Universidade do Minho
Escola de Engenharia

Mariana Isabel da Silva Pinto Ferreira

Detecção e Prevenção de Fraude

Outubro de 2009

Mariana Isabel da Silva Pinto Ferreira **Detecção e Prevenção de Fraude**

UMinho | 2009

Universidade do Minho
Escola de Engenharia

Mariana Isabel da Silva Pinto Ferreira

Detecção e Prevenção de Fraude

Tese de Mestrado
Mestrado em Informática

Trabalho efectuado sob a orientação de:
Professor Doutor Paulo Novais
Professor Doutor Lino Costa

Outubro de 2009

Agradecimentos

Embora uma dissertação seja, pela sua finalidade académica, um trabalho individual, há contributos de natureza diversa que não podem, nem devem, deixar de ser referidos. Por esse motivo, desejo expressar os meus sinceros agradecimentos:

Aos meus orientadores da Universidade do Minho, professor Lino Costa e professor Paulo Novais, cuja orientação, permanente disponibilidade e indicação das importantes fontes de informação, cujo apoio durante a fase de recolha bibliográfica me permitiu arrecadar um manancial de informação que em muito contribuiu para a execução da dissertação.

Ao João Mário Fernandes pela disponibilidade sempre manifestada, ajuda na fase de geração de dados que me permitiu promover o desenvolvimento do gerador de CDRs e pelas questões levantadas ao longo destes meses, que me obrigaram a reflectir mais aprofundadamente sobre determinados aspectos.

À Wedo Consulting, em especial ao Ricardo Marques e à Margarida Almeida que me proporcionaram o tema tratado nesta dissertação. Pelas críticas, sugestões e disponibilidade revelada ao longo do desenvolvimento da dissertação.

Ao Eugénio Rosas e ao Mauro Almeida pela disponibilidade e ajuda em questões pontuais, bem como pela amizade demonstrada.

À Sílvia Santos e à Joana Ribeiro pelo incentivo e apoio nos momentos bons e menos bons, e pela amizade de longa data.

Ao José Monteiro pela paciência, compreensão e pelo ânimo e incentivo que me deram estímulo para desenvolver esta dissertação ao longo destes meses.

Aos meus pais, pelo estímulo e apoio incondicional desde a primeira hora, pela paciência e grande amizade com que sempre me ouviram, e sensatez com que sempre me ajudaram.

Resumo

A detecção de fraude nas telecomunicações passa por conhecer os métodos de fraude existentes e estudar as actividades irregulares dos subscritores. Apenas conseguimos perceber uma intenção fraudulenta após esta ter ocorrido mas surge também a necessidade de perceber se o subscritor irá ter actividades fraudulentas antes de as cometer. Perceber a intenção de um subscritor passa por entender o seu comportamento através das suas actividades na rede. Contudo, nunca é possível determinar com exactidão a ocorrência de fraude. Os dados detectados necessitam de ser analisados por um analista e as actividades do subscritor verificadas para perceber a sua intenção.

Fingerprinting é o método de prevenção de fraude que passa pela detecção de uma similaridade entre o comportamento de um novo subscritor na rede para com os dados recolhidos de actividades de subscritores previamente sob investigação ou submetidos para tal. O *Data Mining* (DM) através do uso de algoritmos específicos de pesquisa tenta descobrir padrões discerníveis e inferir regras para os mesmos [1]. Nesta dissertação é investigado o potencial de algoritmos supervisionados de modelos, de DM, de quatro paradigmas distintos, para implementar o método *fingerprinting*: Raciocínio Baseado em Casos (RBC), Redes Neuronais (RNs), Árvores de Decisão e Redes de Bayes.

Os dados dos subscritores são gerados por uma ferramenta desenvolvida com o objectivo de simular eventos de subscritores, normais e fraudulentos, numa rede de telecomunicações para as fases de treino e teste dos modelos de DM.

Recorrendo a uma ferramenta que implemente algoritmos de DM é possível fazer uma análise comparativa dos resultados obtidos, de experiências de classificação e regressão, avaliando matrizes de confusão, curvas *Receiver-Operating Characteristic* (ROC), precisão e tempos de treino. Importa referir que a análise dos resultados obtidos é condicionada ao facto dos dados serem simulados e os modelos não tirarem partido de um volume real de dados de casos de fraude.

Abstract

Fraud detection in telecommunications is to know the methods of fraud patterns and study the irregular activities of subscribers. We can only realize a fraudulent intent after it occurred but there is also the need to achieve to whether the subscriber will have fraudulent activities before committing them. Understanding the intent of a subscriber is to understand their behavior through their activities on the network. However, it is never possible to exactly determine the occurrence of fraud. The detected data needs to be reviewed by a business analyst and the subscribers' data must be reviewed in order to verify their intent.

Fingerprinting is a method of fraud prevention that involves the detection of a similarity between the behavior of a new subscriber on the network with the data collected from subscribers activities previously under investigation or subjected to such. Data Mining (DM) *tries to discover patterns and infer rules in the data through the use of specific algorithms* [1]. *This dissertation investigates the potential of supervised algorithms of* DM *models from four distinct paradigms, to implement the fingerprinting method: Case-based Reasoning* (CBR)*, Neural Networks* (RNs)*, decision trees and Bayesian networks.*

The subscribers' activities data is generated by a developed tool to simulate events of normal and fraudulent subscribers' activities in a telecommunications network. These data will be used for the training and test of the DM *models.*

Using a tool to implement data mining algorithms allows making a comparative analysis of classification and regression experiments results. The evaluation is made with confusion matrices, Receiver-Operating Characteristic (ROC) *curves, accuracy and training time. It is worth noting that the analysis of results is conditioned to the fact that data is simulated and the models do not take advantage of an actual volume of data fraud.*

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

Este capítulo fornece uma visão geral da dissertação. Começa com uma introdução acerca de fraude nas telecomunicações e descreve a motivação e as condicionantes para a realização deste trabalho. A secção seguinte descreve o problema e as diversas hipóteses de pesquisa abordadas. De seguida, são discutidos os objectivos técnicos e a metodologia da dissertação. Por fim, é apresentada a estrutura da dissertação.

1.1. Fraude nas Telecomunicações

A fraude é caracterizada pelo roubo, tipicamente caracterizado pela prova da intenção onde as perdas resultantes não são na maior parte das vezes recuperáveis e podem ser detectadas através da análise de padrões de chamadas. Fraude é diferente de perda de receitas. A perda de receitas é caracterizada por resultar de buracos operacionais ou técnicos onde as perdas resultantes são muitas das vezes recuperáveis e geralmente detectados através de auditorias ou procedimentos similares.

As telecomunicações são um alvo atractivo para os fraudulentos. Os fornecedores de serviços são atingidos por pedidos de serviços fraudulentos acima de 85% das vezes [2]. São implementados esquemas altamente sofisticados por crime organizado usando *hackers* e algoritmos de aprendizagem.

A associação de controlo de fraude nas comunicações conduziu um estudo e determinou que cerca de 35 a 40 milhões de dólares em perdas se devem a fraude nas telecomunicações em todo o mundo. Grande parte das grandes operadoras de telecomunicações desenvolveu sistemas de gestão de fraude para combater esta ameaça mas um estudo realizado pelo

fórum internacional de acesso irregular a redes concluiu que 10% das operadoras conseguiram estabelecer estratégias de combate à fraude [2].

