Multi-Agente, Ataque Vascular Cerebral, Raciocínio Baseado em Casos, Data Warehouse, Business Intelligence, Programação Lógica.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ACRÓNIMOS

AVC Acidente Vascular Cerebral

BI *Business Intelligence*

DOC *Degree of Confidence*

ETL *Extract, Transform and Load*

RNA Rede Neuronal Artificial

FIPA *Foundation for Intelligent Physical Agents*

QOI *Quality of Information*

RBC Raciocínio Baseado em Casos

SWOT *Strengths, Weaknesses, Opportunities and Threats*

TCP/IP *Transmission Control Protocol/Internet Protocol*

1

INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Devido à importância que a informação representa para as organizações, a sua gestão e a devida utilização para processos de tomada à decisão têm vindo a ser valorizada ao longo dos tempos e tornou-se um processo comum dentro das organizações.

Dia após dia, a quantidade de informação armazenada por uma organização tende a aumentar gradualmente levando a um nível de incapacidade de análise e extração de conhecimento sobre essa mesma informação. Esta controvérsia incide sobre a necessidade de analisar, planear e reagir a determinados eventos que é fulcral para a evolução de uma organização. Devido à enorme evolução tecnológica, organizações na área da saúde têm vindo a desenvolver e a adotar sistemas complexos que suportam a sua atividade diária e enormes quantidades de dados relativamente a tratamentos e pacientes. Os sistemas operacionais que possuem esta base de conhecimento não têm a capacidade de relacionar situações passadas e reagir atempadamente de acordo com as necessidades dos utilizadores. Tal relacionamento possibilita a chegada de informação útil para o suporte à decisão. Perante esta situação, com o objetivo de conciliar a informação histórica, tornar os tratamentos de saúde mais eficazes e eficientes e proporcionar uma resposta rápida, surge a necessidade de minudenciar e transformar este conjunto de dados em informação. É também importante, com base num foco direcionado ao consumidor, integrar este conhecimento, culminando em contenção de custos, conformidade e na elaboração de processos mais complexos sem prejudicar os fatores principais dos cuidados de saúde.

Com a integração dos dados clínicos, as organizações conseguem providenciar ao seus pacientes perceções precisas com ênfase na área da performance e qualidade. Uma adoção de processos e uma solução de *business intelligence* facilita o processo de tomada de decisão possibilitando resultados evidentes na redução de custos, aumento significativo na satisfação dos seus pacientes e, um dos fatores mais relevantes, na otimização e melhor utilização dos seus dados históricos juntamente com uma melhor previsão de diagnósticos.

Na saúde, a utilização de técnicas de *data mining* tem vindo a crescer exponencialmente uma vez que esta possibilita a descoberta de padrões e tendências em diversos contextos. A utilização de técnicas de *data mining* utilizam dados que, aparentemente, não se relacionam entre si e molda-os

e ajusta-os para que sejam interpretados como um conjunto de informação útil para suportar um processo de tomada de decisão.

Com a aplicação destes modelos preditivos, que representem a realidade o mais próximo possível, é esperado que as organizações de saúde consigam redirecionar os seus recursos de forma a reduzir custos e manter ou melhor a tratamento dos pacientes.

1.2 MOTIVAÇÃO E OBJETIVOS

O processo de tomada de decisão envolve o desenvolvimento e análise de alternativas de decisão em resposta a um determinado problema. Em 1950, Hebert Simon e James March introduziram um ferramenta de tomada de decisão para a interpretação do comportamento organizacional. Com base neste estudo, foi possível caracterizar os tipos de decisão em programadas e não programas. A decisões programadas são repetitivas, com uma determinada rotina, estabelecendo regras concretas de acordo com diretrizes das organizações para encontrar a melhor solução. Por outro lado decisões não programadas são excepcionais e não recorrentes.

Os sistemas de suporte à decisão podem assistir os decisores no processamento, acesso, categorização e organização da informação de uma forma fácil. O potencial para melhorar a aprendizagem organizacional é um dos principais fatores destes sistemas, facilitando e promovendo capacidades de decisão e modelação eficazes e eficientes.

Na saúde, o principal objetivo direciona-se para a qualidade, gestão de risco, redução de custos e produtividade.

Existem diversas vantagens no uso de sistemas de suporte à decisão para discutir opção relativamente ao tratamento de um paciente. O fato de colocar num contexto de forma explicita e analítica um determinado problema, leva a pressupostos mais simples e concretos. No entanto, para que haja sucesso na implementação destes sistemas, deve existir a consideração da aceitação dos sistemas por parte dos profissionais de saúde, como entidades importantes no processo de adaptação e tomada de decisão.

O principal objectivo deste trabalho passa pela identificação de características comuns entre os diversos diagnósticos, armazenados numa memória de casos, e um novo caso, para que seja possível auxiliar o processo de tomada de decisão, facilitando o tratamento do paciente e reduzir os custos sem por em causa os cuidados de saúde prestados. A implementação do sistema descrito no documento utiliza diversos conceitos constituintes de um tipico sistema de *business intelligence*.

1.3 METODOLOGIA

Existem alguns conceitos que representam a atividade principal e mais importante desta análise. O conceito base consiste na identificação de grupos diferentes de dados num sistema multiagente. Cada agente tem o seu próprio conhecimento de dados, tornando-os diferentes uns dos outros, o que afeta

a decisão durante o processamento do caso. O primeiro método aplicado ao conjunto de dados de cada agente é o método de *clustering*. Este método envolve a identificação e segmentação dos registos idênticos relacionando os atributos e funções de distância. Assim, um *cluster* é um grupo de registos idênticos entre eles e consequentemente diferentes entre outros grupos. Uma vez segmentado, cada registo é associado ao seu grupo correspondente, definindo o atributo classe do registo.

A segunda metodologia consiste na técnica de classificação utilizando Redes Neuronais Artificiais. A RNA recebe n tuplos, onde n é o numero de entradas, e cada tuplo é composto por três medidas onde uma delas é o grau de confiança (DoC) da entrada correspondente. O objetivo deste método é classificar o sintoma novo com um número de um conjunto de dados gerado com base no conhecimento dos dados de cada agente.

Estas metodologias integram a primeira etapa do Raciocínio Baseado em Casos permitindo ao sistema identificar, com mais precisão, o *cluster* com casos semelhantes. O caso recuperado é obtido com a aplicação de uma expressão matemática calculando a similaridade dentro do próprio *cluster*. Após este processo, o sistema é capaz de apresentar ao utilizador uma solução adequando, tão próximo quanto possível, um tratamento similar, que pode fazer a diferença na recuperação do paciente.

Relativamente à metodologia de investigação foi seguida uma abordagem investigação-ação, um processo composto de várias fases que se desenvolvem de forma contínua identificadas por planificação, ação, observação (avaliação) e reflexão, respetivamente. Trata-se de uma metodologia de pesquisa que procura resolver problemas reais atuando e analisando o seu universo de discurso e desenvolvendo mudanças nas práticas com o objetivo de melhorar os resultados obtidos. Em suma, esta metodologia de investigação é orientada para a melhoria da prática nos diferentes campos de ação (Jaume Trilla, 1998 e Elliott, 1996).

De acordo com Coutinho et al (2009) o conjunto de procedimentos circular inicia novos ciclos que desencadeiam novas esperais de experiências de ação reflexiva.

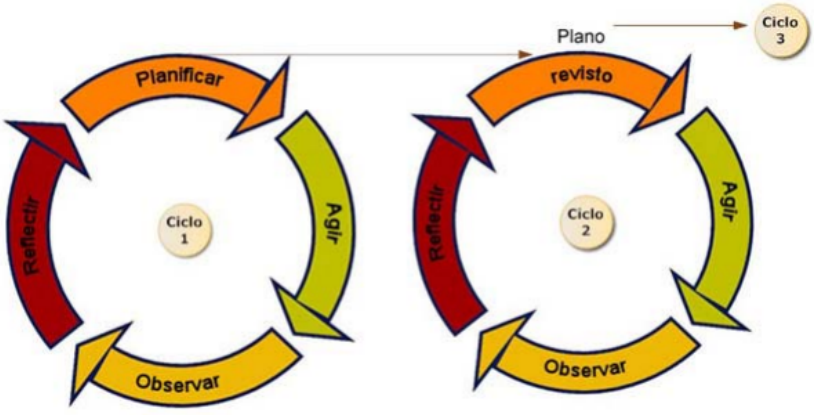


Figura 1: Espiral de ciclos da Investigação-Ação (Coutinho et al., 2009:366)

1.4 TRABALHO REALIZADO

Este modelo de investigação e todo o seu desenvolvimento foram apresentados na conferência internacional sobre sociedade da informação (e-Society 2015), no dia 15 de Março de 2015, de uma forma resumida foram contextualizados e apresentados alguns processos neste documento desenvolvidos [50].

1.5 ESTRUTURA DO DOCUMENTO

Este documento procura apoiar a tomada de decisão através do auxilio de um sistema cuja representação do conhecimento e raciocínio é feita através de um programa lógico, armazenando em memória de casos acontecimentos ocorridos que possam ser reutilizados em situações futuras. Contudo, o sistema apresenta várias técnicas que possibilitam a integração e manipulação da informação possibilitando um funcionalidade fluido e concisa de todo o sistema multiagente. O trabalho esta estruturado em sete capítulos principais, passando pela contextualização da problemática até à implementação e avaliação de resultados.

Parte 2: Estado da Arte

Antes de apresentar todos os procedimentos envolventes no sistema sugerido, é feita uma contextualização sobre temáticas relacionadas com a funcionalidades do sistema. Com o estado da arte pretende-se dar a conhecer os conceitos base que fundamentam as técnicas utilizadas ao longo do desenvolvimento do trabalho. Esta parte contempla os capítulos 2 a 7.

Parte 3: Trabalho Desenvolvido

Capitulo 8: Representação do Conhecimento e Raciocínio

A forma como toda a informação é tratada e interpretada no sistema é apresentada no terceiro capitulo. Através de programação lógica e da definição de predicados, contemplamos um processo de normalização e de representação do conhecimento e raciocínio. Neste capítulo são apresentadas duas medidas que fazem parte do processo da seleção de casos, *Degree of Confidence* (DoC) e *Quality of Information* (QoI).

Capitulo 9: Caso de Estudo

Com todos os conceito introduzido no capítulo 2 e a forma como a representação de conhecimento e raciocínio é feita, no quarto capítulo é elaborado um caso de estudo especifico já com dados relativos ao problema apresentado seguindo todos os passos a considerar na implementação do sistema.

Capitulo 10: Arquitetura

De um modo geral, é apresentado neste capítulo a estrutura que suporta o sistema bem como a comunicação entre os diversos elementos que o constituem. Além disso, são identificados os tipos de agente do sistema multiagente.

Capitulo 11: Implementação

Após toda uma fase introdutório, contextualização da problemática e identificação dos mecanismos utilizados, no capitulo 11 são abordadas as diretrizes seguidas na implementação do sistema de uma forma mais detalhada e técnica.

Capitulo 12: Análise de Resultados

Após a implementação de todos os processos, o capítulo 7 é utilizado como ponto de controlo para refletir sobre os resultados apresentados e produzidos, como produto da implementação de toda a lógica e técnicas aplicadas no sistema, interpretando-os e comparando-os com os casos passados para que seja possível determinar, o mais aproximadamente possível, a viabilidade do sistema.

Capitulo 13: Conclusões e Trabalho Futuro

Finda-se o trabalho com alguns aspetos a considerar como trabalho futuro e melhorias para uma melhor *performance* e precisão dos modelos usados. É também avaliada a utilização dos algoritmos e técnicas utilizadas em todo o documento.

Parte I

ESTADO DA ARTE

2

ACIDENTES VASCULARES CEREBRAIS

Existem diversos tipos de doenças cardíacas, no entanto, cada ano, 15 milhões de pessoas em todo o planeta sofrem de acidentes vasculares cerebrais deixando 5 milhões permanentemente incapacitados e causando 6 milhões de mortes. Apesar dos acidentes vasculares cerebrais representarem a segunda maior causa de morte em todo o mundo também é considerada a segunda em termos de incapacidade, incluindo perda de visão, fala, paralisia ou confusão. Contudo estes factos são considerados para pessoas com idade superior a 60 anos, baixando para o quinto lugar mundial para pessoas entre os 15 e 59 anos. Em países demograficamente desenvolvidos, a idade média em que uma acidente vascular cerebral pode ocorrer é por volta dos 73 anos com uma probabilidade de 1.6 em 1000 (Truelsen1, T. et. al.). A figura seguinte constitui a análise realizada a nível mundial com uma distribuição geográfica relativamente a mortalidade provocado por acidentes vasculares cerebrais *(World Health Organization Global Burden of Disease Program, 2004)*

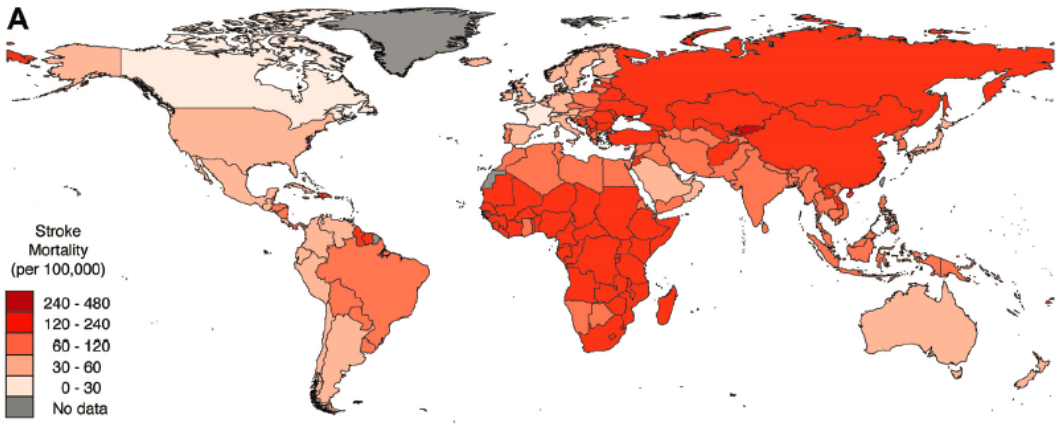


Figura 2: Taxas de mortalidade de AVC por 100 000 habitantes *(World Health Organization Global Burden of Disease Program, 2004)*

De uma forma geral, este acontecimento surge quando o fornecimento de sangue ao cérebro é impedido podendo matar ou incapacidade a vitima. É possível caracterizar um acidente vascular cerebral em isquêmicos, responsáveis por cerca de 80% destes acidentes, em que por causa de coágulos de sangue ou redução do volume das artérias existe um bloqueio do fluxo sanguíneo, e hemorrágicos, quando uma artério explode.

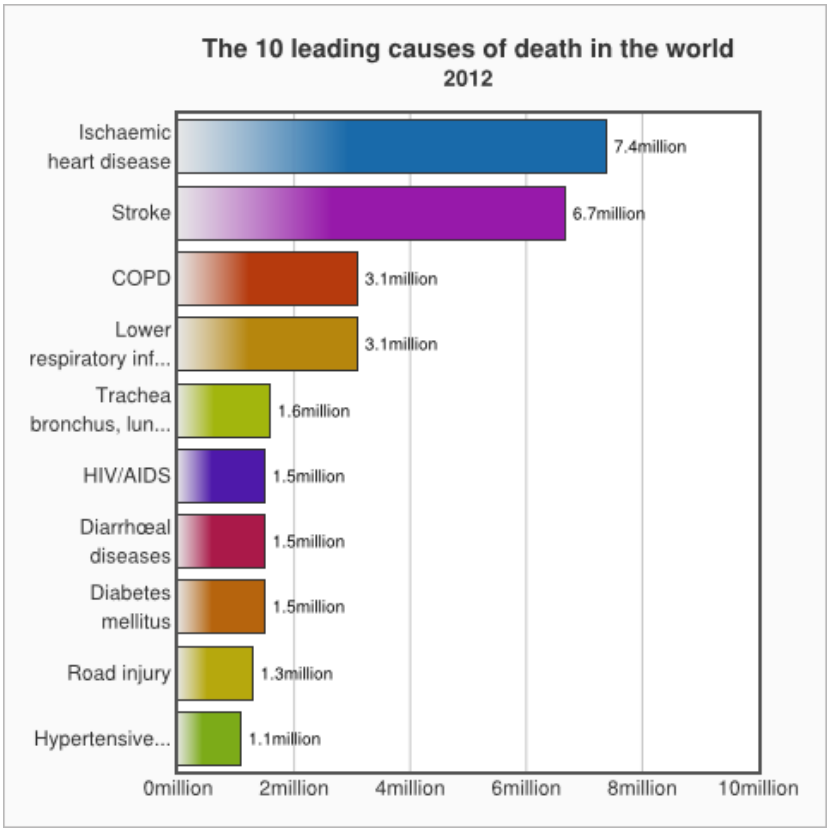


Figura 3: Principais causas de mortalidade no mundo ((*World Health Organization Global Burden of Disease Program, 2004)*

Porém, hábitos de vida não saudáveis e comportamentos que constituem fatores de risco graves são considerados os principais motivos de mortalidade e incapacidade que influenciam bastante este tipo de acontecimentos. Entre os fatores de risco associados com acidentes vasculares cerebrais, há aqueles que são considerados como perigosos. Alguns podem ser tratados ou controlados, como pressão arterial elevada (Go, et al. 2014; Lindgren, 2014), fumador (Bhat et al, 2008; Shah and Cole, 2010; Lindgren, 2014), diabetes *mellitus* (Khoury et al, 2013; Lindgren, 2014), colesterol elevado (Amarenco et al, 2008; Zhang et al, 2012; Lindgren, 2014) e inatividade física (Grau et al, 2009; McDonnell et al, 2013).