Existe uma pressão muito grande para combater efectivamente a fraude. Os accionistas das operadoras pressionam para protegerem as suas receitas. As perdas associadas a outros problemas tais como discrepâncias em interconexão ou dados corrompidos correspondem a cerca de 30% do total das receitas. A fraude é uma das maiores causas singulares de perdas de receitas para as operadoras custando entre 3% a 5% das receitas anuais[2].

A motivação por trás do crime é atribuída a factores como a migração e demografia, inovação das tecnologias, insatisfação dos colaboradores, o desafio, fraquezas operacionais, modelos de negócio pobres, ambição criminosa, lavagem de dinheiro e factores políticos e ideológicos.

As perdas nas receitas originadas pela fraude são aproximadamente iguais à falta de receitas dentro dos sistemas e procedimentos de uma companhia. Os procedimentos de garantia de receitas extraem dados a cada passo da cadeia de receitas submetendo-a a uma verificação de integridade rigorosa. As aproximações baseadas em tempo e custo de processamento de registos de chamadas, *Call Detail Records* (CDRs), por si só estão-se a tornar obsoletas. As operadoras de telecomunicações devem evoluir e investir em métodos de detecção e prevenção de fraude em múltiplas origens inteligentes: serviços, conteúdo, dispositivos de banda larga, relatórios de qualidade dos serviços, etc.

1.2. Motivação

Grandes volumes de dados são coleccionados e armazenados como resultado do aumento do uso dos serviços de comunicações móveis. A informação e conhecimento derivado destes dados podem fornecer às operadoras uma vantagem competitiva em termos de manutenção e retenção do cliente, publicidade e detecção de fraude. Desta forma a fraude nas telecomunicações tornou-se de alta prioridade na agenda da maioria dos operadores de telecomunicações.

Sistemas de análise e detecção de fraude eficazes permitem aos operadores de telecomunicações não perder dinheiro assim como ajudar a restaurar a confiança dos assinantes na segurança das suas transacções. Sistemas de detecção de fraude automatizados permitem aos operadores reagir à fraude através da detecção, negação de serviços e processando os subscritores fraudulentos. O enorme volume de actividades de chamadas numa rede significa que a detecção de fraude e sua análise é um grande desafio.

Em geral, quanto mais avançado é um serviço, mais é vulnerável a fraudes. No futuro, os operadores vão necessitar de se adaptar rapidamente para manterem-se a par da evolução dos métodos de fraude usados. Além disso, o número de pessoas envolvidas na prestação de um serviço tem tendência a aumentar, tornando maiores as possibilidades da fraude se expandir para além do simples caso de um subscritor a tentar defraudar a operadora. Enquanto as aproximações convencionais à detecção de fraude e análise tal como sistemas baseados em regras e valores limite para parâmetros particulares podem ser suficientes para lidar com alguns dos tipos de fraude, os fraudulentos podem mudar as suas tácticas muito facilmente para evitar a detecção.

Portanto, existe a necessidade de considerar aproximações de detecção, prevenção e análise de fraude adaptativas, as técnicas de inteligência artificial providenciam métodos efectivos de resolução de alguns destes desafios.

1.3. Condicionantes

Um CDR é produzido por cada chamada terminada numa rede de telecomunicações móvel. Estes registos contêm informação útil sobre uma chamada efectuada por um subscritor. Além de serem usados para propósitos de facturação constituem uma enorme fonte de dados de onde pode ser extraída informação útil sobre os subscritores. Um exemplo é a detecção de uso anormal da rede móvel de telecomunicações.

Apesar de existirem vários formatos de CDRs este estudo é baseado no *Global System for Mobile Communications* (GSM). Estes CDRs foram gerados de acordo com alguns comportamentos padrão de subscritores fraudulentos através de uma ferramenta desenvolvida durante a fase de desenvolvimento da dissertação. A geração dos dados e não

obtenção de dados reais deve-se a compromissos de confidencialidade por parte das operadoras de telecomunicações para com os clientes.

1.4. Descrição do Problema

Ao longo de um período de tempo, o *Subscriber Identity Module* (SIM) no equipamento de um subscritor gera um padrão extenso de utilização. O padrão de utilização pode incluir chamadas internacionais ou padrões com grandes variações de tempo entre outros. O uso anómalo pode ser detectado dentro de um padrão tal como o abuso de chamadas gratuitas.

O padrão de utilização anormal de uma rede de telecomunicações pode ser classificado como:

- O padrão é intrinsecamente fraudulento, i.e., nunca vai ocorrer durante uma utilização normal;
- O padrão é anormal relativamente ao padrão do historial de chamadas estabelecidas por aquele equipamento.

De forma a conseguir prevenir fraudes do primeiro tipo é necessário ter detectado esses comportamentos e ter conhecimento do historial do uso do SIM. Por isso, pode ser usada uma análise descritiva dos subscritores fraudulentos para ser usada para extracção de conhecimento. A interpretação por modos de *clustering* ou agrupamento de padrões similares pode ajudar a isolar actividades suspeitas de chamadas na rede de telecomunicações. Pode também ajudar a os analistas de fraude nas suas investigações e na análise de padrões de chamadas dos subscritores. Embora os dados das chamadas dos subscritores serem guardados para propósitos de facturação, é interessante saber que não há suposições *a priori* sobre os dados indicativos de padrões de chamadas fraudulentas. Por outras palavras, as chamadas efectuadas para efeitos de facturação não têm qualquer indicativo de fraude. É necessária uma análise posterior para se identificar possíveis actividades fraudulentas. Devido ao enorme volume de chamadas, é virtualmente impossível efectuar uma análise a esses dados sem recorrer a técnicas ou ferramentas sofisticadas. Existe a necessidade de usar técnicas e ferramentas para assistir os humanos de uma forma inteligente na análise de grandes volumes de chamadas. Algumas dessas

técnicas são algoritmos de *Data Mining*, como os que serão investigados nesta dissertação na resolução do problema descrito.

1.5. Objectivo

O mote da dissertação é a detecção e prevenção de fraude no contexto das telecomunicações. No contexto desta dissertação coloca-se a hipótese de, usando processos de *Data Mining*, obter melhores resultados de prevenção de fraude nas telecomunicações. Os objectivos propostos a concretizar com esta dissertação são:

- Criar um cenário de estudo: pretende-se investigar métodos de detecção e prevenção de fraude, dando uma introdução de conceitos de fraude nas telecomunicações para obter conhecimento do negócio e conseguir uma análise objectiva dos resultados;
- Estudar as técnicas de *Data Mining* adequadas: para o objectivo de prevenção de fraude nas telecomunicações, através da identificação de padrões na actividade dos subscritores na rede, pretende-se recorrer às seguintes técnicas:
    - Raciocínio Baseado em Casos: pretende-se estudar o processo de análise de problemas, o modelo da técnica e algoritmos de descoberta de similaridades de casos;
    - Redes Neuronais: pretende-se estudar os conceitos, adaptabilidade e descoberta de padrões e qual o tipo de redes neuronais que mais se adequa à descoberta de padrões arbitrários;
    - Árvores de Decisão: pretende-se estudar os conceitos, estrutura e algoritmos de construção de árvores de decisão de modo a conseguir uma classificação dos subscritores de acordo com o seu comportamento descrito pelos seus eventos na rede;
    - Redes de Bayes: pretende-se estudar o teorema de Bayes e a aplicabilidade da técnica no reconhecimento de padrões e classificação, bem como a implementação do algoritmo de construção de redes *bayesianas*.

- Avaliação dos diferentes modelos: a avaliação dos modelos é efectuada com os resultados obtidos, das experiências de classificação e regressão dos algoritmos dos modelos de *Data Mining* descritos, recorrendo a uma ferramenta que implemente os algoritmos de cada modelo. Devem ser gerados conjuntos de dados para teste e treino dos algoritmos de modo a simular eventos de subscritores, normais e fraudulentos, na rede. Pretende-se perceber a fiabilidade dos modelos recorrendo a técnicas de avaliação de resultados e comparação dos resultados dos modelos como matrizes de confusão, curvas *Receiver-Operating Characteristic* (ROC), precisão e tempos de treino.

Pretende-se com este documento perceber o conceito de *fingerprinting* e estudar modelos comportamentais existentes, de estudo de características e eventos dos subscritores na rede que definam o seu comportamento, de modo a perceber qual o modelo com melhores resultados para ser implementado numa solução de detecção e prevenção de fraude nas telecomunicações desenvolvida pela WeDo Consulting.

1.6. Metodologia da Investigação

A metodologia da investigação da dissertação tem por base o modelo *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) desenvolvida em finais de 1996, pelo consórcio formado pelas empresas NCR (EUA e Dinamarca), DaimlerChrysler AG (Alemanha), SPSS Inc.(EUA) e o Grupo Bancário Holandês (OHRA) [3].

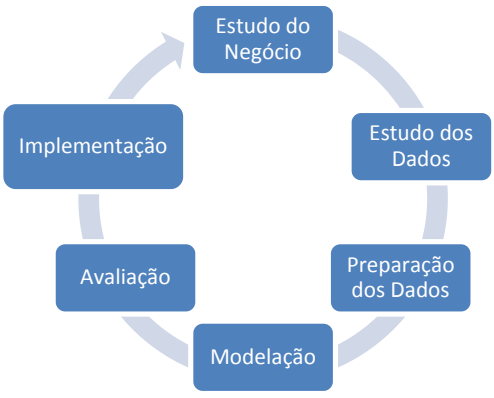

Figura 1 – Ciclo de vida da metodologia CRISP-DM

A metodologia CRISP-DM é descrita em termos de um processo hierárquico, com um ciclo de vida, apresentado na Figura 1, que se desenvolve em seis fases: Estudo do Negócio, Estudo dos Dados, Preparação dos Dados, Modelação, Avaliação e Implementação. As fases não têm uma sequência fixa, dependendo do resultado e desempenho das outras fases ou das tarefas particulares de determinada fase [3].

De seguida são apresentadas cada uma das fases da metodologia com relevância nas características mais proeminentes e como irá ser aplicada ao longo do desenvolvimento da dissertação:

- Estudo do negócio: Centra-se na análise dos objectivos do projecto e nos requisitos (funcionais, técnicos e temporais) segundo a perspectiva do negócio. Este conhecimento é posteriormente utilizado na definição do problema de *Data Mining* e no desenho do plano preliminar para alcançar os objectivos. A fase do estudo de negócio vai compreender uma abordagem à definição de Fraude no contexto das telecomunicações, uma introdução de conceitos e um estudo dos métodos já implementados por ferramentas existentes de detecção e prevenção de fraude. Tem como resultado os objectivos de *Data Mining* e os pressupostos das ferramentas e técnicas a utilizar.

- Estudo dos dados: Começa com a recolha inicial dos dados e prossegue com a sua análise de forma a identificar problemas de qualidade. Esta fase irá passar por demonstrar os critérios para geração de dados de eventos de subscritores de modo a poderem ser usados na fase de modelação.

- Preparação dos dados: Envolve todas as actividades associadas à construção do conjunto final de dados, aquele que será usado pela ferramenta de modelação, agrupado em conjuntos de dados de treino e de teste dos modelos.

- Modelação: Na fase de modelação são seleccionadas várias técnicas de modelação (e.g. Árvores de Decisão, Redes Neuronais Artificiais (RNA), etc.) e os seus parâmetros são ajustados de forma a optimizar os resultados. Inicialmente, foram especificados os problemas e os objectivos do *Data Mining*, mas só nesta fase é que se submetem os dados, previamente preparados, para a modelação, devendo-se

escolher as técnicas mais apropriadas tendo em atenção o tipo de problema, as ferramentas e os objectivos de *Data Mining*.

- Avaliação: Tem como finalidade avaliar a utilidade dos modelos, rever os passos executados na sua construção e verificar se permitem atingir os objectivos de negócio.
- Implementação: Nesta fase é efectuada uma análise aos resultados obtidos da fase de avaliação, verificando qual o modelo ideal a ser implementado numa ferramenta de detecção e prevenção de fraude.

1.7. Estrutura do Documento

Este documento encontra-se organizado em cinco partes lógicas distintas. Na primeira, englobando este capítulo são descritos conceitos básicos de fraude nas telecomunicações, a motivação, condicionantes, descrição do problema e objectivo a atingir com a elaboração da dissertação bem como a metodologia a aplicar no seu desenvolvimento. Na segunda parte são abordados e caracterizados os conceitos de detecção e prevenção de fraude nas telecomunicações, necessários para um entendimento dos capítulos seguintes. O terceiro capítulo é dedicado à abordagem cuidada de um conjunto de modelos e algoritmos de análise de dados comportamentais dos subscritores recorrendo a *Data Mining*. O quarto capítulo consiste na análise dos dados gerados e implementação dos modelos com a ajuda de uma ferramenta de *Data Mining*. Finalmente são discutidos os resultados obtidos com a aplicação dos algoritmos aos dados gerados e conclui-se com uma direcção possível de investigação futura.

Capítulo 2

Detecção e Prevenção de Fraude nas Telecomunicações

A fraude nas telecomunicações é o uso intencional e com sucesso de algum engano, astúcia ou artifício usado para iludir e enganar outra pessoa mesmo que essa pessoa actue pela perda dos seus bens ou lesão jurídica. Mas parece existir um consenso de que a fraude nas telecomunicações, tal como o termo é aplicado, envolve o roubo de serviços ou o abuso deliberado de redes de voz e dados. É ainda aceitável que nestes casos a intenção do autor seja evitar completamente ou reduzir a cobrança legítima dos serviços usados. Em alguma ocasião a evasão da cobrança das chamadas é alcançada enganando os sistemas de facturação e de atendimento ao cliente de modo a cobrarem a pessoa errada.

A fraude é um problema que afecta as grandes operadoras de telecomunicações e as suas receitas anuais. Num mercado competitivo existe pressão para as operadoras aumentarem a sua eficiência, imagem, qualidade de serviço e reduzirem os seus custos.