Contudo existem alguns que não podem ser controlados, como idade, na medida em que pessoas idosas tem mais tendência para acidentes vasculares cerebrais (Sealy-Jefferson et al, 2012; Go et al. 2014; Lindgren, 2014), género, mais comum em homens do que mulheres, apesar de morrerem mais mulheres do que homens, o simples facto de ter sufrido um AVC representa um fator de risco elevado não controlado (Kissela et al, 2012; Lindgren, 2014), etnia (Kissela et al, 2012; Sealy-Jefferson et al, 2012; Lindgren, 2014), entre outras.

Considerado um dos principais fatores relacionados com a mortalidade, esta doença é seguida de perto com o objetivo principal de impedir que aconteça uma vez que, quando diagnosticada, torna-se menos perigosa e mais tratável, comparada com doenças similares (Go et al, 2014).

3

BUSINESS INTELLIGENCE

O comportamento das organizações tem vindo a sofrer alterações drásticas no modo como estas lidam com a informação e sua comunicação para apresentação de resultados. Por um lado está a enorme capacidade de recolha e geração de dados por parte de cliente e/ou mesmo da própria organização utilizada em processos complexos com uma difícil capacidade de compreensão abrangente do ambiente de negócio. Durante vários anos, o termo *business intelligence* evolui significativamente dentro da área dos sistemas de suporte à decisão.

Foi no ano de 1958, que Hans Peter Luhn apresentou os potenciais de BI com apoio de uma definição de inteligência do dicionário "Websters Dictionary", "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal". O artigo, intitulado "A Business Intelligence System"menciona um sistema automático capaz de disseminar informação de várias secções e ainda utilizar máquinas de processamento de dados para abstração e codificação de documentos com perfis de interesse para cada área de ação de uma organização. Após um grande salto evolutivo na área da tecnologia Howard Dresner, uma analista da Gartner definiu o conceito de *business intelligence* como sendo uma variedade de tecnologias que suportam o acesso ao utilizador final para análises de informações quantitativas. Durante os anos seguintes, muitas foram as definições de BI dadas por vários autores tal como (Golfarelli et.al, 2004) que definiu BI como um conjunto de *data warehouse* e também componentes reativos capazes de monitorizar o tempo de processos operacionais críticos para permitir uma sincronização entre a tomada de decisão tática e operacional com a estratégia da empresa. (Kimball et.al. 2013) considera que um sistema de BI deve fazer com que a informação seja fácil de aceder, apresentar informação consistente, adaptar-se a mudanças, apresentar informação sob forma temporal e também um sistema capaz de servir de base autoritário e confiável para uma melhor tomada de decisão. (Stackowiak et.all 2007) considera que BI pode ser definido como o acesso correto a dados ou informação necessários para realizar a melhor decisão de negócio no devido momento. No fundo, todos eles consideram o processo de tomada de decisão um ponto fulcral nos sistemas de BI. Podemos interpretar o termo *business intelligence* como uma área cujo principal objetivo passa pela aquisição e manipulação de um conjunto de informação estruturada e não estruturada para que seja fornecida em forma de informação viável e conhecimento aos diversos agentes de tomada de decisão através de ferramente de processamento analítico. No entanto, este termo é frequentemente interpretado de diferentes formas dependendo do contexto apli-

cado. Muitas organizações definem *business intelligence* como um produto, ferramenta, conhecimento ou informação imprescindível para um bom funcionamento do processo de tomada de decisão das organizações que o integram. Por outro lado esta área surge ao lado das soluções implementadas pelas organizações para o processamento de dados dos seus sistemas operacionais bem como a sua representação como forma de transmitir conhecimento e informação acerca da performance da sua organização que facilite o processo de decisão.

Embora estas definições sejam utilizadas, em alguns casos, separadamente, elas complementam-se na medida em que os produtos usados para extrair informação e conhecimento dos sistemas operacionais das organizações são o motor de uma solução típica de *business intelligence* que no seu todo definem esta área como um conjunto agregador de processos, tecnologias, arquiteturas e representações de informação e conhecimento para que no seu conjunto sejam capaz de responder ao principal objetivo, suportar e fundamentar o processo de tomada de decisão.

3.1 ARQUITETURA DE SISTEMAS BI

O sistemas de *business intelligence* são de várias tarefas principais, tal como exploração avançada de dados, integração, agregação, análise multidimensional proveniente de várias fontes de dados. Contudo o desenvolvimento deste tipo de sistema requer mais tempo de desenvolvimento e análise do que os restantes sistema de gestão de informação pelo facto de serem sistema bastante mais complexos.

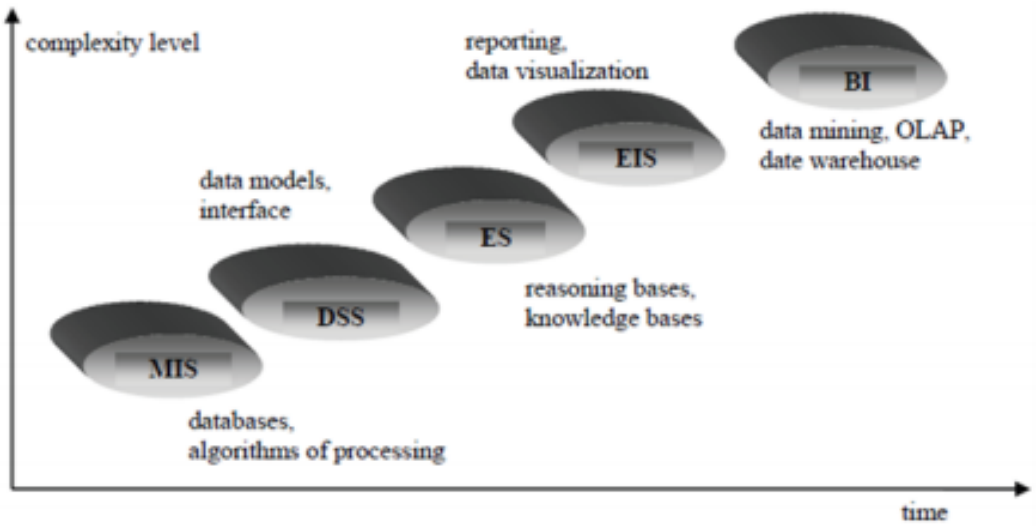


Figura 4: Desenvolvimento de sistemas de gestão de informação (Olszak and Ziemba, 2004)

Seguindo a abordagem de Kimball, existem quatro componentes distintos e separados que constituem o ambiente principal de um sistema de *business intelligence*: sistemas operacionais de origem, sistemas de ETL, área de apresentação dos dados e aplicações de *business intelligence*.

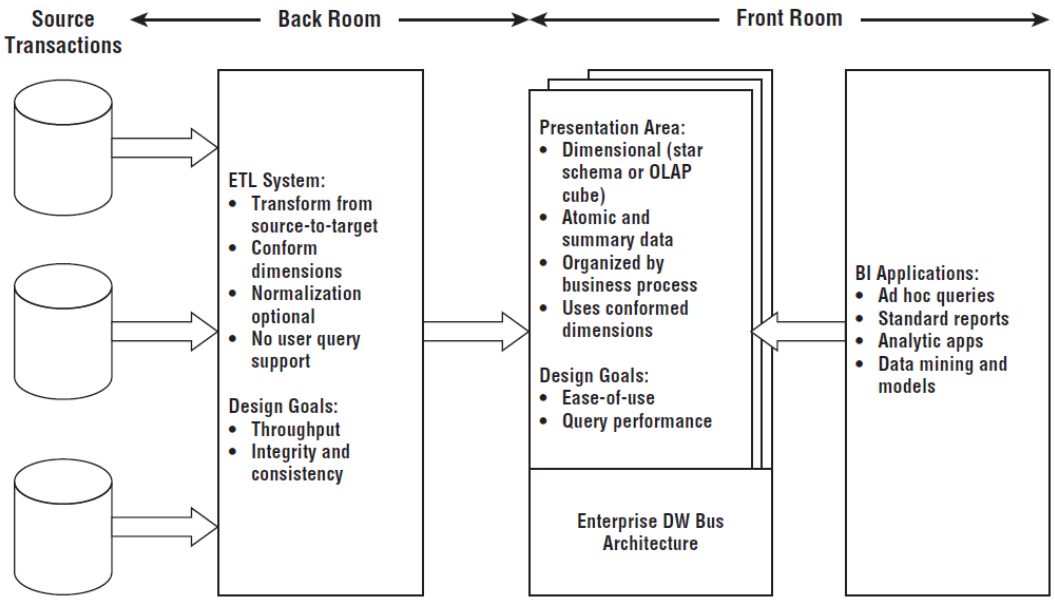


Figura 5: Arquitetura de um sistema BI (Kimball et.al. 2013)

3.1.1 Sistemas Operacionais de Origem

Sistemas complexos concebidos para suportar processamento de transações de negócio pré-definidas (*Online Transaction Processing*) capazes disponibilizar a informação no seu estado mais atual para as transações diárias e processar rapidamente ações simples mantendo a integridade do sistema. Estes sistemas são considerados como sendo um fator externo ao *data warehouse* uma vez que não existe nenhum ou escasso controlo sobre os mesmos. Uma vez que estes sistemas mantêm poucos dados históricos, uma boa implementação dos componentes seguintes reduz significativamente a responsabilidade da representação de dados históricos.

3.1.2 Sistemas ETL (*Extract, Transformation, and Load*)

A modelação do processo de ETL (*Extract Transform Load*) engloba uma combinação complexa de processos e tecnologias que consomem uma parte significativa do esforço de desenvolvimento de um sistema de BI e requer as habilidades dos arquitetos de *data staging* juntamente com o administrador do *data warehouse*, para a parte do desenho e planeamento do ETL, e dos programadores da *data staging* para a parte da implementação. Este componente estabelece a ponte entre os sistemas operacionais de origem e a área de apresentação de dados. Sendo este um processo crucial no desenvolvimento de um sistema de *data warehousing* é bastante importante planeá-lo e desenhá-lo antes de dar inicio à fase de implementação, detalhando todas as suas etapas, começando na extração, passando depois pelos processos de transformação e limpeza dos dados e por fim fazer o carregamento destes para o *data warehouse*. Depois de desenvolvido e implementado, este processo, não é um evento

de um tempo único, pois serão adicionados novos dados ao *data warehouse* periodicamente, daí a importância de bom desenho e planeamento antes de começar a desenvolver o ETL.

A primeira fase, extração, consiste no processo de obter os dados, através da definição e compreensão de meta dados, provenientes de um ou vários sistemas operacionais, para o área de retenção onde serão transformados utilizados diversas técnicas de limpeza e enriquecimento de dados, tais como, correção de palavras, interpretação de valores desconhecidos e não duplicação de dados, constituindo a segunda fase do processo de ETL. Por ultimo, os dados são carregados para as estruturas multidimensionais, um processo bastante critica uma vez que deve deixar o sistema num ponto de integridade e consistência.

3.1.3 Área de Apresentação de Dados

Tudo que o utilizador tem acesso através das aplicações de *business intelligence* está na área de apresentação. Uma estrutura que apresenta e armazena os dados em esquemas dimensionais, organizando-os e disponibilizando-os para consulta direta aos seus utilizadores. Por isso, esta área de apresentação e a consulta dos seus dados devem ser estruturados de acordo com as necessidades do negócio e suas medidas em vez de serem direcionados para um único departamento. Uma vez que existe uma abstração, em termos físicos, deste componente com os sistemas operacionais, este deve alinhar o seus modelo de acordo com recolha e definição dos dados dos componentes anterior. No entanto, estes sistema devem manter os seus dados de acordo com determinadas heurísticas (W. H. Inmon 2002).

- **Orientados ao Assunto**: A estrutura deve ser orientada para as principais áreas da organização que foram definidos na análise dos sistemas operacionais e ambiente da organização.

- **Integrados**: Um dos principais aspetos destes sistemas, uma vez que os dados convergem de várias origens num só esquema, é serem integrados incorporando a capacidade de converter, reformatar e sumarizar os dados para que a imagem final da organização seja única e concisa.

- **Não-volátil**: Normalmente, os dados num sistema operacional são atualizados de forma transacional representando o ultimo estado da organização, porém, na área de apresentação os dados são carregados e consultados em grandes quantidades mas, de forma geral, não atualizados sendo considerados carregamentos estáticos. Esta característica permite armanezar dados históricos.

- **Variante no tempo**: Implica que todas a unidades dentro destes sistemas seja precisa num determinado momento dependendo explicitamento de fatores temporais. Existe sempre, independentemente do contexto, algo que identifique temporalmente o registo numa linha temporal.

3.1.4 **Aplicações de Business Intelligence**

Como suporte aos dados que a área de apresentação providencia, as aplicações de BI auxiliam o processo analítico utilizado na tomada de decisão. Estas ferramentas dão uso aos dados presentes no componente anterior sob forma de consultas podendo resultar em consultas simples ou até mesmo processo mais complexos como data mining.

4

MACHINE LEARNING

Com a utilização da computação natural existe a necessidade de adaptar aos sistemas computacionais comportamentos característicos dos seres humanos como a aprendizagem. A aprendizagem é fulcral para o conhecimento e inteligência humana, do mesmo modo, também se torna essencial na construção de sistemas inteligente para a resolução de problemas fornecendo a capacidade de adaptação no ambiente, ou seja, o sistema aprende através da recolha e adaptação da informação recolhida ou soluções passadas, adequando a sua futura decisão de acordo com os acontecimentos sucedidos e informação retida. “Computer program that improves its performance at some task through experience.” Tom Mitchell (1997 - Machine Learning).

A aprendizagem por computador representa a capacidade que os sistemas possuem em aprendem automaticamente a partir da informação baseada em experiências passadas, a qual é previamente fornecida, treinada, para possibilitar a realização de predições sobre informação futura. Devido à sua enorme potencialidade esta tecnologia tem vindo a ser valorizada em diversas áreas do conhecimento.

Os métodos de aprendizagem por computador são divididos em duas fases principais:

1. Treino: Esta é executada a partir de um conjunto de regras e dados previamente conhecidos, ao qual se atribui a designação de algoritmo de aprendizagem ou treino, para formação de um modelo.

2. Aplicação: Dependente da fase anterior, o modelo gerado é aplicado para tomar decisão relativamente a dados futuros, novos.

TIPOS DE APRENDIZAGEM

Tendo em conta a primeira fase dos métodos, existem vários tipos de aprendizagem que distinguem se distinguem pela maneira como a própria aprendizagem é feita.

- **Aprendizagem supervisionada**: Na aprendizagem supervisionada todos os registos contêm um atributo classe (atributo que classifica o registo, corretamente identificado). Esta aprendizagem dá uso aos valores de entrada (atributos não classe) e o valor de saída (atributo classe) associando-os a um modelo de aplicação nos processos de predição. Os tipos de aprendizagem supervisionada mais comuns são a classificação e a regressão.

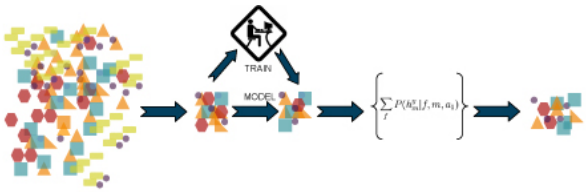


Figura 6: Modelo do processo de aprendizagem supervisionada

- **Aprendizagem não supervisionada**: Em contrapartida da supervisionada, a aprendizagem não supervisionada não possui o atributo classe, apenas os valores de entrada. Devido à falta atributo classificador no registo esta tenta descobrir padrões ou fatores de similaridade nos dados que seja pertinentes à análise. Posto isto, o *clustering* é uma das técnicas mais utilizadas quando estamos perante esta situação.

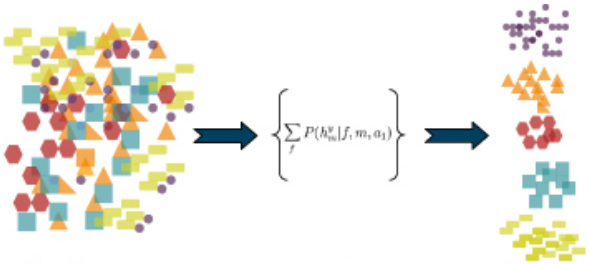


Figura 7: Modelo do processo de aprendizagem não supervisionada

- **Aprendizagem por reforço**: Esta aprendizagem é do carácter individual de uma agente que procurar aprender um determinado comportamento resultante de interações de tentativa erro contextualizado num ambiente específico, ou seja, o agente está ligado ao ambiente através de perceção e ação.

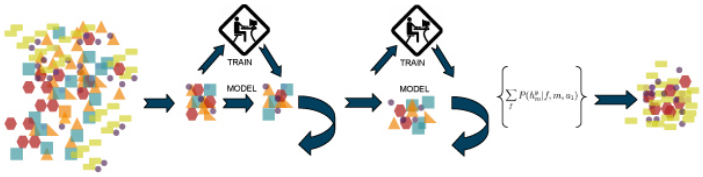


Figura 8: Modelo do processo de aprendizagem por reforço

5

DATA MINING

Consiste, de uma forma sucinta, no processo de encontrar informação pertinente como padrões, associações, mudanças, anomalias e estruturas, tornando este conceito a ferramenta mais utilizada para extração de conhecimento. Este processo é constituído por diferentes etapas:

- **Preparação**: Tudo tem início na fase da preparação onde os dados são devidamente preparados para serem utilizado na fase subsequente. São selecionados os dados importantes retirando aqueles que apresentem inconsistência e pré-processados (representação adequada dos dados para serem submetidos);
- **Data Mining**: Esta etapa consiste na extração de conhecimento, transformando os dados, para que a informação importante seja analisada mais facilmente;
- **Análise de Dados**: É realizada uma análise sobre o resultado obtido na etapa precedente. Esta análise pode ser apresentada, por exemplo, em gráficos com a análise do comportamento do mesmo.