O comportamento e a intenção dos subscritores podem ser estudados através dos dados registados das suas actividades na rede. Um subscritor novo na rede pode ser identificado como fraudulento quando se verifica um comportamento similar ao de um previamente detectado como tal. As motivações que podem levar um subscritor a ser fraudulento são inúmeras tais como: ganância, malícia, vingança, ameaças, dívidas, entretenimento, ingenuidade ou mesmo ser uma forma de ganhar dinheiro. Os ataques realizados pelos fraudulentos podem ser efectuados através de possibilidades e permutações infinitas e variar desde abusos simples a ataques organizados envolvendo ou instigados por grupos de criminosos externos. Estabelecer identidades no ponto de acesso a serviços é o método mais

efectivo de reduzir tentativas de fraude. Efectuar acções retrospectivas em vez de perseguir uma recuperação financeira consome mais tempo, é mais custoso e menos efectivo.

As medidas protectoras a tomar devem passar por:

- Verificação de subscritores para prevenir o retorno de antigos fraudulentos;
- Saber quais os seus registos de pagamentos antigos e comportamentos na rede;
- Verificar as listas internas de fraudulentos registados;
- Agrupar novos subscritores e monitorizar o uso dos serviços;
- Automatizar todo o tipo de trabalho intensivo de verificação;
- Usar ferramentas de gestão de fraude em *real-time* para detectar actividades fraudulentas, disparar alarmes, notificar os analistas de fraude e gerar relatórios.

A natureza da fraude nas telecomunicações está em transformação constante. Os *service providers* olham para os planos de preço fixo para convergir em serviços baseados em voz, conteúdo e dados. Servem uma variedade de serviços para os clientes baseada numa combinação complexa de redes *wireline*, *broadband* e *wireless*. A emergência destas redes teve impacto nos modelos de negócio e distorceram os conceitos de fraude. Fraude em *roaming* tornou-se uma das causas principais de perdas financeiras para as operadoras de telecomunicações. Várias companhias estão a olhar para a gestão de fraude mantendo uma base e dados com toda a informação dos clientes. Providenciam soluções para eliminar o risco de fraude com soluções de monitorização de fraude na rede [4].

As soluções escaláveis providenciam funcionalidades em *real-time* que processam todo o tipo de transacções incluindo voz, SMS, *Multimedia Message Service* (MMS), *General Packet Radio Service* (GPRS), etc. para assegurar que os casos relevantes de fraude e *high usage* disparam alertas em segundos. Múltiplos motores analíticos providenciam combinações de processamento *real-time*, *fingerprinting* e perfis de utilização para obter capacidade analítica na indústria. As companhias estão a investir em soluções de modo a eliminar o risco inerente em cada uma das fases do subscritor incluindo:

- Durante a fase do processo de pré aquisição, antes do custo de instalação;

- Durante a fase do período de subscrição onde a vulnerabilidade de fraude é maior;
- Ao longo do processo de aquisição, incluindo uma validação da identificação do subscritor;
- No decorrer da relação com o subscritor, com vigilância nas alterações que se podem traduzir em abusos de utilização dos serviços da rede.

A fraude nas telecomunicações pode ser descrita como uma qualquer actividade pela qual é obtido um serviço que não se tem intenções de pagar [5]. Esta definição de fraude pode ser apenas detectada uma vez que já aconteceu. É importante saber distinguir detecção e prevenção de fraude [6]. A prevenção de fraude é impedir que seja cometida fraude. Nenhum método de prevenção é perfeito e é normalmente um compromisso entre eficácia e conveniência de utilização. A detecção de fraude é em contrapartida a identificação de fraude depois de esta ter sido cometida.

As técnicas de fraude evoluem rapidamente e sempre que um método de detecção é conhecido os fraudulentos adaptam as suas estratégias. O desenvolvimento e a troca de métodos de detecção de fraude é limitada no sentido em que não faz sentido descreverem os métodos em detalhe uma vez que fornece aos fraudulentos a informação que necessitam para evitar serem detectados [7].

Para conseguir obter sucesso na detecção e prevenção de fraude, Sistemas de Suporte à Decisão (DSS) têm muito a ver com sistemas de detecção de fraude. O utilizador procura padrões de comportamento, associações e anomalias e toma as decisões da empresa baseadas nestes resultados. Procurar padrões de comportamento permite às operadoras orientarem projectos de *marketing* e vendas com maior adesão. Se os subscritores usam mais serviços de mensagens as operadoras podem concentrar-se nessa área em termos de publicidade.

Procurar associações permite às operadoras obter informação menos óbvia de como segmentar os seus clientes. Procurar nos dados por anomalias é onde as operadoras encontram as surpresas maiores e mais úteis. É também onde os sistemas de detecção de fraude e DSSs têm mais em comum. As mesmas regras aplicam-se à detecção de fraude

bem como a DSSs. O analista ou os algoritmos de Inteligência Artificial (IA) procuram anomalias, ou padrões de comportamento, associações ou sequências de eventos.

Os métodos de detecção e prevenção de fraude podem ser supervisionados ou não [6]. Os métodos supervisionados são aqueles em que são usadas amostras de detalhes de comportamentos de subscritores normais e fraudulentos para construir modelos que permitem aos sistemas atribuir novas observações de cada uma das classes. Estes métodos podem apenas identificar actividades fraudulentas que já ocorreram anteriormente. Os métodos não supervisionados procuram nos dados comportamentos que sejam fora do normal.

As anomalias encontradas por sistemas de detecção de fraude podem incluir exemplos como um número de telemóvel fazer chamadas no estrangeiro pela primeira vez depois de seis anos de subscrição. Este exemplo pode ser uma regra de negócio para prevenção e detecção de fraude. Muitas operadoras já conhecem muitas das regras possíveis mas não usam estas regras de uma forma automatizada em sistemas de detecção de fraude. Se as operadoras questionassem estas transacções de uma forma automatizada obteriam uma redução de fraude.

As ferramentas automatizadas que podem ser usadas para facilitar a detecção de fraude devem incluir implementações de DSSs, *On-Line Analytical Processing* (OLAP), análise estatística e inteligência artificial, tal como modelos heurísticos, redes neuronais, algoritmos de *data mining* e Sistemas de Análise Táctica (TAS).

O primeiro passo passa por criar a infra-estrutura de um sistema de detecção de fraude para codificar estas regras de modo a ser capaz de descobrir potenciais fraudes a tempo de identificar e prevenir potenciais perdas.

O segundo passo é criar um sistema de detecção de fraude capaz de descobrir informação desconhecida nos dados da operadora. Este processo denomina-se de *Knowledge Discovery in Databases* (KDD) ou *Data Mining* (DM). Isto não é o mesmo que dizer que KDD ou DM são detecção de fraude. São ferramentas que podem ser usadas para determinar padrões, sequências de eventos, associações e segmentos de comportamento para usar em sistemas de prevenção e detecção de fraude [8].

2.1. Tipos de Fraude

De acordo com especialistas da indústria, existem pelo menos 200 tipos de fraude nas telecomunicações, e a guerra constante contra estas actividades são batalhas sem fim para as companhias de telecomunicações [2].

As ferramentas de detecção e prevenção de fraude permitem aos operadores lidar com roubos de identidades e fraudes de subscrição de uma forma eficaz. A fraude deve ser detectada e prevenida antes e depois da activação do subscritor na rede. Os roubos de identidades e fraudes de subscrição são as duas maiores ameaças de fraude nas telecomunicações actualmente. Alguns dos exemplos mais comuns de fraude são:

- Manipulação de contas:
- Transferência de crédito/controlo de registos;
- Renúncia de taxas e depósitos;
- Concessão de acesso a serviços Premium;
- Criação de contas com dados fictícios;
- Roubo:
- Roubo de equipamento terminal: *Personal Identification Numbers* (PINs), *vouchers*, palavras-chave e informação dos clientes;
- Manipulação de pagamentos:
- Modificar dados de cobranças;
- Dar créditos ilegais;
- Roubar pagamentos a clientes;
- Perder crédito;
- Ataques técnicos:
- Acesso à rede sem autorização;

- Abuso de dados de autenticação;
- Corrupção de dados na rede.

Seguidamente são detalhados alguns dos tipos de fraude mais usados pelos fraudulentos.

2.1.1. Fraude de Subscrição

A fraude de subscrição prevalece desde o roubo ou manufacturação de identidades, sem existir a necessidade de um fraudulento atacar a encriptação da rede ou sistemas de autenticação. É uma forma de, com menos recurso a tecnologias, conseguir uma melhor hipótese de não ser apanhado. A fraude de subscrição é um dos métodos preferidos por fraudulentos para a fraude de *roaming* digital.

O *modus operandis* de um fraudulento de subscrição é fazer-se passar por uma pessoa ou empresa confiável. O fraudulento consegue assim aceder a qualquer rede, em qualquer lugar (1G, 2G ou 3G). Tipicamente o primeiro passo é usar a fraude de subscrição para ter acesso à rede mãe. Desta forma aparecem na rede como um subscritor aceite pela rede digital e pelo sistema de autenticação.

Em grande parte dos casos os fraudulentos lidam com revendedores corruptos ou grupos internos no fornecedor dos serviços de forma a criarem as contas de subscrição. Obtêm privilégios de *roaming*, por exemplo fazendo-se passar por pequenas empresas ou comportando-se como um pagador regular por um período de tempo (conhecido como "adormecido"). O fraudulento desloca-se para uma rede estrangeira beneficiando do serviço de *roaming* e gera um volume muito elevado de chamadas prolongadas em sucessões rápidas, usualmente em múltiplos dispositivos móveis. Para o reconhecer, as operadoras estão a aprender os padrões de fraude de subscrição, incluindo os indicadores comuns de alterações de morada nos primeiros 15 a 30 dias da abertura de conta. Se um subscritor novo se desvia substancialmente do padrão de comportamento dos novos subscritores e usa serviços numa quantidade excessiva, é lançado um alerta. Certos serviços são mais vulneráveis a fraude de subscrição tal como a revenda de chamadas de longa distância e *roaming* [2].

As técnicas mais comuns para minimizar o impacto e perdas são análises baseadas em gamas de limites, análise com regras de inferência, análise baseada em perfis e redes neuronais [2].

2.1.2. Fraude Interna

A fraude interna representa 8,2% dos incidentes mas gera 40,3% das perdas de receitas que é igual aos seguintes quatro tipos de fraude combinadas: *roaming* (11,4%), pré pagamento (10,8%), subscrição (11,6%) e *premium* (13,2%). A motivação para tal fraude é causada pelas empresas por não acompanharem o trabalho dos empregados, por má gestão de práticas de contabilidade, objectivos irrealistas e empregados insatisfeitos [2].

Os métodos de evitar perdas internas são através de:

- Técnicas de Auditoria Assistidas por Computador (CAATs) permitem aos investigadores obter uma visão geral das operações de negócio, desenvolver compreensões das relações de todos os elementos dos dados, e facilmente conseguir ver detalhes de áreas de interesse específicas.
- Revisão de registos para procurar a existência de transacções duplicadas, transacções em falta e outras anomalias tais como:
- Comparar os endereços dos empregados com os dos revendedores para identificar empregados que possam ser revendedores;
- Procurar números de cheques duplicados para identificar cheques ou facturas desaparecidas;
- Identificar companhias de revendedores que tenham mais que um código de revendedores ou mais do que um endereço de e-mail. Estas listagens podem representar revendedores fantasma;
- Verificar conflitos:
- Procurar relações entre potenciais novas contratações e unidades de negócio para descobrir conflitos de interesse ou actividades ilegais;

- Procurar na informação de pagamentos por fornecedores que não estão listados em qualquer base de dados comercial – é um possível indicador de que não são legítimos; ou
- Operar com endereços que tenham sido associados a fraude no passado.

É necessário implementar métodos eficazes para mitigar os riscos. Acções tais como ter auditores bem treinados, melhorar as práticas de contratações e verificar o *curriculum vitae,* acompanhar e julgar quando são identificados como fraudulentos, não ter qualquer ponto de falha e formações eficazes para empregados que lidam com dinheiro vai minimizar os riscos.

Conduzir investigações internas é também uma solução para minimizar riscos. O objectivo da investigação interna é reunir factos suficientes para levar a uma condenação. Os membros das equipas requerem independência das equipas operacionais e não podem ser escolhidas das equipas de desenvolvimento.

2.1.3. Fraude de Parceiros

Os revendedores (parceiros) podem representar um risco devido aos activos limitados para acções de apoio à recuperação. Os parceiros podem representar falsamente as transacções. Normalmente têm acessos significativos a sistemas *Operational Support Systems* (OSS) e *Business Support Systems* (BSS) ou conhecem como o negócio funciona. Recrutamento temporário cria oportunidades de fraude. Os distribuidores podem ter acesso a códigos de autenticação de activações, podem criar contas de clientes, apresentar volumes de vendas falsos, etc. A necessidade de gerir relações é absolutamente necessária e requer acções apropriadas por parte de várias entidades:

- Operadores e revendedores comerciais:
- Realizar acordos de interconexão a abordar questões de fraude;
- Cooperar com os parceiros na identificação de fraudulentos;
- Usar tecnologias de prevenção de fraude tal como autenticação;

- Fornecedores de conteúdo ou de logística e instalações de BSS:
- Conduzir testes de proficiência periodicamente;
- Determinar contratualmente quem gere o risco;
- Oferecer assistência;
- Ajudar os revendedores com grandes volumes de contas endividadas.

Os fraudulentos fazem-se passar por subscritores de contas de telecomunicações. A fraude de interconexão é também uma ameaça e é necessário retê-la [2]:

- Os contratos de interconexão devem abordar questões de fraude;
- O crime organizado opera além fronteiras:
- Fraude de subscrição de chamadas internacionais;
- Fraude de *roaming* internacional;
- O fraudulento pode ser uma operadora de telecomunicações:
- Realizam as operações a partir de hotéis de telecomunicações partilhados;
- Pagam as contas nos primeiros meses e depois têm volumes de contas *offshore* por pagar consideráveis;
- O Fórum Internacional de Acesso Irregular à Rede (FIINA), compreende especialistas de fraude do histórico das operadoras e grandes revendedores no mundo para trocar informações de fraude, estima que a média de perdas de receitas anuais em fraude das operadoras de telecomunicações é de 6%.

2.1.4. Fraude de Rede Fixa

As redes fixas estão a evoluir e tornam-se cada vez mais atractivas para os fraudulentos. Melhorias tais como migração dos circuitos de multiplexação por divisão de tempo (em que num único canal transmitem-se vários sinais digitais simultaneamente) para sistemas de multiplexação de pacotes (IP), acesso digital de grande velocidade (*Asymmetric Digital*

Subscriber Line (ADSL), cabo, satélite de banda larga), multiplexação por software em vez de hardware, a ênfase de focar em conteúdo em vez de transporte estendendo o acesso ao Sistema de Sinalização por Canal Comum Número 7 (SS7), supervisão de redes para operadoras e parceiros e alterações em planos de numeração permitem aos fraudulentos com novos meios de aumentar a sua presença.