Em data mining podem ser identificadas duas principais categorias distintas de tarefas que diferem no modo em como a sua análise em relação aos dados é feita:

1. *Descriptive data mining*
2. *Predictive data mining*

O primeiro método encara os dados e suas análises realizadas no passado para introspeção de como abordar futuros procedimentos. Esta análise permite determinar qual o motivo de sucesso ou insucesso de procedimentos anteriores. Por outro lado, a análise preditiva transforma os dados em informação importante que por sua vez é usada para determinar um resultado provável de um evento no futuro.

Num sistema de *data mining* é possível realizar diversas tarefas:

- **Descrição de classes**: Esta funcionalidade pode ser integrada numa análise para tornar mais concisa a ideia utilizada, hipótese ou factos observados;
- **Associação**: É a funcionalidade de determinar que dados estão relacionados, ou seja, descobrir regras de associação condicionadas a valores de atributos que ocorrem juntos num conjunto de

dados. Aplica-se aos casos em que se pretende estudar preferências, afinidades com o objetivo de ajustar o problema ao utilizador;

- **Classificação**: Consiste em examinar uma determinada característica nos dados e atribuir uma classe previamente definida. Estes dados podem ser associados a classes ou conceitos através de um processo de caracterização ou de discriminação. A discriminação caracteriza-se pelo resultado obtido através da atribuição de um valor a um atributo no registo, em função de um ou mais atributos. Na caracterização é realizada a sumarização de um atributo em relação a um ou mais atributos. Ao criar uma árvore de classificação os dados podem ser extraídos de forma a determinar a probabilidade de associação de um valor segundo o conjunto de atributos;

- **Previsão**: Os modelos de previsão ou também designados por modelos de regressão respondem normalmente a perguntar numéricas. Este modelo determina prevê os valores de alguns dados ou a distribuição de valores de certos atributos de um conjunto de dados. No fundo, envolve a descoberta de um conjunto de atributos relevantes para o atributo de interesse e prediz a distribuição do valor baseado no valor do conjunto de dados semelhantes ao(s) objeto(s) selecionado(s);

- **Segmentação**: A análise de *clusters* ou de segmentação consiste em identificar possíveis agrupamentos nos dados, onde o agrupamento é um conjunto de objetos que são semelhantes uns aos outros. Diferentes medidas de similaridade, baseadas em funções de distância, podem ser especificadas para diferentes contextos de aplicação;

5.1 CLASSIFICAÇÃO

A classificação consiste, como o próprio nome indica, na determinação da classe dos registos do modelo pretendido, para tal, examina-se as características de cada registo (com classe devidamente preenchida) a fim de, a partir de algoritmos de aprendizagem, analisar os registos fornecidos de modo a treinar o modelo para que posteriormente seja possível classificar novos casos (modelo casos de teste) de acordo com o modelo gerado. Este processo identifica-se com a aprendizagem supervisionada.

Após a tarefa de análise e treino do modelo os registos são classificados como sendo corretamente classificados e incorretamente classificados, através dos quais é gerada a matriz confusão. Esta é simplesmente uma matriz quadrada cuja sua diagonal corresponde às classificações corretas.

5.1.1 Árvores de Decisão

Uma árvore de decisão é um fluxograma (*flow-chat*) semelhante a uma estrutura de árvore, onde cada nó interno denota um teste em um atributo, cada ramo (subárvore) representa o resultado do teste e cada folha representa a distribuição dos registros. A sua utilização recomenda o treinamento do método, utilizando-se várias amostras nos dados, até que se conheça as melhores regras para segmentação do conjunto de dados.

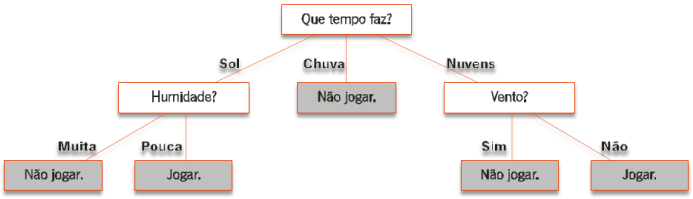


Figura 9: Exemplo de árvore de decisão

5.1.2 Redes Neuronais Artificiais

As redes neuronais artificias são um sistema computacional de base conexionista para a resolução de problemas. Este é concebido com base num modelo simplificado do sistema nervoso central dos seres humanos e definida por uma estrutura interligada de unidades computacionais, designadas como neurónios, com capacidades de aprendizagem.

O neurónio é identificado pela sua posição na rede e caracterizado pelo valor de estado utilizando o axónio como via de comunicação que o pode ligar a qualquer neurónio incluindo o próprio, sendo que este canais de comunicação podem variar ao longo do tempo conduzindo a informação num só sentido. Ao terminal resultante da conexão entre dois neurónios diz-se sinapse no qual o seu valor determina o peso do sinal a entrar no neurónio: **excitativo**, **inibidor** ou **nulo**.

Neurónio Artificial

Como já menciona, um neurónio artificial ou nodo é a unidade de processamento nas operações realizadas numa RNA. Estes neurónios produzem um valor de saída calculado a partir de um conjunto, ou de único, de valores de entrada. Um neurónio artificial pode ser dividido em diversos elementos de um neurónio, desde o conjunto das conexões, um integrador e uma função de ativação.

As conexões associadas a um neurónio são identificadas por um peso que determinada o efeito excitatório de acordo com valores positivos ou negativos. Recebido um sinal de uma determinada conexão, este estímulo é multiplicado pelo peso associado a essa mesma conexão. Complementar às conexões de entrada poderá existir uma conexão extra designada bias permitindo identificar tendências num processo.

De uma geral, existem uma operação que agrega as conexões com o objetivo de reduzir os argumentos de entrada num único argumento para que possa ser parametrizado numa função de ativação devolvendo um valor computacionalmente calculado de modo a que seja transferido como valor de transferência para os neurónios da camada seguinte.

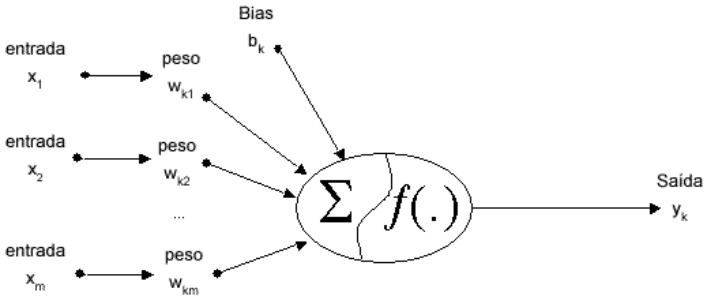


Figura 10: Representação do neurónio artificial

Arquitetura

Numa Rede Neuronal Artificial, um nodo pode não ser suficiente, independentemente da quantidade de valores de entrada. Uma rede neuronal artificial pode ser caracterizada pela sua arquitetura que se distinguem pela distribuição e pela forma como os neurónios se relacionam, define a estrutura.

Redes *feedforward*

Uma Rede Neuronal Artificial é formada por neurónios artificiais em que estes são organizados por camadas e em que a conexão, comunicação, se realiza de uma forma unidirecional entre as diversas camadas. Por norma, estas redes são estáticas produzindo um conjunto de valores de saída ao contrário de uma sequência para um determinado valor de entrada.

Em termos de memórias estas redes não guardam qualquer tipo de informação sobre o estado da rede anterior o que faz com que a sua resposta a um determinado valores de entrada seja independente desse estado anterior.

Camadas múltiplas

As redes com camadas múltiplas contêm uma única camada de saída que são alimentados diretamente por uma camada anterior. Estas camadas anteriores são designadas por camadas escondidas até ao momento de início do processo que se realiza numa única camada de entrada. Cada camada possui nós e cada nó está totalmente interligado por pesos com todos os nós da camada subsequente.

Camada única

Uma rede que apenas contém uma camada de entrada e uma camada de saída em que o seu funcionamento é idêntico à das camadas múltiplas.

Redes recorrentes

Estas redes, ao contrário das *feedforward*, são sistemas dinâmicos na forma em que quando apresentado um determinado padrão de entrada, os valores de transferência de um neurónio são novamente calculados, ou seja, modificados. A comunicação destas redes não tem direção única podendo ser implementados diversos caminhos dependendo do tipo de problema presenteado.

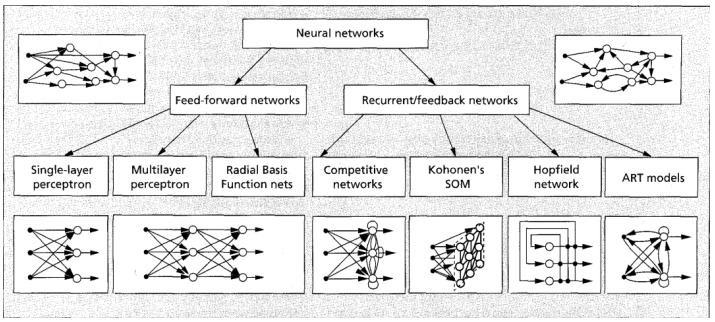


Figura 11: Tipologias de uma Rede Neuronal Artificial

Teorema Bayesiano

A abordagem bayesiana avalia a probabilidade de um registo pertencer a uma determinada classe, dados os valores de atributos observados para o registo. Quando a classificação é do modo categórica e não probabilística a classe com o maior valor estimado é selecionada.

5.2 REGRESSÃO

A regressão é similar à classificação mas trabalha com valores numéricos e o seu objetivo é verificar como o valor de um ou vários atributos (valores de entrada) influenciam o resultado do atributo selecionado (valor de saída) para cálculo. Em áreas médicas, industriais e químicas, entre outras, é necessário saber se um ou mais atributos têm influência entre eles sem ainda estar definido um atributo classe e é aqui que a regressão tem um papel importante.

Em suma a regressão permite prever o valor de um atributo tendo em conta todos os dados antigos para o efeito. A regressão pode ser dividida em duas categorias, sendo elas:

- **Regressão linear simples**: tem como objetivo calcular a relação de apenas um valor de entrada para o valor de saída;
- **Regressão linear múltipla**: tem por base calcular a relação entre vários valores de entrada para um valor de saída.

Neste documento devido ao facto da existência de vários atributos, apenas foi usada a regressão linear múltipla, no entanto vários métodos foram implementados.

5.3 SEGMENTAÇÃO

Clustering (segmentação) consiste na identificação e criação de grupos de registos idênticos. Um grupo (*cluster*) é, então, um conjunto de registos similares entre si e por consequência diferente dos outros grupos. Enquanto na classificação é necessário categorizar os dados previamente, no *clustering* não é necessário existir uma categorização porque tem apenas a finalidade de identificar grupos de dados similares o que faz com que esta técnica se enquadre na aprendizagem não-supervisionada.

Com a utilização desta técnica é possível agrupar registos idênticos segundo uma medida de proximidade os diversos atributos de cada registo. A figura 7 a disposição dos registos segundo dois atributos de três espécies de uma flor.

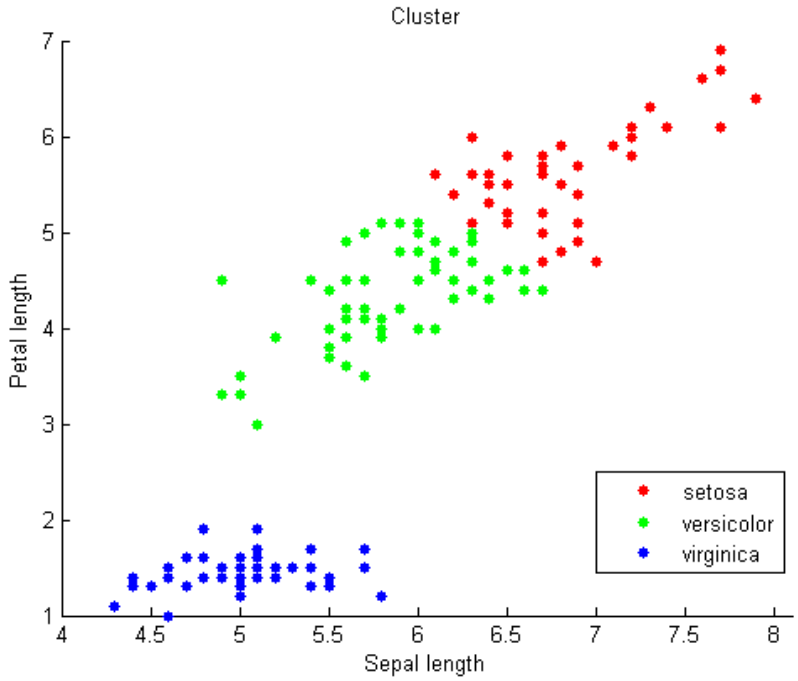


Figura 12: Distribuição dos segmentos segundo o comprimento da pétala e da sépala

6

SISTEMA MULTIAGENTE

Existem várias vertentes na área da Inteligência Artificial, entre elas, o Inteligência Artificial Distribuída que procura relacionar vária entidade independentes, num sistema comum, interagindo com o domínio envolvente. Uma das características deste sistemas passa pela gestão do comportamento e de ações das entidade, também designadas por agentes, que constituem um sistema multiagente, conduzindo a uma resolução de problema paralela com opiniões e resoluções diferentes de acordo com os interesses e objetivos dos agentes. Todo este avanço no âmbito da Inteligência Artificial tem vindo a ser possível devido á evolução das técnicas utilizadas tornando estes sistemas mais robustos e complexos. Por isso, existe um interesse acrescido na utilização de técnicas de aprendizagem por computador (Weiß and Sen, 1996; Sen, 1996), noutras palavras, o sistema é inteligente o suficiente para lhe atribuir conceitos cognitivos para caracterizar, perceber, analisar e prever comportamentos.

Um dos principais problema dos sistemas tradicionais é que nenhum deles é capaz de saber o que fazer se que previamente lhes sejam definidos um conjunto de regras, mas, se uma nova situação aparecer, que não faça parte do conjunto de regras, o sistema não funcionará.

Um agente é um sistema computacional que é colocado num determinado ambiente, e é capaz de agir autonomamente no seu ambiente de modo realizar o seus objectivo (Wooldridge and Jennings (1995)). No que diz respeito à inteligência, este deve ter propriedades reativas, que lhe permitam percecionar mudanças que ocorrem no ambiente e responder atempadamente, deve ser proativo, demonstrando comportamentos orientado para o objetivo e tomando iniciativa com o intuito de satisfazer os seus objetivos, e deve ainda possuir uma capacidade social para que possa interagir com outros agentes.

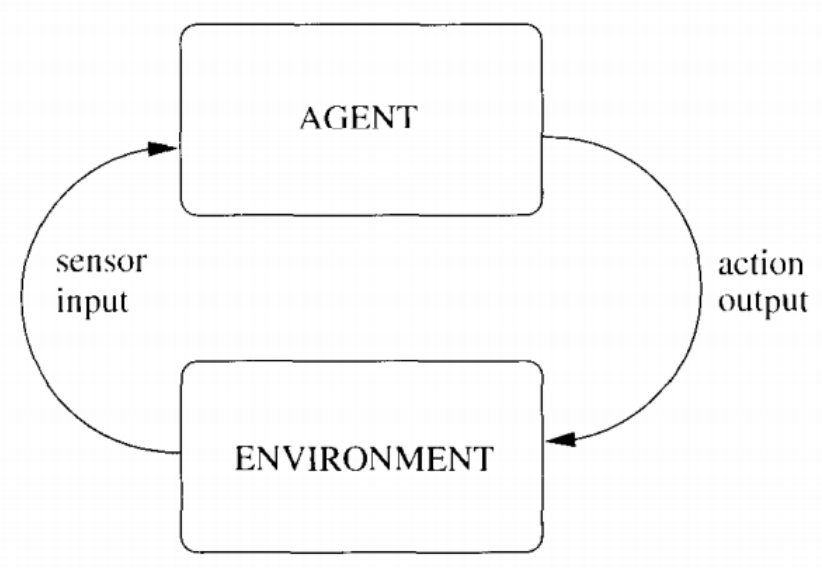


Figura 13: Modelo ilustrativo de um agente inteligente onde recebe informação de sensores e produz ações que afetam o ambiente (Wooldridge 2002)

Porém, a definição de agentes pode dividir-se na sua noção fraca ou forte de acordo com as características do agente (Wooldridge & Jennings, 1995). Um agente designado como fraco um conjunto mínimo de características:

- Autonomia, permite que este opere no seu ambiente sem intervenção de terceiros controlando as ações e estado de conhecimento interno;
- Reatividade, o que possibilita aos agentes percecionrem eventos ocorridos no seu ambiente e responderem de acordo com as mudanças;
- Pro-atividade Tal como na definição de inteligência, os agentes devem revelar iniciativa conduzindo as suas ações à concretização dos seus objetivos;
- Sociabilidade na medida em que deve-se relacionar com outros agentes para a resolução de problemas, competindo ou cooperando.

Por outro lado, complementarmente à definição de agente fraco, um agente forte deve apresentar características cognitivas com a capacidade de desenvolver a sua própria consciência (Wooldridge & Jennings, 1995) (Ferber, 1999), (Nwana, 1996), (Russell & Norvig, 1995).

6.1 ARQUITETURA REATIVA

Determinado tipos de agentes não relacionam eventos passados com a sua decisão atual, levando a um processo de tomada de decisão baseado no presente. A decisão deste tipo de agentes é puramente reativa e em tempo real, geralmente, com um suporte limitado de informação. Como tal, esta

arquitetura não necessita de uma representação simbólica do ambiente nem de forma de raciocínio complexas, apenas tendo em conta dados recebidos dos sensores (Brooks).

6.2 ARQUITETURA DELIBERATIVA

Os agentes mantêm uma representação interna do seu universo de discurso cujo explicito estado mental e conhecimento pode ser modificado por um raciocínio logíco (Newell & Simon, 1976). Requer a capacidade de representar ações e suas consequências derivadas sem que sejam realizadas, ou seja, existe um relação entre a percepção/ação e memória, no sentido que o agente pondera usando casos históricos

6.3 ARQUITETURA HIBRIDA

Procura-se numa arquitetura hibrida colmatar os problemas existentes nas arquiteturas exclusivamente reativas ou deliberativas. Os sistema apenas reativos tem uma noção limitada do seu universo de discurso o que torna difícil implementar comportamentos direcionados ao objetivo do agente, por outro lado, sistema apenas deliberativos realçam mecanismos de raciocínio bastante complexos diminuindo o desempenho em termos reativos. Esta arquitetura procura combinar características das arquiteturas reativas e deliberativas resultando numa arquitetura por camadas. Este tipo de arquitetura encorpora reatividade, deliberação, cooperação e adaptabilidade organizando as camadas de forma hierárquica que interagem entre elas aumentando a robustez e o poder computacional, uma vez que as camadas reativas e deliberativas podem trabalhar paralelamente com intuito de aproveitar, principalmente, a capacidade de resposta e perceção no universo de discurso.

6.4 COORDENAÇÃO

As ações realizadas por agentes influenciam o ambiente, alterando-o de forma dinâmica, o que faz com que este nunca seja estático. Contudo, num ambiente multiagentes é necessário que cada agente percecione a ação realizada por outro agente para que a ação de um determinado agente não interfira ou sobreponha a ação de um outro agente no mesmo ambiente, da mesma equipa. A coordenação pressupões que exista trabalho em conjunto com o intuito de atingir um objetivo comum. “Processo pelo qual um agente raciocina acerca das suas ações locais e das ações previstas dos outros para tentar assegurar que a comunidade atue de modo coerente” (Jennings, 1996). Em suma, a coordenação de agentes preocupa-se em controlar atividades e regular fluxos operacionais num sistema multiagente sendo a comunicação um dos processo principais. A coordenação pode caracterizada de acordo com o contexto do sistema multiagente:

- **Sistemas Cooperativos:** Um sistema cooperativo constitui um conjunto de agentes com o objetivo de realizar atividades de mutuo interesse. Os agentes partilham o seu conhecimento e coordenam as suas atividades em grupo de forma a aumentar a utilidade global do sistema deixando as preferências individuais com prioridade secundária.

- **Sistemas Competitivos:** Ao contrário dos sistema cooperativos, os agentes, num sistema competitivo tentam realizar em primeiro lugar atividades que satisfaçam os seus interesses pessoais, considerando as atividades da comunidade secundárias, podendo não partilhar toda a informação que possuem. Este sistemas são utilizados principalmente em contexto de aquisição de bens ou serviços em que o processo de tomada de decisão dos agentes é individual. A coordenação nestes sistemas passa pela persuasão de outros agente para a realização dos objetivos do agente persuasor.

7

RACIOCÍNIO BASEADO EM CASOS

Durante alguns anos, progressos na área de Inteligência Artificial basearam-se na utilização e desenvolvimento de sistema baseados em regras. Estes sistemas possuem um definição explicita e extensiva do conhecimento do domínio em que atuam. No entanto, o desenvolvimento de tais processos consumem uma enorme quantidade de tempo uma vez que toda a base de conhecimento deve ser transformada em regras descrevendo o conhecimento relativo à forma de fazer em vez da forma como deveria funcionar. Tal implementação revela problema de adaptabilidade tornando este sistemas, muitas das vezes, obsoletos, sendo incapazes de lidar com problemas fora do âmbito definido em forma de regras.

As bases do raciocínio baseado em casos, na área da Inteligência Artificial, tiveram origem em trabalhos realizados por Roger Schank sobre organização de memória dinâmica e com foco na reutilização de situação passadas simulando o raciocínio humano. Schank define o conhecimento sobre o mundo como MOP (memory organization packets), pacotes de memória organizados que representam episódios do nosso quotidiano que sejam suficientemente significantes para serem relembrados. Por outro lado, o primeiro sistema que suporta raciocínio baseado em casos foi desenvolvido por Janet Kolodner na Universidade de Yale integrando a teoria apresentada por Schank, ”memory-based reasoning model and memory based expert systems”.

A solução de problemas utilizando raciocínio baseado em casos em diversas áreas de resolução de problemas, tal como planeamentos, diagnósticos, entre outros e foi com a introdução da problemática e do modelo apresentado por Schank que os sistemas RBC começaram a surgir.

O raciocínio baseado em casos procura resolver problemas adaptando soluções de casos anteriores a problemas idênticos. Esta metodologia não necessita explicitamente de um modelo de domínio, o que torna mais fácil a procura de soluções que é feita pela seleção de casos históricos, e tem a capacidade de aprender através da aquisição de conhecimento através de casos facilitando a manutenção de grandes volumes de dados. Em suma, um RBC procura solucionar novos problemas relembrando situações passadas reutilizando a sua informação e conhecimento que é desenvolvido sob uma teoria de aprendizagem e recordação baseada em experiências antigas de uma forma dinâmica, evoluindo a estrutura da memória. Este processo é similar ao raciocínio humano em certos processos de tomada de decisão.

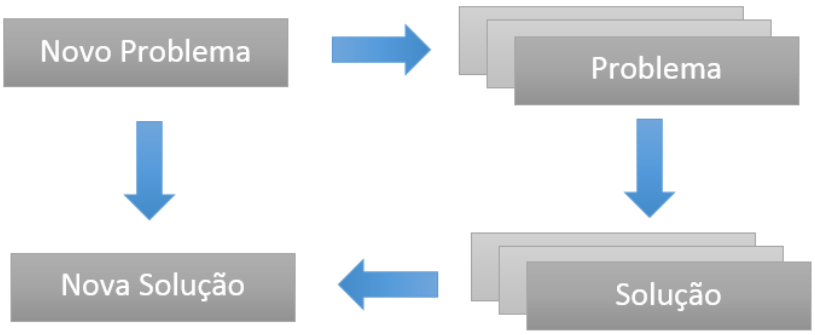


Figura 14: Modelo lógico Raciocínio Baseado em Casos

A aprendizagem é uma das características mais importantes num RBC. Este processo ocorre de um forma natural. Quando um problema é resolvido com sucesso, este é retido sob forma de conhecimento/experiência com o objetivo de vir a ser reutilizado em situações futuras, no entanto, não só os casos com sucesso são memorizados, todos os restantes que resultam em falha são armazenados com a condição de que devem ser identificados quais os problemas resultantes da falha. Contudo, existem várias abordagem que enfatizam, de um certo modo, umas fases do ciclo de um RBC do que outras, por exemplo, uns preferem armazenar uma maior quantidade de casos na memória de casos enquanto outros preferem realçar o processamento aperfeiçoando os algoritmos de seleção e reutilização. Além disso, alguns métodos podem ser totalmente autónomos enquanto outros podem depender, em grande parte, da interação com utilizadores.

- **Raciocínio Baseado em Exemplos:** Esta abordagem procura classificar corretamente um determinado exemplo não classificado, utilizando a classe com mais similaridade como solução do problema. O processo de classificação utiliza casos passados em vez de regras pré-definidas.

- **Raciocínio Baseado em Memórias:** Uma metodologia que utiliza uma grande quantidade de casos para o conhecimento e utilizado o raciocínio como forma de pesquisa em memória, dando maior importância à organização e acesso, em paralelo (fator de diferenciação dos outros métodos) da memória.

- **Raciocínio Baseado em Casos:** Nesta abordagem é considerado que o caso contem um certo grau de informação e uma certa complexidade relativamente à sua organização interna. Este método pressupões uma capacidade de adaptação e modificação de soluções encontradas em casos anteriores.

- **Raciocínio Baseado em Analogias** Ao contrário do raciocínio baseado em casos, esta abordagem utiliza casos de diferentes domínios utilizando-os de forma idêntica ao raciocínio baseado em casos. Tem como principal objetivo a transferência de soluções de domínios diferentes para solucionar o problema atual.

7.1 TIPO DE CONHECIMENTO

Richter distingue quatro fatores que representam tipo de conhecimento sobre o domínio do problema: o vocabulário, medida de similaridade, adaptação do conhecimento e os casos.

- O vocabulário inclui o conhecimento necessário para a escolha das características utilizadas para selecionar um caso. Tais características devem ser uteis para a seleção do caso e serem o mais seletivas possível para evitar a que dois ou mais casos sejam retornados. Este tipo de conhecimento deve ser bem estruturado para que o processo de aprendizagem seja eficaz e fácil para representar novos casos.
- As medidas de similaridade incluem um representação quantitativa do conhecimento usado para escolher de modo mais eficiente um caso com a maior nível de características semelhantes entre o caso selecionado e o novo caso. Um conhecimento bastante útil quando utilizado em processos de classificação em que a estrutura do caso é bastante complexa.
- A adaptação do caso selecionado ao novo caso é dado pelo conhecimento das regras do domínio organizacional tendo este a capacidade de avaliar o caso escolhido de acordo com as necessidade do novo caso. Normalmente este processo é realizado por intervenção de terceiros que possuem mais conhecimento organizacional e das regras utilizadas.
- Os casos são acontecimento que ocorreram no passado e representa o conhecimento, adquirido ao longo do tempo devido à experiência do sistema, determinado pelo vocabulário escolhido. Uma das preocupações na aquisição de casos passa pela filtragem na aprendizagem de modo a evitar o sobrecarregamento da memória de casos o que dificulta a recuperação de soluções para novos casos.

7.2 MODELO

Os sistemas suportados por RBC podem ser representados como processos cíclicos constituído, principalmente, por quatro ações principais:

- Selecionar e Recuperar casos previamente vivenciados cuja similaridade seja a mais próxima possível.
- Reutilizar os casos copiando-os e integrando-os numa nova solução.
- Rever ou Adaptar a solução com o objetivo de solucionar o novo problema.
- Reter a nova solução depois de avaliada e confirmada, constitui o processo de aprendizagem.

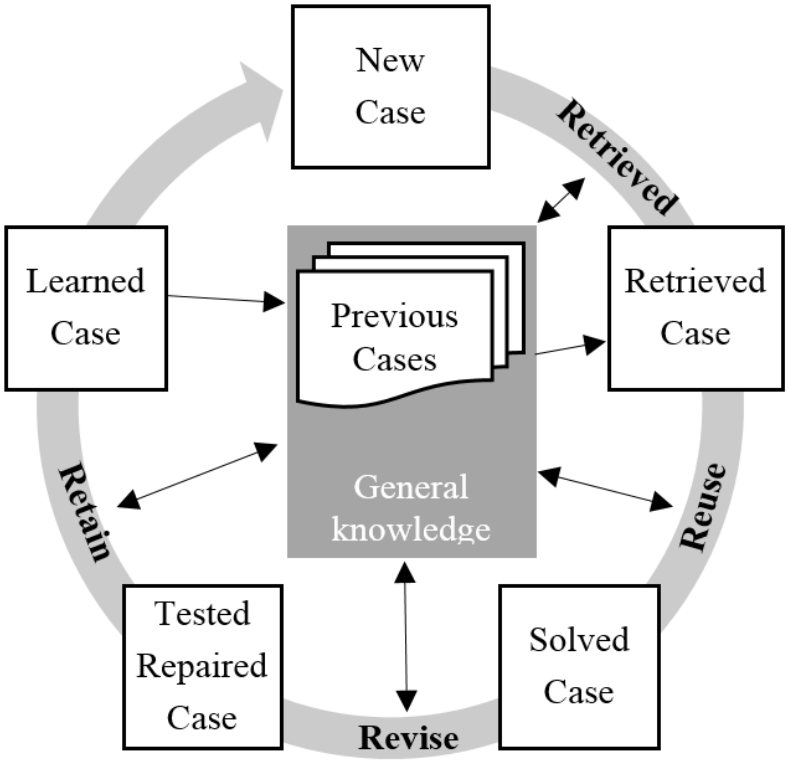


Figura 15: Modelo Raciocínio Baseado em Casos

7.2.1 **Selecionar e Recuperar**

A ocorrência de um novo caso leva à aplicação de algoritmos de recuperação de casos que dependem de índices e da organização da memória de casos utilizando critérios que determinam o método de procura e os casos potencialmente uteis para resolver o novo problema encontrado. Este processo pode ser divido em duas fases:

- **Recuperação de casos:** Tem por objetivo recuperar os melhores casos que possam suportar os processos seguintes e que contêm o potencial de melhor se ajustar aos novos casos. Esta recuperação é feita através da utilização de características do novo caso como indexes para a base de casos e depende bastante do modelo de memória e dos procedimentos utilizados. Para realizar este processo existem vários algoritmos que podem ser utilizados.

- **Selecionar o melhor subconjunto:** Devido à possibilidade de vários casos serem recuperados da base de casos, esta fase consiste na seleção dos melhores casos comparando os casos determinando o grau de similaridade entre eles.

Este processo é considerado o mais complexo de todo o ciclo, uma vez que, mal implementado pode prejudicar as restantes etapas. As técnicas para seleção e recuperação de casos mais investigadas passam por *k-nearest neighbors (k-NN)*, árvores de decisão e abordagens indutivas. Independentemente

da técnica usada, todos os algoritmos necessitam de considerar a similaridade como fator de decisão na seleção do caso mais apropriado. A eficácia de uma medida de similaridade é determinada pela utilidade de um caso recuperado na resolução de um novo problema (Sankar et. al.)

- **Distância Euclidiana Ponderada** Um dos métodos de distância mais comuns e utilizados é baseado na localização de objetos num espaço Euclidiano.

- ***k-nearest neighbors (k-NN)*:** Este método permite a recuperação dos casos cuja soma do peso das características do caso recuperado é superior aos restantes na mesma base de casos. As características podem ainda ser consideradas mais importantes com a aplicação de pesos extra no processo de comparação.

7.2.2 Adaptação do Caso

Uma vez que, raramente, as soluções anteriores não correspondem ou se adquam na totalidade às novas situação, é necessário a adaptação de um caso passando por processos de transformação da solução do caso recuperado numa solução apropriada ao problema atual. Estas adaptações procuram diferenças significativas entre o casos selecionado e o caso atual aplicando um conjunto de formulas sugerindo uma nova solução. O modelos de adaptação podem ser classificados em diferentes categorias (Cunningham et al., 1994; Watson and Marir, 1994; Kolodner and Leake, 1996; Wilke and Bergmann, 1998)::

- **Nula:** O tipo de adaptação mais simples que consiste na copia integral da solução recuperada para a resolução do novo caso sem qualquer tipo de modificação. Uteis para tarefas de classificação ou raciocínios complexos com soluções simples. A maior parte dos sistemas de RBC funcionam com base nesta adaptação.

- **Substituição:** Consiste na substituição de atributos que são inválidos ou que se contradizem porque entram em conflito com os novos requisitos do problema atual. Esta método não altera a estrutura da solução mas apenas o valor dos seus atributos.

- **Transformação:** Ao contrário do método de substituição, a transformação envolve mudanças na estrutura da solução, alterando elementos, adicionando ou removendo caraterísticas com base em certas condições. Estas alterações são fundamentadas com base em regras estabelecidas com base no conhecimento do domínio (Kolodner, 1993).

- **Generativa:** A técnica mais complexa que tem por objetivo reajustar toda o raciocínio que levou à solução selecionada, muita das vezes conhecida como analogia derivacional (Veloso and Carbonell, 1994). São muitas das vezes usados, não para resolver o problema desde o inicio para para gerar soluções para as partes do problema que são incompatíveis com a versão selecionada.

7.2.3 **Aprendizagem e Manutenção de Casos**

Neste modelo apresentar, o processo de aprendizagem constitui o ultimo passo do ciclo após a produção e aceitação da nova solução para o novo caso. Num modo geral, o caso é armazenado apropriadamente na memória de casos para futura utilização possibilitando ao sistema uma capacidade de aprendizagem baseada na experiência da resolução de problemas.

Parte II

TRABALHO DESENVOLVIDO

8

REPRESENTAÇÃO DO CONHECIMENTO E RACIOCÍNIO

Existem diversas abordagens propostas para representação de conhecimento e raciocínio que usam o paradigma da Programação Lógica, principalmente na área da teoria dos modelos (Kakas et al, 1998; Gelfond and Lifschitz, 1988; Pereira and Anh, 2009), teoria da prova (Neves, 1984; Neves et al, 2007). Ao longo deste trabalho é utilizado a abordagem demonstrada na teoria da prova como forma de extensão à linguagem de programação lógica para representação de conhecimento e raciocínio. Esta abordagem, programação lógica estendida, é um conjunto de clausulas finitas representadas na forma:

$$p \leftarrow p_1, ..., p_n, not\ q_1, ..., not\ q_m \tag{1}$$

$$?(p_1, ..., p_n, not\ q_1, ..., not\ q_m)\ (n, m \geq 0) \tag{2}$$

em que ? é um átomo de domínio que denota falsidade.p_i, q_j e p são literais clássicos, podendo ser divididos em átomos positivos ou átomos antecedidos pelo sinal de negação $\neg$ (Neves, 1984). Cada programa está associado a um conjunto de abdutivos (Kakas et al. 1998; Pereira and Anh, 2009) representados sob forma de exceções para as extensões dos predicados que constituem o programa. Esta metodologia introduz um mecanismo de procura eficaz, formalizando a taregas de representação do conhecimento e raciocínio para a resolução de problemas.