Os tipos de fraude mais comuns que afectam as linhas de telecomunicações fixas são: navegação nas linhas, roubo de identidades, alterações dos serviços ou cobrança de serviços sem autorização ou conhecimento por parte do subscritor, venda de chamadas, *hack* de *Private Branch Exchange* (PBX) e activações (fraude interna). A menos comum é a fraude de pré pagamento e de engenharia social. A fraude de subscrição pode-se definir como sendo quando um fraudulento ganha acesso a um serviço, normalmente usando identidades falsas. Também está relacionada ao contorno das verificações de crédito mediante subscrição. Em alguns casos, uma empresa ou pessoa individual paga pelo serviço no primeiro mês de subscrição antes de aumentar o seu volume de chamadas por um ou dois meses deixando depois de pagar as facturas.

Os ataques físicos às redes são usualmente efectuados através de telefones públicos, exploração de linhas, vandalismo nos terminais, linhas de Rede Digital Integrada de Serviços (RDIS) com ataques através de SS7, etc. A fraude de *Premium Rate Services* (PRS) é cometida através da criação de um PRS "legítimo" de tarifas altas noutro país para o qual são realizadas chamadas a cobrar na operadora de onde as chamadas são originadas. Em alguns casos são organizados escritórios para ligar para estes números e manter a ligação durante a noite.

Premium rate SMS (toques, imagens e pagamentos) e *premium rate Wireless Application Protocol* (WAP), para serviços de informação, são também conhecidos como riscos de fraude.

Os melhores métodos de eliminar este tipo de fraude são recorrer a *Fraud Management Systems* (FMS), organizações de gestão de fraude, auditores internos, verificação de *revenue assurance*, verificações de correlação, monitorização de tráfego e sistemas actualizados de OSS/BSS.

2.1.5. Fraude de Rede Móvel

A segurança das redes móveis é uma preocupação maior do que a das redes fixas. Aceder fisicamente a uma rede é mais fácil do que através da rede fixa. A autenticação e privacidade (através de encriptação) são as maiores preocupações. Tanto o tráfego dos serviços de voz como o da comunicação de computadores sofrem dos mesmos problemas de segurança. À medida que a indústria migra das redes 2G para 3G todo o tráfego de comunicação é efectuado através de pacotes de informação que transportam voz, dados ou vídeos. A utilização de estações e pontos de acesso ilícitos são uma ameaça. A encriptação é considerada de grande importância para as operadoras, equipamentos móveis e cartões SIM. No caso das operadoras usam standards de segurança *Universal Mobile Telecommunication System* (UMTS) baseados em 802.11i para encriptação de segurança nos pontos de acesso como *Secure Socket Layer* (SSL) para o tráfego de internet e IP Sec para os túneis através de *Virtual Private Networks* (VPN).

Os fabricantes de equipamentos e cartões SIM são responsáveis pela autenticação e armazenamento seguro de chaves de autenticação. Todas as transacções de terminais GSM têm sido asseguradas através do uso de algoritmos de chave pública e simetria nos cartões SIM.

A clonagem e a instalação de chips nos equipamentos para gerarem Números de Série Electrónicos (ESN) e de Identificação (MNI) aleatórios comprometem a segurança das redes móveis.

As melhores formas de reduzir a fraude em redes móveis é através da instalação de sistemas de gestão de fraude para alertar, bloquear e analisar tráfego, implementar tecnologias de autenticação e *fingerprinting*, instalar tecnologias de inactivação dos equipamentos em caso de roubo, recorrer a sistemas de encriptação para evitar *sniffing* de ESN e MNI e implementar práticas anti-fraude de subscrição.

2.1.6. Fraude de Pré-Pagos

Existem vários métodos para os fraudulentos explorarem os serviços pré-pagos:

- usar cartões de crédito perdidos ou roubados;
- os engenheiros internos conseguem aceder a sistemas de cobrança;
- usar cartões, números PIN e códigos de carregamento roubados;
- procurar dados de telemóveis legítimos e usá-los noutros telemóveis.

As operadoras podem reduzir estes métodos da seguinte forma:

- fortalecer a segurança da empresa e vigiar atentamente os empregados;
- usar um sistema de activação de cartões nos pontos de venda a efectuar na altura da compra;
- não permitir acesso à rede a cartões roubados e gerar os códigos PIN e de carregamento na venda;
- instalar sistemas de gestão de fraude com funcionalidades de log e alertas.

Os subscritores pré pagos precisam de ter regulamentações de carregamento, tal como datas de expiração para obrigar a efectuar um carregamento dentro de um período específico de tempo. Pode também ser uma forma de evitar a fraude limitando a activação e carregamentos nos pontos de venda.

Existem riscos associados aos métodos de carregamento. Para reduzir o impacto da fraude de pré pagamento as operadoras estão a desenvolver técnicas tal como:

- cobrança em tempo imediato;
- utilização de sistemas de rateamento;
- combinação dos sistemas de pré e pós pagos num só com geração de logs quando existe alguma alteração;
- activação nos sistemas dos pontos de venda;
- migração para sistemas inteligentes.

2.1.7. Fraude de *Roaming*

Os subscritores dos serviços de *roaming* efectuam chamadas noutras operadoras de telecomunicações como visitantes. A facturação é enviada por ficheiros *Transferred Account Procedure* (TAP) para a rede de que o subscritor é assinante.

Esta fraude é similar à fraude de subscrição em que o perpetrador não tem intenção de pagar pelos serviços usados. O fraudulento aproveita-se do atraso de tempo de identificação e notificação à outra operadora para efectuar a tarifação da chamada.

Um dos métodos usados é o roubo de telemóveis a turistas que normalmente só reportam o roubo dias depois quando voltam a casa.

À medida que as operadoras começam a prevenir-se de métodos fraudulentos, o problema migra para as redes parceiras de *roaming* em que tiram partido das falhas dos algoritmos de autenticação usados pelos cartões SIM nas redes móveis GSM. Os fraudulentos conseguem clonar cartões SIM bem como números ESN e MNI.

A associação GSM está a fazer um esforço para minimizar o impacto da fraude de *roaming*. Foi desenvolvido um Sistema de Segurança de Acreditação (SAS) concebido para minimizar as auditorias de segurança efectuadas pelas operadoras nos fornecedores. Criaram documentos de referência de *roaming* para o Grupo de Especialistas do Ramo de *Roaming* Internacional (IREG) para as operadoras poderem trocar entre si. Estes documentos permitem às operadoras partilharem experiências e conhecimentos.

A comunidade de vendedores trabalha constantemente em técnicas para prevenir ou minimizar os impactos da fraude de *roaming* através de:

- *clearinghouses*: providenciam informação tal como detalhes das chamadas, carregamentos e dados de tarifação;
- *High Usage Reports* (HUR): enviam sumários das chamadas dos subscritores que excedem limites e falta de crédito;
- troca de CDRs de *roaming* (RoamEx): fornece detalhes das chamadas e dados específicos de fraude.