Os modelos e raciocínios qualitativos têm vindo a ser analisados em pesquisas relacionadas com a Inteligência Artificial e teóricas de base de dados devido ao facto de, cada vez mais, ser necessário oferecer suporte no processo de tomada de decisão (Halpern, 2005; Kovalerchuck and Resconi, 2010). Devido a esta necessidade, surge uma medida capaz de representar a qualidade da informação de um sistema relativamente à imagem que este tem da representação do conhecimento e raciocínio. A medida *QoI*, *Quality-of-Information*, relativamente à extensão do predicado i, será representada pelo

valor de verdade no intervalo $[0, 1]$, ou seja, se a informação é conhecida (positiva) ou falsa (negativa) o QoI da extensão do predicado i é 1. Para situações em que a informação é desconhecida, a medida de QoI é dada pela expressão:

$$QoI_i = \lim_{N \to \infty} \frac{1}{N} = 0 \ \ (N > 0) \tag{3}$$

onde N representa a cardinalidade do conjunto de clausulas ou termos do predicado i que representa a incompletude sob consideração. Para situações em que a extensão do predicado i é desconhecido mas pode ser retirado a partir de um conjunto de dados, o valor é dado por:

$$QoI_i = \frac{1}{Card} \tag{4}$$

em que $Card$ denota a cardinalidade do conjunto de adutíveis para i se disjunto, caso contrário, o QoI é dado por:

$$QoI_i = \frac{1}{C_1^{Card} + ... + C_{Card}^{Card}} \tag{5}$$

onde C_{Card}^{Card} é um subconjunto de $Card$ elementos escolhido de conjunto de $Card$ elementos. Por outro lado, com o objetivo de representar a importância de um atributo k na extensão do predicado i é considerado o elemento w_i^k normalizado pelo fator:

$$\sum_{1 \leq k \leq n} w_i^k = 1, \ \forall_i \tag{6}$$

sendo $\forall$ o quantificador universal. Com isto, é possível definir uma função de pontuação para o predicado i, $V_i(x)$ de modo a que para um valor $x = (x_1, ..., x_n)$, com base nos atributos do predicado i, se possa ter:

$$V_i(x) = \sum_{1 \leq k \leq n} w_i^k * \frac{QoI_i(x)}{n} \tag{7}$$

permitindo que se possa definir:

$$predicate_i(x_1, ..., x_n) :: V_i(x) \tag{8}$$

que representa a qualidade do predicado i relativamente a todos os predicados que constituem o programa. A função para avaliar o programa lógico através da formula:

$$LP_{ScoringFunction} = \sum_{i=1}^{n} V_i(x) * p_i \tag{9}$$

onde p_i significa a relevância do predicado i em relação aos restantes predicados. Todos os pesos das extensões dos predicados devem ser normalizados pela expressão:

$$\sum_{i=1}^{n} p_i = 1, \forall_i \tag{10}$$

onde $\forall$ representa o quantificador universal. Para estabelecer o universo de discurso, de acordo com a informação dada nos programas lógicos que reforça a informação sobre o problema em consideração:

$$predicate_i - \bigcup_{1 \leq j \leq m} clause(x_1, ..., x_n) :: QoI_i :: DoC_i \tag{11}$$

sendo $\bigcup$ e m, respetivamente, união e cardinalidade da extensão do predicado i. Complementarmente, o DoC_i representa a confiança depositada no valor do atributo de uma termo em particular da extensão dos predicados i. Supondo que o universo de discuso é dado pela extensão dos predicados:

$$f_1(...), f_2(...), ..., f_n(...) \; where \; (n \geq 0) \tag{12}$$

Os argumentos que forma uma clausula são todos os atributos de um determinado evento, assumindo que uma clausulo representa um acontecimento. Os valores dos argumentos podem ser desconhecidos, membros de um conjunto, pertencentes a um determinado intervalo, ou qualificativos de uma determinada observação. Considerando a primeira clausula poderá conter valores no intervalo $[20, 30]$ num domínio entre 0 e 50, a segunda clausula é representada por um valor preciso com um domínio compreendido entre 0 e 10 e um valor desconhecido definido como terceira clausula representado por $\perp$ em que o intervalo do seu domínio é $[0, 100]$. O extensão do predicado f_1 é dado por:

$$f_1 : x_1, x_2, x_3 \rightarrow \{\ 0, 1\ \} \tag{13}$$

em que ″ { ″ e ″ } ″ representam um conjunto, onde ″0″ e ″1″ denotam, respetivamente, valores de verdade $true$ e $false$. Portanto:

$$\begin{aligned}
&\{\\
&\quad \neg f_1(x_1, x_2, x_3) \leftarrow not\ f_1(x_1, x_2, x_3)\\
&\quad f_1(\ \underbrace{[20, 30],\quad 5,\quad \bot}_{valor\ dos\ atributos\ para\ x_1,x_2,x_3}\ ) :: 1 :: DoC\\
&\quad \underbrace{[0, 50]\quad [0, 10]\quad [0, 100]}_{dominio\ dos\ atributos\ para\ x_1,x_2,x_3}\\
&\quad ...\\
&\}
\end{aligned}$$

Após estabelecidos os termos da extensão do predicado, é necessário definir todos os argumentos num intervalo continuo. Nesta fase o domínio do argumentos é considerado, no entanto, uma vez que o terceiro argumento é desconhecido, este irá ser representado na sua amplitude total, cobrindo todas as possibilidades do domínio.

$$\begin{aligned}
&\{\\
&\quad \neg f_1(x_1, x_2, x_3) \leftarrow not\ f_1(x_1, x_2, x_3)\\
&\quad f_1(\ \underbrace{[20, 30],\quad [5, 5],\quad [0, 100]}_{valor\ dos\ intervalos\ dos\ atributos\ x_1,x_2,x_3}\ ) :: 1 :: DoC\\
&\quad \underbrace{[0, 50]\quad [0, 10]\quad [0, 100]}_{dominio\ dos\ atributos\ para\ x_1,x_2,x_3}\\
&\quad ...\\
&\}
\end{aligned}$$

Concluído o segundo passo, é possível calcular o *Degree of Confidence*, DoC, para cada atributo que constituem os argumentos do termo. Este calculo é normalizado num intervalo $[0,1]$ de acordo com o procedimento de normalização dado pela expressão:

$$\frac{(Y - Ymin)}{(Y_{max} - Ymin)} \tag{14}$$

$$\{$$
$$\neg f_1(x_1, x_2, x_3) \leftarrow not\ f_1(x_1, x_2, x_3)$$
$$x_1 = \left[\tfrac{20-0}{50-0}, \tfrac{30-0}{50-0}\right]\ \ x_2 = \left[\tfrac{5-0}{5-0}, \tfrac{10-0}{10-0}\right]\ \ x_2 = \left[\tfrac{0-0}{100-0}, \tfrac{100-0}{100-0}\right]$$
$$f_1(\underbrace{[0.4, 0.6], \quad [0.5, 0.5], \quad [0, 1]}_{\substack{valor\ dos\ intervalos\ normalizados\ dos \\ atributos\ para\ x_1,x_2,x_3}}) :: 1 :: DoC$$
$$\underbrace{[0,1] \quad [0,1] \quad [0,1]}_{\substack{dominio\ normalizado\ dos \\ atributos\ para\ x_1,x_2,x_3}}$$
$$...$$
$$\}$$

O *Degree of Confidence* é obtido pela equação $DoC = \sqrt{1 - \Delta l^2}$ onde Δl^2 representa o comprimento do intervalo dos argumentos normalizados.

$$\{$$
$$\neg f_1(x_1, x_2, x_3) \leftarrow not\ f_1(x_1, x_2, x_3)$$
$$f_1(\ \underbrace{0.98, \quad 1 \quad , \quad 0}_{\substack{valor\ de\ confianca\ dos \\ atributos\ para\ x_1,x_2,x_3}}\ ) :: 1 :: DoC$$
$$\underbrace{0.4, 0.6] \quad [0.5, 0.5] \quad [0, 1]}_{\substack{valor\ dos\ intervalos\ normalizados\ dos \\ atributos\ para\ x_1,x_2,x_3}}$$
$$\underbrace{[0,1] \quad [0,1] \quad [0,1]}_{\substack{dominio\ normalizado\ dos \\ atributos\ para\ x_1,x_2,x_3}}$$
$$...$$
$$\}$$

Por fim, dado o exemplo, temos que o *DoC* para a extensão do predicado f_1, $f_1(0.98, 1, 9)$ é calculado pela soma de todos os *DoC* de cada termo da extensão do predicado dividido pelo total de termos de f_1, $(0.98 + 1 + 0)/3$, assumindo que todos os atributos tem o mesmo peso.

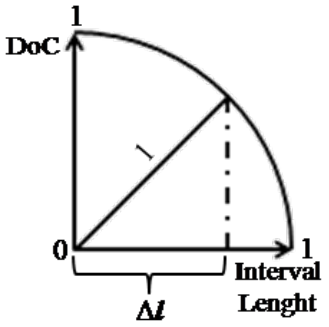


Figura 16: Avaliação do Grau de Confiança

9

CASO DE ESTUDO

Com o objetivo de exemplificar a abordagem apresentada para resolução de problemas, foi utilizado um modelo relacional de base de dados, uma vez que este providencia a ferramenta que corresponde às expetativas do sistema (Liu and Sun, 2007) e considerado como o gênese da abordagem de programação lógica para representação de conhecimento e raciocínio (Neves, 1984).

Como caso de estudo, considera-se o cenário onde a base de dados relacional é apresentada sob forma de extensões dos relacionamentos, que representam situações on alguém tem de gerir informação sobre predisposição e deteção de risco para acidentes vasculares cerebrais.O sintomas agregam um conjunto de atributos que são caracterizados pela sua ocorrência no caso. Nesta situação, em que se procura simular acontecimento do quotidiano o mais próximo possível, os atributos podem ser opcionais ou definidos com um certo grau de incerteza, portanto considera-se que hajam alguns dados incompletos. Por exemplo, na entidade *Stroke Predisposition*, o atributo *Systolic Blood Pressure* no primeiro caso é desconhecido enquanto que o atributo *Risk Factors* varia no intervalo $[1, 2]$.

Os valores apresentados no atributo *Lifestyle Habits* e *Risk Factors* da entidade *Stroke Predisposition* resultam do somatório das entidade correspondentes, variando entre $[0, 6]$ e $[0, 4]$, respetivamente.

$$stroke : Age, Gen_{der}, P_{revious}S_{troke}E_{pisodes}, B_{lood}S_{ystolic}P_{ressure}, Chol_{esterolLDL},$$
$$Chol_{esterolhHDL}, Trigly_{cerides}, L_{ifestyle}H_{abits}.R_{isk}F_{actors} \rightarrow \{\ 0, 1\ \}$$

onde 0 (zero) e 1 (um) representam, respetivamente, os valores de verdade *false* e *true*. Podemos definir a extensão do predicado *stroke* como:

$$
\begin{aligned}
&\{\\
&\quad \neg stroke(Age, Gen, PSE, BSP, Chol_{LDL}, Chol_{HDL}, Trigly, LH, RF)\\
&\quad \leftarrow not\ stroke(Age, Gen, PSE, BSP, Chol_{LDL}, Chol_{HDL}, Trigly, LH, RF)\\
&\\
&\quad stroke(\underbrace{69, \quad 0, \quad 1, \quad \bot, \quad 131, \quad 49, \quad 200, \quad 4, \quad \{\,1,2\,\}}_{valor\ dos\ atributos}) :: 1 :: DoC\\
&\\
&\quad \underbrace{[22,95][0,1][0,1][70,200][50,250][20,90][90,600][0,6][0,4]}_{dominio\ dos\ atributos}\\
&\quad ...\\
&\}
\end{aligned}
$$

Frequentemente os dados providenciados ou obtidos são incompletos, atributos com valores errados, o simples fato de não serem relevantes ou necessários para análise, atributos com baixa influencia preditiva e também a existência de inconsistência. Os dados são designados como dados não processados com um nível de qualidade que pode ser melhorado através do uso de determinadas técnicas. Comparado com o processo de *data mining* o sistema incorpora um processo de normalização, transformando todos os valores dos argumentos em intervalos contínuos e também o DoC de todas as clausulas da extensão do predicado *stroke*, alteração linear dos dados de origem num determinado intervalo. A normalização é particularmente útil para algoritmos de classificação que envolvem redes neuronais artificiais e o seu principal objetivo é mapear os dados de acordo com uma determinada escala.

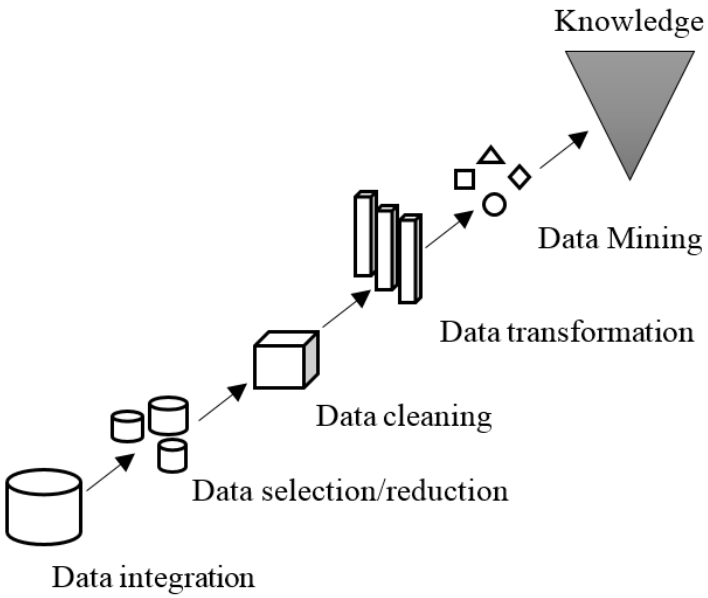


Figura 17: Modelo concetual do processo de *data mining*

{

$\neg stroke(Age, Gen, PSE, BSP, Chol_{LDL}, Chol_{HDL}, Trigly, LH, RF)$
$\leftarrow not\ stroke(Age, Gen, PSE, BSP, Chol_{LDL}, Chol_{HDL}, Trigly, LH, RF)$

$stroke(\underbrace{1, \quad 1, \quad 1, \quad 0, \quad 1, \quad 1, \quad 1, \quad 1, \quad 0.968}_{valor\ de\ confiancado satributos}) :: 1 :: DoC$

$\underbrace{[0.64, 0.64][0, 0][1, 1][0, 1][0.4, 0.4][0.41, 0.41][0.22, 0.22][0.67, 0.67][0.25, 0.5]}_{valor\ dos\ atributos\ normalizados}$

$\underbrace{[0, 1][0, 1][0, 1][0, 1][0, 1][0, 1][0, 1][0, 1][0, 1]}_{dominio\ dos\ atributos\ normalizados}$

...

}

Tabela 1: Entidade com informação do paciente

#	Age	Gender	Previous Stroke Episode	Systolic Blood Pressure	CholesterolLDL	CholesterolHDL	Triglycerides
1	69	F	1	$\perp$	111	49	200
...	...	...	...	...	...	...	...
n	32	M	0	120	$\perp$	$\perp$	$\perp$

Tabela 2: Entidade com valores de predisposição de AVC

#	Age	Gender	Previous Stroke Episode	Systolic Blood Pressure	CholesterolLDL	CholesterolHDL	Triglycerides	Lifestyle Habits	Risk Factors
1	69	F	1	$\perp$	111	49	200	4	$[1,2]$
...	...	...	...	...	...	...	...	...	...
n	32	M	0	120	$\perp$	$\perp$	$\perp$	6	$[0,1]$

#	No Smoking	Exercise	Breakfast	Vegetables/Fruit	Low Salt	Low Sugar
1	1	0	1	1	0	1
...	...	...	...	...	...	...
n	1	1	1	1	1	1

Tabela 3: Entidade de Hábitos de Vida

#	Diabetes	Obesity	Hypertension	Long-term Medicaments
1	1	0	0	$\perp$
...	...	...	...	...
n	0	0	0	$\perp$

Tabela 4: Entidade de Fatores de Risco

O esquema em forma de tabelas acima apresentado representa uma extensão ao modelo relacional da base de dados. No atributo *Previous Stroke Episode*, na entidade que contem a informação do paciente, 0(zero) e 1(um) denotam, respetivamente, *não ocorrência* e *ocorrência*. Nas entidades que cotem dados sobre os hábitos de vida e fatores de risco 0(zero) e 1(um) denotam, respetivamente, *sim* e *não*. A coluna *Gender* representa, em todas entidades, respetivamente, *feminino* e *masculino*.

Vários estudos mostram de as Redes Neuronais Artificiais podem ser usadas com sucesso para modelar dados e identificar relacionamentos complexos entre valores de entrada e saída (Caldeira et al, 2011; Vicente et al, 2012; Salvador et al, 2013). Como exemplo, é utilizado o ultimo caso

apresentado nas entidades das tabelas acima apresentadas, onde se pode ter uma situação em que é necessária uma avaliação da predisposição para acidente vascular cerebral.

$$
\begin{aligned}
&\{\\
&\quad \neg stroke(Age, Gen, PSE, BSP, Chol_{LDL}, Chol_{HDL}, Trigly, LH, RF)\\
&\quad \leftarrow not\ stroke(Age, Gen, PSE, BSP, Chol_{LDL}, Chol_{HDL}, Trigly, LH, RF)\\
&\\
&\quad stroke(\underbrace{32, \quad 1, \quad 0, \quad 120, \quad \bot, \quad \bot, \quad \bot, \quad 6, \quad [0,1]}_{valor\ dos\ atributos}) :: 1 :: DoC\\
&\\
&\quad \underbrace{[22,95][0,1][0,1][70,200][50,250][20,90][90,600][0,6][0,4]}_{dominio\ dos\ atributos}\\
&\quad ...\\
&\}
\end{aligned}
$$

De acordo com o formalismo apresenta acima, quando a transição para intervalos contínuos estiver completa, é possível ter os argumentos do predicado *stroke* normalizados no intervalo $[0, 1]$ para que seja possível obter o *DoC* de cada clausula ou termo.

$$
\begin{aligned}
&\{\\
&\quad \neg stroke(Age, Gen, PSE, BSP, Chol_{LDL}, Chol_{HDL}, Trigly, LH, RF)\\
&\quad \leftarrow not\ stroke(Age, Gen, PSE, BSP, Chol_{LDL}, Chol_{HDL}, Trigly, LH, RF)\\
&\\
&\quad stroke(\underbrace{1, \quad 1, \quad 1, \quad 1, \quad 0, \quad 0, \quad 0, \quad 1, \quad 0.97}_{valor\ de\ confiancado satributos}) :: 1 :: DoC\\
&\quad \underbrace{[0.14,0.14][1,1][0,0][0.38,0.38][0,1][0,1][0,1][1,1][0,0.25]}_{valor\ dos\ atributos\ normalizados}\\
&\\
&\quad \underbrace{[0,1][0,1][0,1][0,1][0,1][0,1][0,1][0,1][0,1]}_{dominio\ dos\ atributos\ normalizados}\\
&\quad ...\\
&\}
\end{aligned}
$$

Os cálculos, de uma forma normalizada, para os valores de entrada para a RNA são apresentados no capítulo 3. O resultado traduz a predisposição para um acidente vascular cerebral e a confiança que o agente em tal acontecimento.

Este trabalho incorpora registos de 250 pacientes (duzentos e cinquenta termos ou clausulas que constituem a extensão de predicados *stroke*), com uma média de idade de 57 anos, compreendidos entre o 31 e 87 anos. A distribuição de géneros é de 47% e 53% para pacientes femininos e masculinos, respetivamente.

De forma a garantir a significância estatística dos resultados alcançados, 20 execuções foram aplicadas a todos os testes. Em cada simulação, os dados disponíveis foram divididos aleatoriamente em duas partições exclusivas, o conjunto de dados relativos a treino (67%) e os restantes dados para simulação e testes (33%). O algoritmo de retro-propagação foi usado no processo de aprendizagem da RNA. Na saída na camada de pré-processamento foi utilizada a função *identity* e *sigmoid* nas restantes camadas.

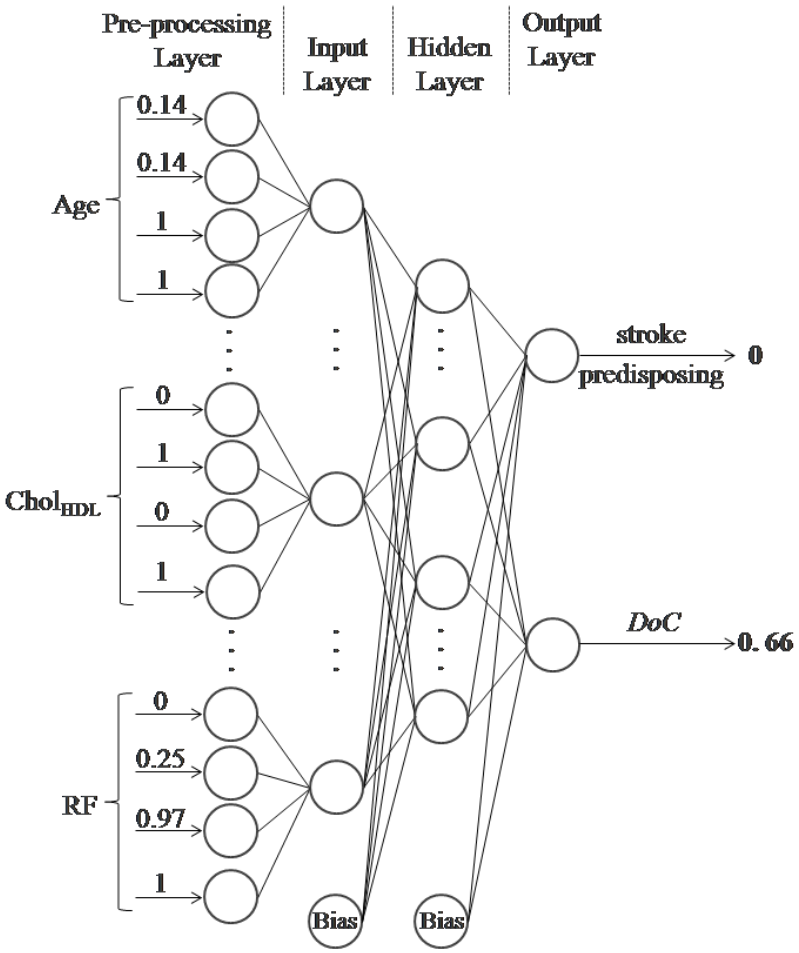


Figura 18: Topologia da Rede Neuronal Artificial

10

ARQUITETURA

A interoperabilidade pode ser realizada através de várias metodologias que permitem a comunicação entre todos os diferentes cenários e ambientes. Tal como analisado anteriormente, um SMA compreende um grupo de entidade que cooperam com o objectivo de resolver problemas que estão para além das capacidades humanas. Essas entidade, agente inteligentes, atuam num determinado universo de discurso, produzindo um determinado efeito sendo caracterizadas pela sua flexibilidade de desenvolver uma capacidade autónoma sobre o ambiente. Os agentes, neste sistema multiagente, reagem segundo informação providencia por outras entidades. No entanto, de acordo com as arquiteturas mencionados no capítulo 2, e com toda a metodologia utilizada para a seleção da informação, os agentes agem de acordo com regras e representações simbólicas explicitas do seu ambiente, cujo estado de conhecimento pode variar através do raciocínio logico-matemático descrito nos capítulos 3 e 4, incorporando neste sistema multiagente capacidades de uma arquitetura deliberativa.

Porque, a representação ideal, para este sistema, passa pela combinação de características reativas e deliberativas, os agentes foram desenvolvidos para responder o mais rápido possível a novos casos mas também para processar a informação de uma forma deliberativa, concedendo-lhes as capacidades de raciocínio para o processo de tomada de decisão.

Os agentes foram divididos em camadas lógicas de acordo com as suas características, funcionalidades e processamento de tarefas resultando num total de quatro agentes completamente diferentes. Os agentes podem ser mais reativos do que deliberativos ou vice-versa.

10.1 ESTRUTURA GLOBAL

A estrutura deste sistema é disposta em camadas distintas em que a comunicação estabelecida, com uma ontologia específica, é implementada via canais TCP/IP usando *sockets*. Para assegurar o fluxo de dados, foi implementa uma mensagem de reconhecimento para cada ação envolvendo dois ou mais agentes diferentes. Este tipo de mensagem ajuda o controlo do fluxo de dados e controlo a participação do agente no ciclo de vida do sistema.

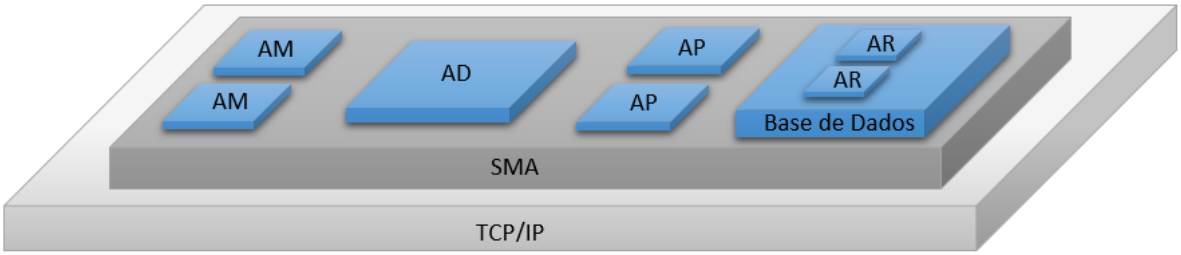


Figura 19: Esquema global da arquitetura

10.2 AGENTE DE MONITORIZAÇÃO

Este agente tem um relacionamento condição/ação com o sistema, construindo uma estrutura de sintomas dinâmica obtidos pelo utilizador.

Após a definição da estrutura de sintomas, este envia todo o seu conhecimento para o agente de decisão ficando em modo de espera, aguardando uma solução ao problema. Um agente com este tipo de comportamento simula a recolha de dados de um sistema tradicional com capacidade comunicativa entre o sistema e o utilizador. O agente de monitorização tem apenas características reativas uma vez que depende sempre de respostas e alterações do ambiente em que se enquadra.

Algorithm 1 Comportamento do agente ao detetar novos casos

```
1:  procedure LISTEN
2:      case ← user
3:      if case is null then return false
4:      else
5:          goto CollectSymptoms(case)
```

Algorithm 2 Recolha e envio de sintomas para construir um novo caso

```
1:  procedure COLLECTSYMPTOMS(case)
2:      symptoms ← empty
3:      while no more symptom do
4:          symptom ← user symptom
5:          if symptom is null then return false
6:          else
7:              symptoms ← symptom
8:      goto SendMessage(symptoms, case)
9:      ack ← waitACK
10:     if ack is null then return false
11:     else
12:         waitResponse
13:         goto SendSolution(solution, case)
```

10.3 AGENTE DE DECISÃO

O Agente de decisão é responsável pelo processo de tomada de decisão. A sua principal funcionalidade passa por avaliar todos os agentes de processamento, numa determinada resolução de problemas, e decidir qual o resultado/solução que deverá ser aceite e enviada para o agente de monitorização. Além disso, este agente é responsável pela gestão do fluxo de mensagens no sistema.

A decisão da solução ideal avalia independentemente os resultados mas também considera o fator de confiança que o agente decisor tem por cada um dos agentes de processamento envolventes no processo.

Algorithm 3 Comportamento de um agente de decisão ao receber um novo caso

```
1:  procedure LISTEN
2:      agents ← enrolled agents
3:      case ← case from monitor
4:      if case is null then return false
5:      else
6:          for each agent do
7:              goto SendCase(case, agent(i))
8:              ack ← waitACK
9:              if ack is null then return false
10:             else
11:                 solutions ← Solution(case, agent)
12:         goto EvalSolutions(solutions, case)
```

Algorithm 4 Avaliação e envio de soluções

```
1:  procedure EVALUATE(solutions)
2:      agents ← enrolled agents
3:      case ← case from monitor
4:      if case is null then return false
5:      else
6:          for each solution do
7:              goto Eval(solution(i))
8:              if solution(i) is best solution then return solution(i)
9:              else
10:                 continue
```

10.4 AGENTE DE PROCESSAMENTO

Todo o processo deliberativo e raciocínio é implementado no agente de processamento, partilhando informação com o agente de recursos para retornar todas a informação pertinente para a resolução do problema com o objetivo de encontrar a melhor solução ao problema. Este agente integra um

conjunto de processos de classificação e segmentação como forma de raciocínio lógico. O agente de processamento desempenha um papel importante no sistema uma vez que retorna soluções face ao problema apresentado. Contudo, diferentes agentes podem retornar diferentes soluções ou até mesmo soluções falsas.

Algorithm 5 Sequência de métodos utilizados no agente de processamento

```
1:  procedure SOLUTION(case)
2:      cases ← cases from database
3:      segments ← Segmentation(cases)
4:      caseSegment ← Classify(case, segments)
5:      for each casesOnSegment do
6:          solution ← Similarity(case, casesOnSegment)
7:          if solution(i) is best solution then return solution(i)
8:          else
9:              continue
10:     return solution
```

10.5 AGENTE DE RECURSOS

Um agente simples, cuja principal funcionalidade é apoiar o agente de processamento, otimizando a procura de casos na memória, que neste caso, é representada numa base de dados. Contudo, este agente dispões de um processo de aprendizagem comandado pelo agente de processamento para aquisição de novos casos.

Algorithm 6 Envio dos casos armazenados

```
1:  procedure LISTEN
2:      cases ← empty
3:      for each case in memory do
4:          cases ← case(i)
5:      return cases
```

10.6 F.I.P.A

A FIPA é uma organização internacional que se dedica a promover a indústria dos agentes inteligentes, através do desenvolvimento de especificações abertas que suportam a interoperabilidade entre agentes e a computação baseada em agentes.

As especificações FIPA representam um conjunto de normas que promovem a interoperabilidade entre agentes heterogêneos e os serviços que eles podem representar e podem ser divididas em diferentes categorias: agente de comunicação, agente de transporte de mensagem, agente de gestão, arquitetura e

aplicações. Destas categorias, a comunicação entre agentes é a categoria central no coração do modelo de sistema multiagente FIPA.

- As especificações do agente de transporte de mensagem gerem a representação e o transporte das mensagens através dos diferentes protocolos de rede, tais como wireline e wireless;
- As especificações do agente de comunicação possuem as linguagens de comunicações possíveis, tipo de mensagens, protocolos de interação, representação do conhecimento e a teoria dos atos comunicativos.
- As especificações do agente de gestão focam-se no controlo e gestão dos agentes presentes na plataforma e fora dela;
- As especificações da arquitetura abstrata lida com as entidades abstratas necessárias para construir os serviços do agente e o seu ambiente;
- As especificações aplicacionais são exemplos de áreas onde os agentes podem ser implantados. Representam a ontologia e o serviço que descreve as especificações para um determinado domínio.

O seguinte diagrama representa, de uma forma geral, o modo como os agentes interagem uns com os outros, trocando mensagens sobre o protocolo definido. Note-se que em cada mensagem troca entre dois agentes existe sempre uma mensagem de reconhecimento que informa o agente que iniciou a comunicação da sua receção.

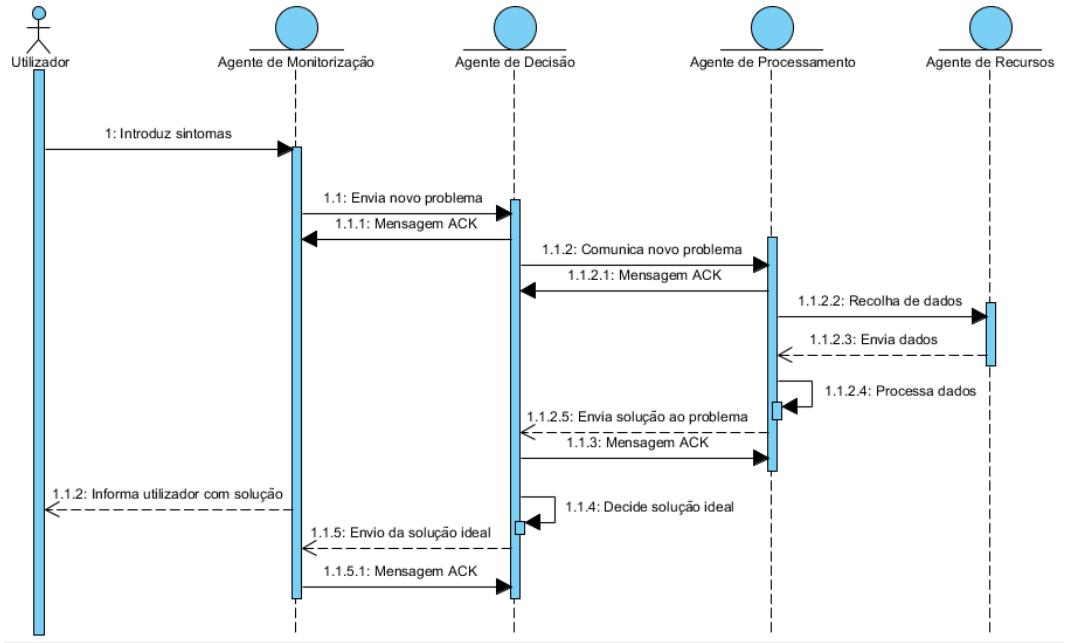


Figura 20: Representação da comunicação entre os agentes

11

IMPLEMENTAÇÃO

11.1 MODELAÇÃO DIMENSIONAL

A simplicidade é a característica que levou à implementação de técnicas de modelação dimensional, no entanto existem outras razões para o desenvolvimento deste tipo de estruturas para além da sua simplicidade e rapidez. Como tal, a implementação deste modelo dimensional foi realizada seguindo a abordagem de um esquema em estrela (os modelos dimensionais implementados em sistemas de gestão de base de dados relacionais são definidos como *star schemas* devido à sua semelhança com uma estrela). Esta abordagem faz com que o sistema possas ser usado para fazer suposições seguras e fundadas mas também para tornar as interfaces do utilizador mais compreensíveis com tempos de processamento e resolução mais eficientes.