2.1.8. Fraude de Serviços de Conteúdo de Valor Acrescentado

Existem muitos riscos associados a conteúdos maiores que os custos das chamadas. Demora tempo para que o departamento financeiro consiga chegar a acordo com os subscritores para receber os pagamentos relativos ao uso de serviços de conteúdo de valor acrescentado. Existem ainda um grande número de parceiros envolvidos simultaneamente durante a mesma sessão. As leis de *copyright* são diferentes regionalmente e de país para país o que tornam a cobrança ainda mais complicada.

Os ladrões de conteúdos estão usualmente associados a redes de troca de ficheiros de música e vídeo. Existem vários métodos de reduzir o risco de fraude através de encriptação de conteúdos e do uso de canais seguros.

2.2. Métodos de Detecção e Prevenção de Fraude

Alguns dos métodos de detecção e prevenção de fraude existentes e usados por FMS são detalhados de seguida.

2.2.1. Sistemas Baseados em Regras

A grande maioria das ferramentas de detecção de fraude são baseadas em regras ou pelo menos têm componentes de detecção baseados em regras. Uma abordagem baseada em regras permite detectar fraude com uma taxa de falsos alarmes baixa [9].

As regras para lançar alarmes são desenhadas manualmente por especialistas de fraude de modo a gerarem alarmes que servem de alerta de possíveis casos de fraude. As regras são geralmente implementadas/processadas por motores de regras que providenciam funcionalidades para escrever regras ou validá-las. Assim que lançam um alerta este deve ser analisado por um analista de fraude.

Uma ferramenta baseada em regras pode facilmente lançar alarmes. Normalmente é combinado com outras estratégias e/ou modelos tais como *profiling* ou redes neuronais para obter melhores resultados.

2.2.2. Análise de Redes Sociais

A análise de redes sociais emergiu como um paradigma chave nas ciências da sociologia, tecnologia e informação. Este paradigma baseia-se no ponto de vista de os atributos de um indivíduo na rede são menos importantes do que as suas relações com outros indivíduos na rede. Explorar a natureza e a força destas ligações pode ajudar a compreender a estrutura e dinâmica de redes sociais e explicar fenómenos reais desde a eficiência organizacional até à propagação de informação e doenças [10].

Uma relação social entre dois amigos, no contexto das telecomunicações, é baseada na duração das chamadas de voz, frequência das chamadas, etc. que são trocadas durante um certo período de tempo.

A análise de uma rede social de subscritores de uma operadora permite ao operador perceber quais as redes sociais do típico fraudulento, ou identificar possíveis fraudulentos que estejam na rede social de um subscritor reconhecido como fraudulento. Uma outra aplicabilidade possível, no contexto de churn (mudança de operadora), será perceber a probabilidade dos restantes subscritores que pertençam à rede social de alguém que mudou de operadora mudarem também.

2.2.3. Profiling

A ideia por trás de fazer *profiling* de um subscritor é acumular o comportamento passado de um subscritor de modo a construir um perfil ou um dicionário de valores esperados do comportamento do subscritor. Este perfil contém sumários numéricos de alguns aspectos do comportamento ou alguns padrões de comportamento.

O comportamento futuro do subscritor pode então ser comparado com este perfil de modo a examinar a sua consistência com o comportamento considerado normal ou algum desvio do seu perfil o que pode indicar algum tipo de actividades fraudulentas.

Um problema importante é que nunca podemos ter a certeza de que foi cometida fraude. Esta análise deve ser tratada como um método que nos providencia um alerta ou uma pontuação suspeita. A análise deve providenciar algum tipo de medidas para mostrar que

uma observação é mais suspeita de fraude do que outras. Deve ser então focada maior atenção em observações com pontuações maiores [7].

2.2.4. *Fingerprinting*

As perdas triviais podem-se tornar num problema significativo. Enquanto que não se pode travar a fraude no negócio das telecomunicações, há formas de reduzir a fraude como por exemplo, através de *fingerprinting*. Funciona na base de que cada indivíduo tem um padrão único de chamadas. Se, por exemplo, um cliente que omitiu o pagamento das suas facturas volta à rede com outro nome, as chamadas efectuadas para a família, amigos e colegas pode ser usada para cruzar com a informação previamente armazenada e etiquetada como fraudulenta. Não é só o subscritor que pode ser uma ameaça de fraude. Existem também internamente causas prováveis de fraude. A questão mais flagrante é a integridade dos empregados [11].

Fingerprinting de chamadas permite a identificação de indivíduos ou comunidades que foram previamente identificados como sendo interesse de investigação e monitorização. Permite a identificação de fraudulentos que podem ter mudado a sua identidade, mas mesmo mudando de identidade os seus hábitos de comunicações não mudam.

É um benefício particular para as operadoras móveis, especialmente da segmentação de pré-pagos. Vão conseguir detectar clientes recorrentes nos novos subscritores de pré-pagamento. Normalmente são identificados como novos subscritores em vez de clientes existentes com novos equipamentos a tomarem partido de ofertas de novos contratos. Ao identificar comportamentos de comunicação, incluindo chamadas tradicionais, SMS e outros serviços, a operadora pode criar perfis para verificar se está a ocorrer alguma actividade fraudulenta.

Neste tipo de fraude os indivíduos subscrevem-se na operadora com identificações falsas. As facturas, possivelmente após um período de normalidade acabam eventualmente por não ser pagas. Após a desconexão tentam novamente registar-se com uma identificação diferente. Uma solução possível para este problema recai sobre a assunção de o fraudulento ter um padrão específico, como uma assinatura, que é definida pelos eventos efectuados.

Estes eventos podem ser na sua maioria chamadas efectuadas, alguns dos números para os quais ligou ou apenas com um certo grau de unicidade que pode permitir identificar o fraudulento. *Fingerprints* (identificação de padrões) são armazenadas numa base de dados interna, feita de sequências de chamadas efectuadas de contas específicas que são cruzadas com tráfego de novas contas.

2.3. *Fraud Management Systems*

As técnicas dos FMS mais comuns são detecção baseada em regras, técnicas de aperfeiçoamento e descoberta de regras, estudo das detalhes dos clientes e dos seus comportamentos, redes neuronais, auditorias, uso de PINs para realizar chamadas, ajustar limites de capacidades de chamadas e serviços de *roaming*.

As tecnologias mais comuns associadas a FMS são *fingerprinting* de frequências rádio, autenticação (chaves simétricas entre os telemóveis e as estações), sistemas digitais com encriptação, sistemas de mediação IP e de facturação para complementarem a aquisição de dados, chaves públicas *Public Key Infrastructure* (PKI), vigilância SS7, *software* anti-vírus e anti-*trojans*, uso de *firewalls* e assinaturas digitais.

De seguida são apresentados alguns dos FMS existentes no mercado usados por operadoras de telecomunicações.

2.3.1. WeDo Consulting

O Fraud:RAID é o novo sistema de gestão de fraude da WeDo Consulting, tendo sido especificamente desenhado para combater a fraude no sector das telecomunicações em todas as suas vertentes actuais, mantendo ao mesmo tempo a escalabilidade e flexibilidade para o combate à fraude nos serviços de nova geração e convergência.