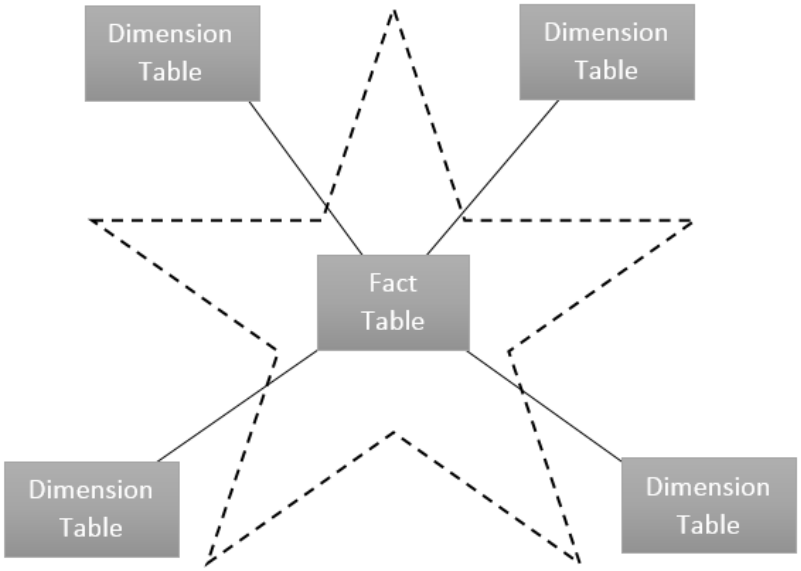


Figura 21: Esquema em estrela

Cada processo de negócio, representado por um modelo dimensional, contem uma tabela de fatos (*Fact Table*) que por sua vez contem as medidas numéricas dos eventos desse processo e as dimensões de análise na altura desse acontecimento (*Dimension Table*). Uma tabela de factos armazena as medi-

das de desempenho dos eventos do processo de negócio. Um principio fundamental para a modelação dimensional passa por considerar um evento como um único registo na tabela de factos correspondente. Adicionalmente, integradas nas tabelas de factos, as tabelas dimensionais contem o contexto associado ao evento do processo de negócio. A modelação dimensional desenvolvida e considerada neste projeto segue uma abordagem discriminada em quatro fases distintas. As fases são sequências e passam pela seleção do processo de negócio, declaração do grão, escolha das dimensões a aplicar a cada tabela de factos e por último a identificação dos factos numéricos que virão a ser preenchidos em cada tabela de factos.

11.1.1 Processo de Negócio

Podemos considerar um processo de negócio como uma atividade de baixo nível de uma organização, tal como a realização de procedimentos de saúde. Estes processos normalmente são expressados sob forma de verbos de ações porque representam atividades e são suportados pelos sistema operacionais.

É através da análise dos requisitos, previamente recolhidos sobre os sistemas operacionais, ou no caso deste trabalho, sobre o conjunto de dados, que se identificam os processos de negócio que fazem parte da modelação. O conjunto de dados apresenta um conjunto total de 76 atributos dos quais apenas 20 foram utilizados. Cada registo contem dados sobre o levantamento realizado a um determinado paciente sobre a sua condição e predisposição para doenças cardíacas. Em detalhe, a caracterização do processo de negócio pode ser visualizada na primeira parte da tabela de caracterização do data mart.

Identificação: Tratamento
Descrição Geral: Com base na atividade realizada em diversos hospitais, destaca-se o interesse sobre a relação entre os custos de tratamentos efetuados a paciente que sofreram de acidentes vasculares cerebrais. Esta análise possibilita um suporte à tomada de decisão de modo que haja uma melhor perceção dos custos e quais os tratamentos efetuados numa ocorrência

11.1.2 Declaração do Grão

Após identificar o processo de negócio é necessário decidir sobre a granularidade, que diz respeito ao nível de detalhe existente no modelo dimensional, ou seja, consiste na especificação de um registo individual da tabela de fatos. Esse modelo deve ser desenvolvido com base na informação mais atómica recolhida na análise de requisitos. Quanto mais atomicidade houver nas medidas mais é o nível de detalhe sobre a informação conhecida, ou seja, quanto mais baixo for o nível de granularidade mais robusto é o desenho do modelo, possibilitando uma melhor análise, melhores repostas a interrogações inesperadas não limitando o acesso a informação num nível de mais detalhe.

Tendo em consideração todos estes aspetos e a análise feita sobre o modelo de negócio definiu-se o grão como sendo o seguinte.

"Duração e custos associados aos tratamentos de um paciente que sofreu de um acidente vascular cerebral num determinado dia realizados num hostpital."

Após a escolha do processo de negócio e da escolha do grão asseguramos que o *data warehouse* é orientado ao assunto, modelado de acordo com as necessidades de uma organização, assegurando a primeira característica definida por William Inmon.

11.1.3 Escolha das Dimensões

Uma vez que um sistema *data warehousing* deve variar no tempo uma das dimensões base, sem analisar qualquer processo de negócio, deve ser a de tempo, calendário. Isto permite descobrir tendências e analisar grandes quantidades de dados e ainda a capacidade de mudança do próprio sistema ao longo do tempo.

Com a definição do grão podemos identificar as dimensões principais associadas à tabela de factos respetiva. É necessário construir uma tabela de fatos com um conjunto de dimensões robustas que representem todas as descrições possíveis que assumem valores no contexto do evento (Kimball et. al. 2013)

*"Duração e custos associados aos tratamentos de um **paciente** que sofreu de um **acidente vascular cerebral** num determinado **dia** realizados num **hospital**."*

Neste sistema destaca-se, na definição do grão, entidades como paciente, hospital e dia (calendário). Complementarmente à escolha das dimensões, pode-se identificar os atributos individuais de cada uma para as aperfeiçoar.

A relação entre dimensões e tabelas de factos constitui a granularidade dessa mesma tabela de factos podendo ser visualizada na tabela seguinte sobe a forma de matriz de decisão.

Tabela 5: Matriz de Decisão

Dimensões \ Tabela de Factos	AVC-F
Calendário	✓
Paciente	✓
Hospital	✓
Tratamento	✓

Cada dimensão, independentemente do seu tipo, foi identificada e caracterizada de acordo com a sua função e o seu esquema. A tabela seguinte descreve todas as dimensões presentes neste projeto.

Tabela 6: Caracterização das Dimensões

Nr	Identificação	Descrição	Esquema (Tipo)
1	Calendário	Esta é a dimensão temporal. Acolhe todos os atributos que sustentem análises ao longo do tempo, como data, mês, semana, ano, etc.	Normal
2	Paciente	Identificação e caracterização do pacientes com os respetivos dados pessoais	Normal
3	Hospital	Caracterização dos hospitais através do nome.	Normal
4	Tratamento	Identificação dos vários tipos de tratamento	Normal com Tabela de Ponte

11.1.4 Identificação dos Fatos

Esta última fase permite identificar e caracterizar os factos presentes em cada tabela e devem ser tantos quantos os encontrados dentro do contexto do grão declarado, no entanto, como nem todos as medidas entram em conformidade com o grão definido este pode vir a ser alterado de acordo com medidas de extrema importante para a organização. Os factos são determinados através da interrogação às medidas que estão na base do processo de negócio.

Similar ao processo de declaração do grão foram identificadas medidas nas diversas tabelas de factos.

*"**Duração** e **custos** associados aos tratamentos de um paciente que sofreu de um acidente vascular cerebral num determinado dia realizados num hospital."*

Tabela 7: Fatos da tabela de fatos

Nr	Identificação	Domínio	Tipo(Função)	Descrição	Exemplo
1	duracao	Numérico	A	Tempo de tratamento em minutos	1000
2	custo	Décimal(20, 5)	A	Custo associado aos tratamentos	286,12

11.1.5 Modelo Concetual

Através da notação de Golfarelli & Rizzi foi possível organizar os processos realizados nas seções anteriores resultando num modelo concetual de fácil perceção capaz de transparecer, sob a forma de hierarquias, as várias vertentes do sistema bem como a diversas organizações dentro de cada hierarquia.

A figura seguinte representa o modelo concetual relativo aos eventos de AVC onde se denota, de forma multidimensional, todas as abordagens relativamente às medidas, ou seja, os diferentes pontos de vista sobre determinada informação.

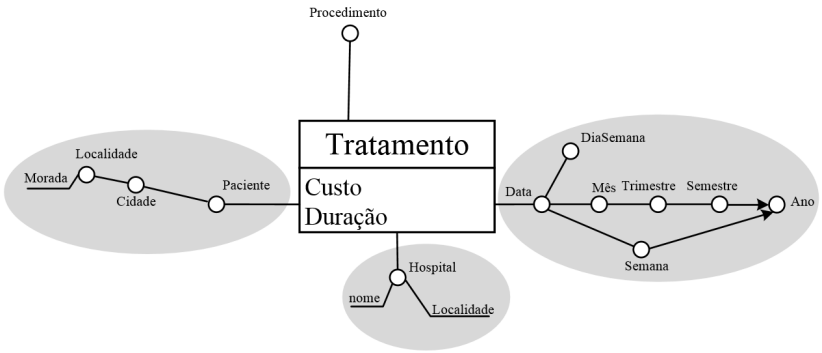


Figura 22: Modelo concetual

11.1.6 Modelo Dimensional

Com o objetivo de modelar os dados para que estes possam ser visualizados de forma simples e direta proporcionando uma perspetiva de análise ampla de todo processo dentro do modelo de negócios.

O modelo desenvolvido é orientado ao assunto, relacionando os factos de acordo com as diferentes perspetivas. De acordo com o mapeamento e análise realiza sobre o sistema operacional alvo, que serve como fonte de dados, construiu-se um modelo em que o esquema consiste num floco de neve devido à necessidade de normalização de certas dimensões.

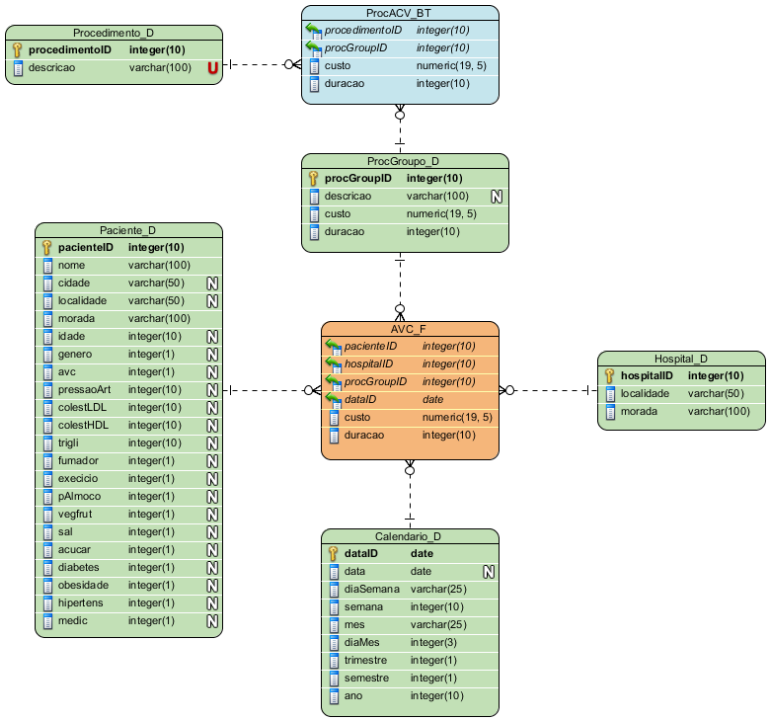


Figura 23: Modelo Dimensional

11.2 SISTEMA MULTIAGENTE

O sistema multiagente foi desenvolvido através da linguagem de programação JAVA. O projeto em si, dentro de uma abordagem mais técnica é subdividido em partes que representam níveis de abstração diferentes tendo em conta uma boa Engenharia de Software. Estes níveis são definidos através de pacotes criados na ferramenta Eclipse. Dentro disso, existem então os seguintes níveis de abstração:

- **entity:** Aqui são definidas as principais entidades que fazem parte da análise de requisitos e que, no fundo, podem ser no seu conjunto um diagrama de classes.
- **dal:** Representa o acesso aos dados contidos na base de dados, *Data Access Layer*, aqui é possível estabelecer uma conexão entre a aplicação e o sistema de gestão da base de dados.
- **dao:** De uma maneira geral, este nível representa as funcionalidades de cada uma das entidades, contudo não as implementa.
- **bll:** *Business Logic Layer*, é neste nível de abstração que a funcionalidades identificadas no nível anterior são implementadas de modo a obter informação significante e necessária para a manutenção da aplicação.
- **agentes:** Todos os agentes que integram o sistema multiagentes são identificados e funcionalmente construídos neste nível.
- **gui:** *Graphical User Interface*, toda a aplicação é apresentada por uma interface que aqui é detalhada em seções.
- **controllers:** Eventos de controlo e apoio ao nível anterior. São executadas ações conforme o tipo de evento solicitado.

A constituição, em termos de interface, representa separadamente os três tipo de agentes que interagem com a aplicação, agente de monitorização, agente de decisão e agente de processamento (o agente de recursos é omitido uma vez que está integrado no agente de processamento). No que diz respeito ao agente de monitorização, ilustrado na figura seguinte, foi desenvolvido um formulário capaz de, dinamicamente, recolher os sintomas de casos novos. Após o envio do novo caso, o agente fica em modo de espera até que uma solução para o caso submetido seja apresentado.

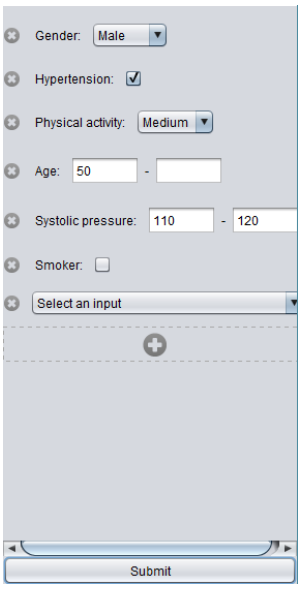


Figura 24: Formulário para recolha de sintomas

O agente de decisão é responsável pelo processo de decisão através do qual avalia os agentes selecionados na participação da resolução do problema. Após receber o novo caso, o agente de decisão reenvia o mesmo caso para os agente de processamento aguardando uma solução de cada participante. Os resultados dessa avaliação são representados em forma de gráficos para que haja uma constante monitorização e gestão do sistema por parte dos utilizadores.

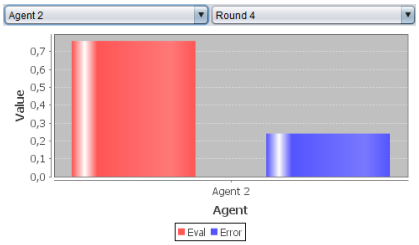


Figura 25: Avaliação dos agentes

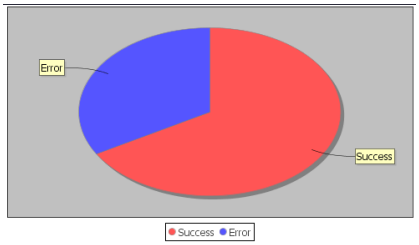


Figura 26: Desempenho do sistema

O processo deliberativo e raciocínio do sistema é implementado nos agentes de processamento com o apoio dos agentes de decisão.

Para alcançar a solução é utilizado um processo de aprendizagem não supervisionado. O algoritmo *Expectation Maximization (EM)*, ideal para resolver problemas com estimativas de parâmetros em redes neuronais artificiais, além disso, este algoritmo tem uma grande performance quando fornecida informação parcial ou valores desconhecidos. Existem duas fazes que constituem este algoritmo, *E-step*, onde os dados desconhecidos é estimada resultando em valores expectáveis, *M-step*, que maximiza o similaridade entre os dados re-estimando os parâmetros do modelo.

A integração desta técnica é considerada uma otimização melhorando o processo de seleção usado nos sistemas de raciocínio baseado em casos reduzindo o numero de casos a considerar.

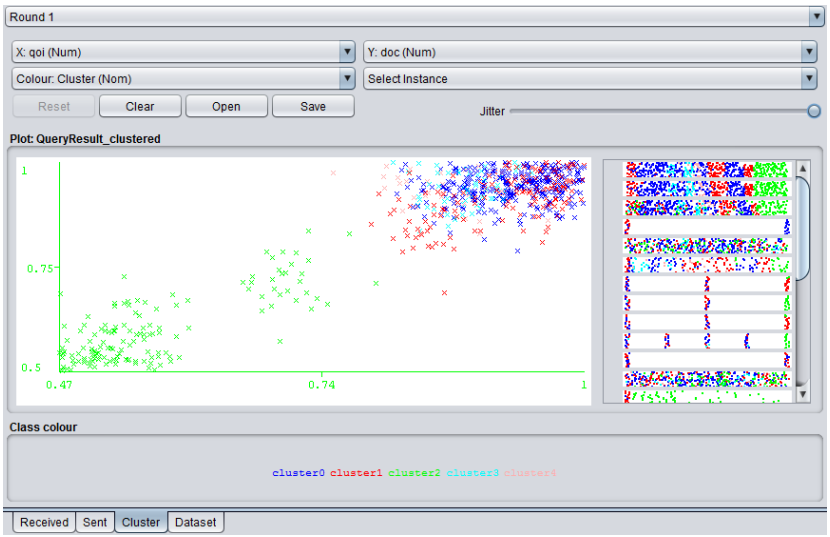


Figura 27: Estrutura dos *clusters* de um agente

Apresentado na arquitetura do sistema, a comunicação estabelecida é especificada sobre canais TCP/IP, com a troca de mensagem de reconhecimento entre cada mensagem.

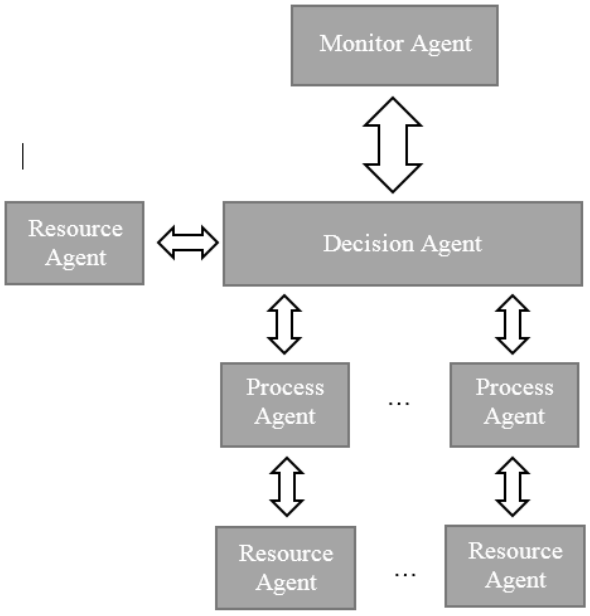


Figura 28: Via de comunicação dos agentes

12

DISCUSSÃO DE RESULTADOS

Como forma de analisar o trabalho desenvolvido foi aplicada uma análise SWOT que é composto por quatro pontos essenciais que denotam as iniciais da ferramente de análise, *Strenghts* (Pontos fortes), *Weaknesses* (Pontos fracos), *Opportunities* (Oportunidades) e *Threats* (Ameaças). Esta análise, desenvolvida pela Universidade de Stanford é apresentada sob forma de matriz com a representação da sigla (SWOT).

Figura 29: Matriz SWOT

Este tipo de análise permite permite realizar um síntese do trabalho desenvolvido, identificando os elementos chave para facilitar a abordagem estratégica e estabelecer prioridades para melhorar a resolução de um determinado problema. Uma vez que esta metodologia permite avaliar os riscos e problemas a resolver de uma determinada solução é também bastante fácil identificar pontos fortes e oportunidades que visam colmatar essas falhas na solução.

A análise SWOT, desenvolvida na escola de design de *Harvard*, procura dividir e analisar as capacidade internas com o objetivo de identificar força e fraquezas e o ambiente externo analisando oportunidades e ameaças.

O pontos fortes consistem nas características positivas e vantagens que devem ser exploradas com o objetivo de atingir as metas estabelecidas. Por outro lado temos as características fracas que podem inibir ou dificultar o desempenho dos processos prejudicando o funcionamento geral. A nível externo, são consideradas oportunidades, características com potencial de influenciar positivamente o processo de desenvolvimento cujo controle não seja exercido pela entidade sob análise, ao contrário das ameaças que influenciam negativamente esse mesmo processo de desenvolvimento.

A análise SWOT deste projeto é apresentada na tabela seguinte.

	Ajuda (a alcançar o objetivo)	Atrapalha (a alcançar o objetivo)
Interna (organização)	• A enorme quantidade de dados armazenada nos sistemas operacionais de saúde possíbilitam a recolha de diversos casos de estudo. • A representação de conhecimento e raciocínio permite representar os problemas em estruturas consistentes e dinâmicas com diversa informação melhorando a *performance* de ferramentas preditivas. • Tradução de medidas sem pré-processamento e de dificil representação em estrutura interpretáveis. • Armazenamento de dados representando/transparecendo informação.	• O diagnóstico de acidentes vasculares cerebrais pode ser bastante complexo. • Dividir a informação por diferentes agentes pode dificultar o processo de decisão devido à falta de caracteristicas comuns ou poucos casos. • O sistema desenvolvido necessita de uma quantidade de casos razoável para poder iniciar e sugerir soluções com um grau de confiabilidade bom. • O sistema não substitui na integra a decisão do utilizador. • Nao existe verificação a nível da integridade e veracidades dos dados a não ser o grau de confiança e qualidade da informação.
Externa (ambiente)	• Utilização do sistema em ambiente hospitalar como auxilio para o processo de tomada de decisão com o objetivo de aumentar a base de conhecimento. • Disponibilização de dados em grande escala poderá facilitar e complementar a análise do processo.	• A resposta dos utilizadores do sistema é essencial para o seu processo de aprendizagem. • Os dados podem ser considerados um obstáculo quando não são obtidos de forma correta e mesmo quando são incompletos. • Restrição no acesso a dados. • O preenchimento de dados incoerentes pode levar a solução sem qualquer tipo de relacionamento com o novo problema.

13

CONCLUSÕES E TRABALHO FUTURO

Hoje em dia, as organizações de saúde lidam com uma quantidade consideradamente elevada de informação, o que faz com que sejam incapazes de analisar os relacionamentos, e outros aspetos, entre os dados. O uso de técnicas de *data mining* providencia e melhora o desempenho, relacionando a informação, do processo de tomada de decisão. Quando exposto a estes processos, todos os registos históricos representam uma parte do relacionamento do conhecimento cuja existência poderá influenciar o resultado de novos casos. O diagnóstico de predisposição para acidentes vasculares cerebrais demonstrou ser um tarefa difícil, uma vez que os valores dos argumentos da extensão dos predicados que como um conjunto descrevem a doença, não são totalmente representados por dados objetivos. Estas características colocam esta problemática na área dos problemas que podem ser enfrentados por metodologias baseadas em inteligência artificial e técnicas de resolução de problemas.

Este trabalho apresenta a fundação de um quadro coerente que utiliza técnicas sem precedentes e poderosas de representação do conhecimento e raciocínio para definira estrutura da informação e dos seus mecanismos computacionais. Esta constatação baseia-se em:

- Dados são diferentes de informação e conhecimento;
- A tradução de medidas sem pré processamento em termos de informação incompleta, desconhecida e até mesmo contraditória em estruturas interpretáveis e acionáveis é um desafio;
- Análise podem detetar marcadores e alvos candidatos, sem pré-conceção, ou seja, por exemplo, saber como informações pessoais e fatores de risco podem afetar a predisposição para doenças hepáticas.

Este método apresenta uma nova abordagem que pode revolucionar as ferramentas de previsão em todos os sentidos, tornando-as mais completas do que a existentes metodologias e viáveis para a resolução de problemas.

Uma vez que os valores e qualidade dos dados variam consideravelmente, é possível dividir todos os registo em grupos diferentes de acordo com as suas características de forma a melhorar a análise seguinte. Esta abordagem de representação de conhecimento e raciocínio permite o uso de valores normalizados dos limites do intervalo das cláusulas ou termos dos argumentos que constituem a extensão do predicado $stroke$ e o seus valores de DoC, para entrada da rede neuronal artificial. Com a

RNA implementada, todos os casos podem ser corretamente classificados considerando o modelo e os dados históricos, em que o resultado traduz o risco de predisposição de um paciente para um acidente vascular cerebral bem como a confiança dessa análise.

Face à implementação e uso das técnicas de *data mining*, representação de conhecimento e raciocínio e raciocínio baseado em casos foi possível simular, num sistema real, a predisposição para acidentes vasculares cerebrais. Uma vez que este sistema funciona com base na experiência e confirmação dos resultados dos utilizadores, os resultado podem variar de acordo com a utilização. No entanto, o sistema incorpora uma capacidade de aprendizagem característica dos sistemas de raciocínio baseado em casos que possibilita avaliar e repreender soluções erradas.

Em suma, os resultados obtidos pelo sistema foram satisfatórios uma vez que todos os agentes devolveram respostas, com soluções adequadas, para um novo caso com um grau de similaridade bastante próximo em relação às características entre o novo caso e os existentes.

Em termos de trabalho futuro, estima-se utilizar a mesma abordagem de representação de conhecimento e raciocínio noutras áreas ou noutro âmbito, uma vez que se apresentou uma técnica universal para traduzir o universo de discurso capaz de ser adaptado em diferentes circunstâncias.

BIBLIOGRAFIA

[1] Isken, M. and Rajagopalan, B. Data mining to support simulation modeling of patient flow in hospitals, Journal of Medical Systems, 26, 2002, pp. 179-197.

[2] Soni, S. Vyas, O.P. Using Associative Classifiers for Predictive Analysis in Health Care Data Mining, International Journal of Computer Applications, 2010.

[3] Azari, A. An Ongoing Research Projection Dynamic Prediction of Length Of Stay. 2013 IEEE International Conference on Healthcare Informatics (ICHI)

[4] Liu, P. Lei, L. Yin, J. Zhang, W. Naijun, W. El-Darzi, E. Healthcare Data Minig: Prediction Inpatient Length of Stay, 3rd International IEEE Conference Intelligent Systems, September 2006

[5]Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, second Edition, (2006).

[6]Ian Written and Eibe Frank, "Data Mining, Practical Machine Learning Tools and Techniques",3rd Ed., Morgan Kaufmann, 2011.

[7]LAROSE, D. T. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley and Sons, Inc, 2005.

[8]S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Data Preprocessing for Supervised Leaning", International Journal of Computer Science, 2006, Vol 1 N. 2 ISSN 1306-4428 Timonthy C. Havens. Clustering in relational data and ontologies, July 2010

[9] Kimball, R., Ross, M. The Data Warehouse Toolkit Second Edition - The Complete Guide to Dimensional Modeling, John Wiley & Sons, Inc., 2ª Edição, 2002;

[10] Kimball R. Reeves L., Ross M., Thornthwaite W., The Data Warehouse Lifecycle Toolkit - Expert Methods for Designing, Developing, and Deploying Data Warehouses, John Wiley & Sons, Inc, 1998;

[11] Kimball R. Caserta J., The Data Warehouse ETL Toolkit – Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data, John Wiley & Sons, Inc, 2004;

[12] Golfarelli M., Rizzi S., Maio R., Conceptual Design of Data Warehouse from E/R Schemes, Published in the Proceedings of the Hawaii International Conference On System Sciences, 1998;

[13] Amarenco P. Labreuche J, Touboul PJ. 2008. High-density lipoprotein-cholesterol and risk of stroke and carotid atherosclerosis: a systematic review. In Atherosclerosis, Vol. 196, pp. 489–496;

[14] Bhat, VM., Cole, JW., Sorkin, JD., Wozniak,. MA, Malarcher, AM., Giles, WH., Stern, BJ., Kittner, SJ. 2008. Dose-response relationship between cigarette smoking and risk of ischemic stroke in young women. In Stroke, Vol. 39, pp. 2439–2443;

[15] Caldeira, A., Arteiro, J., Roseiro, J., Neves, J., Vicente, H. 2011. An Artificial Intelligence Approach to Bacillus amyloliquefaciens CCMI 1051 Cultures: Application to the Production of Antifungal Compounds. In Bioresource Technology, Vol. 102, pp. 1496-1502;

[16] Carneiro, D., Novais, P., Andrade, F., Zeleznikow, J., Neves, J. 2013. Using Case-Based Reasoning and Principled Negotiation to provide decision support for dispute resolution. In Knowledge and Information Systems, Vol. 36, pp. 789-826;

[17] Cortez, P., Rocha, M., 2004. Evolving Time Series Forecasting ARMA Models. In Journal of Heuristics, Vol. 10, pp. 415-429;

[18] Gelfond M., Lifschitz V., 1988. The stable model semantics for logic programming. Logic Programming – Proceedings of the Fifth International Conference and Symposium. pp. 1070-1080;

[19] Go A. S. et al, 2014. Heart disease and stroke statistics — 2014 update: a report from the American Heart Association. In Circulation, Vol. 129, pp. e28–e292;

[20] Grau A. J. et al, 2009. Association between recent sports activity, sports activity in young adulthood, and stroke. In Stroke, Vol. 40, pp. 426–431;

[21] Halpern, J., 2005. Reasoning about uncertainty. MIT Press, Massachusetts, USA;

[22] Kakas A. et al, 1998. The role of abduction in logic programming. Handbook of Logic in Artificial Intelligence and Logic Programming. Vol. 5, pp. 235-324;

[23] Dinevski, D. et. al. Clinical Decision Support Systems

[24] AL-Gamdi, A., Albeladi, K., AlCattan, R., 2014 Clinical Decision Support System in HealthCare Industry Success and Risk Factors, Vol. 11

[25] Bhatt, G., Zaveri, J., 2001. The enabling role of decision support systems in organizational learning. Decision Support Systems 32. Vol. 11, pp.297 – 309

[24] Asemi, A., Safari, A., Zavareh, A., 2011. The Role of Management Information System (MIS) and Decision Support System (DSS) for Manager's Decision Making Process . International Journal of Business and Management. Vol. 6, No. 7

[25] Wilke, W., Bergmann, R., 1998. Techniques and Knowledge used for Adaptation during Case-Based Problem Solving.

[24] Delany, S., Cunningham, P. The Application of Case-Based Reasoning to Early Software Project Cost Estimation and Risk Assessment

[26] Pal, S., Shiu, S., 2004. Foundations of Soft Case-Based Reasoning, Wiley-Interscience

[27] Kolodner, J., 1992. An Introduction to Case-Based Reasoning. Artificial Intelligence Review. Vol. 6, pp. 3-34

[28] Aamodt, A., Plaza, E., 1994. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications. IOS Press, Vol. 7: 1, pp. 39-59

[29] Masethe, H., Masethe, M., 2014. Prediction of Heart Disease using Classification Algorithms. Proceedings of the World Congress on Engineering and Computer Science. Vol. 2, pp.22–24

[30] Letham, B., Rudin, C., 2013. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model.

[31] Wadhonkar, M., Tijare, P., Sawalkar, S., 2013. Classification of Heart Disease Dataset using Multilayer Feed forward backpropogation Algorithm. Internacional Journal of Application or Innovation in Engineering and Management. Vol. 2, Issue 4

[32] Gopal, S., Radhakrishna, Sadhana, 2015. Cardiovascular Disease Dataset Exploration Using Hive and R. International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 5, Issue 4

[33] Chitra, R., Manju, T., Priya, K., 2015. Predictive Model Of Stroke Disease Using Hybrid Neuro-Genetic Approach. International Journal of Research in Science and Engineering. Vol. 3, Issue 2

[33] Neves, J., Ribeiro, J., Pereira, P., Alves, V., Machado, J., Abelha, A., Novais, P., Analide, C., Santos, M., Fernández, Delgado, M.: Evolutionary intelligence in asphalt pavement modeling and quality-of- information. Progress in Artificial Intelligence.

[34] Storms, P., Grant, T. Agent Coordination Mechanisms for Multi-National Network Enabled Capabilities. Progress in Artificial Intelligence.

[35] Muller, J., 1998. Architectures and applications of intelligent agents: A survey. The Knowledge Engineering Review, Vol. 13: pp. 353-380

[36] Stone, P., Veloso, M., 2000. Multiagent Systems: A Survey from a Machine Learning Perspective. Autonomous Robotics, Vol. 8: 3

[37] Luhn, H., 1958. A Business Intelligence System. IBM Journal

[38] Chee, T., Chan, L., Chuah, M., Tan, C., Wong, S., Yeoh, W., 2009. Business Intelligence Systems: State-of-the-art review and contemporary applications. Symposium on Progress in Information & Communication Technology

[39] Khan, R., Quadri, S., 2012. Dovetailing of Business Intelligence and Knowledge Management: An Integrative Framework. Information and Knowledge Management, Vol. 2: 4

[40] Mozaffarian, D., et. al., 2014. Heart Disease and Stroke Statistics—2015 Update

[41] Neves J., 1984. A logic interpreter to handle time and negation in logic data bases. Proceedings of the 1984 Annual Confer-ence of the ACM on the Fifth Generation Challenge, pp. 50-54.

[42] Neves J. et al, 2007. The halt condition in genetic programming. In Progress in Artificial Intelligence – Lecture Notes in Computer Science, Vol. 4874, pp. 160-169.

[43] Lindgren, A., Risk Factors. 2014. Oxford Textbook of Stroke and Cerebrovascular Disease. Oxford University Press, Oxford, pp. 9-18.

[44] Khoury J. C. et al, 2013. Diabetes mellitus: a risk factor for ischemic stroke in a large biracial population. In Stroke, Vol. 44, pp.1500–1504.

[45] McDonnell M. N. et al, 2013. Physical activity frequency and risk of incident stroke in a national US study of blacks and whites. In Stroke, Vol. 44, pp. 2519–2524.

[46] Shah R. S. and Cole J. W., 2010. Smoking and stroke: the more you smoke the more you stroke. In Expert Review of Car-diovascular Therapy, Vol. 8, pp. 917–932.

[47] Seshadri S. et al, 2006. The lifetime risk of stroke: estimates from the Framingham Study. In Stroke, Vol. 37, pp. 345–350.

[48] Zhang Y., et al, 2012. Total and high-density lipoprotein cholesterol and stroke risk. In Stroke, Vol. 43. pp. 1768–1774.

[49] Coutinho, C. P., Sousa, A., Dias, A., Bessa, F., Ferreira, M. J., & Vieira, S. (2009). Investigação-acção: metodologia preferencial nas práticas educativas. Revista Psicologia, Educação e Cultura, 13:2, pp.355-379.

[50] José Neves, Henrique Vicente, Nuno Gonçalves, Ruben Oliveira, António Abelha and José Machado, Artificial Neural Networks in Stroke Predisposition Screening, in Proceedings of the International Conference on e-Society 2015, Madeira, Portugal, 2015.