O Fraud:RAID é o primeiro sistema de gestão de fraude especificamente desenhado para ser completamente configurável pelo operador, permitindo assim à equipa de gestão de fraude a configuração dinâmica das técnicas de detecção de fraude do Fraud:RAID de forma a visarem novos casos de fraude à medida e quando estes forem surgindo.

O Fraud:RAID suporta técnicas padrão de correlação, definição de perfis e estatísticas. Estas técnicas são, regra geral, específicas e permitem ajudar as empresas a detectarem e rastrearem potenciais clientes fraudulentos a partir do momento em que estes subscrevem os serviços do operador e durante todo o seu ciclo de vida.

O Fraud:RAID incorpora um GUI baseado em Web com controlo de acesso, *dashboards* e relatórios configuráveis pelo utilizador que permitem conceder às pessoas certas das equipas e de toda a operadora de telecomunicações uma visão do estado e êxito da investigação da equipa. Os controlos de acesso do Fraud:RAID permitem determinar a informação que é apresentada, aos membros da equipa e aos colegas para que todos tenham a informação que necessitam com o nível de detalhe adequado.

O Fraud:RAID permite a que as equipas de analistas de fraude observem e rastreiem todos os aspectos das actividades fraudulentas e análise de forma fácil, numa ferramenta completa de gestão de casos, proporcionando assim total visibilidade bem como um histórico completo das actividades fraudulentas. Esta visão única permite eliminar clientes fraudulentos e ajudam a aconselhar sobre como eliminar as lacunas técnicas de dos processos do negócio que são exploradas para fins fraudulentos que prejudicam as operadoras de telecomunicações.

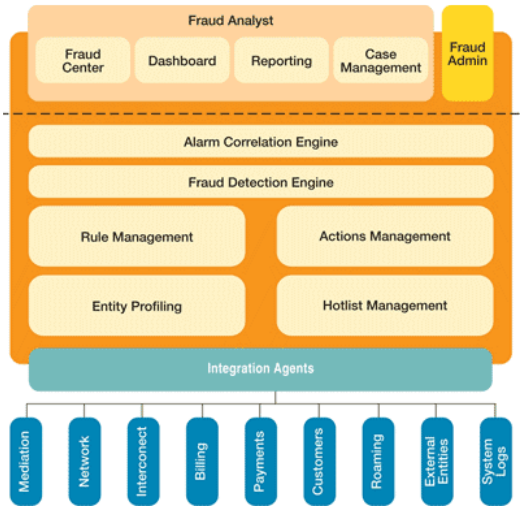

Figura 2 – WeDo Consulting FRAUD:RAID arquitectura

2.3.2. Azure

Azure está a implementar uma tecnologia de *fingerprinting* de chamadas em *real-time* para complementar os motores de detecção de fraude baseado em regras e inteligência artificial permitindo às operadoras identificar padrões de fraude simultaneamente novos como convencionais. O *fingerprinting* de chamadas alimenta directamente o processo de construção de casos de fraude adicionando aos dados que já estão interpretados e integrados para providenciar os operadores com alarmes de fraude.

O director dos serviços de controlo de fraude da Azure disse que os fraudulentos estão continuamente a tentar defraudar as redes. Ao alterar a funcionalidade de *fingerprinting* de chamadas de ser um sistema de análise (*reporting*) para um motor de detecção em *real-time* no sistema de controlo de fraude permitem aos operadores terem a capacidade de detectar fraude. Ao trabalhar em *real-time* o sistema assegura que as operadoras podem agir rapidamente para prevenir fraude e assegurar que não perdem dinheiro.

O *churn* de clientes pré pagos pode também ser reduzido pois as operadoras podem gerir os clientes sem saberem sem saberem quem eles são. Ao monitorizar os eventos, identificar padrões e aplicando *business intelligence* este conhecimento pode ser aplicado como parte do serviço de apoio ao cliente para manter subscritores.

2.3.3. Agilis

A solução da Agilis, denominada de Netmind, é um sistema de *revenue assurance* e gestão de fraude que lida com pontos de risco durante o ciclo de vida de um subscritor para redes terrestres, banda larga e GSM ou *Code Division Multiple Access* (CDMA). A solução de fraude nas telecomunicações é independente da tecnologia e adapta-se rapidamente para evoluir o modelo de negócio num ambiente de *service provider*.

2.3.4. Centaur – PT Inovação

A solução de gestão de fraude Centaur permite um combate eficaz à fraude através da extensão e evolução dos mecanismos dos sistemas de gestão de fraude convencionais, nomeadamente:

- Integração expedita das diversas fontes de informação e incorporação transparente de diferentes técnicas de detecção;
- Flexibilidade dos processos de detecção na adaptação à especificidade de cada cliente ou sistema e aos novos métodos de fraude;
- Processos de prioritização multi-variável orientam os analistas para as situações críticas:
- Disponibilização centralizada de toda a informação necessária ao processo de investigação;
- Valorização do tráfego independente permite a fidedigna replicação da taxação dos distintos serviços e planos comerciais do operador.

2.4. Sumário

A problemática de detecção e prevenção de fraude nas telecomunicações foi apresentada e discutida apresentando a sua evolução histórica. Apresentou-se os tipos de fraude mais comuns e a distinção entre detecção e prevenção de fraude. Foram revistas publicações existentes na área de fraude nas telecomunicações, assim como metodologias e ferramentas existentes dando maior foco ao método de prevenção de fraude *fingerprinting* com descoberta de padrões de fraude em registos de chamadas. Uma aproximação que nunca foi efectuada. No próximo capítulo são apresentados modelos de descoberta de padrões através de técnicas de *Data Mining.*

Capítulo 3

Data Mining Aplicado à Detecção e Prevenção de Fraude nas Telecomunicações

Um dos maiores problemas das operadoras hoje em dia é ser capaz não só de armazenar e gerir a quantidade de dados gerados pelas aplicações e pontos de apoio ao cliente, mas também valorizar esses dados. Tal implica usar ferramentas para armazenar os dados, geri-los, avaliá-los, percebê-los e explorá-los para poder reagir da forma mais adequada.

Conhecer os subscritores reveste-se de grande importância para o negócio das operadoras. Há grande quantidade de dados, de diversas origens, relativos a todas as transacções, preferências ou padrões de comportamentos. Estes dados são muito numerosos e muito complexos e se queremos pessoas orientadas ao negócio para serem capazes de lidar com eles é necessário providenciar não só métodos complexos de processamento mas também intuitivos de usar.

O processo de Descoberta de Conhecimento em Bases de Dados (DCBD) segue uma cadeia complexa em que é necessário dominar cada passo desde a aquisição até à aprendizagem e geração de conhecimento (Figura 3):

- Aquisição dos dados: depende dos pontos de contacto com os subscritores, onde a interacção é em qualquer nível: *switches* (CDR*s*), facturação, apoio ao cliente, pontos de venda, etc;
- Armazenamento dos dados: este é sem dúvida um problema técnico, mas crucial. Uma vez que lidamos com milhões de subscritores durante intervalos de tempo desde meses a anos é importante ter meios de armazenamento dos dados;

- Gestão dos dados: onde deve começar o sistema inteligente, permitindo aos utilizadores acederem a todos os dados. As bases de dados de *marketing* são estruturadas num *datamart*, usando uma representação conveniente para pedidos automáticos;
- Processamento de dados: lidar com funções matemáticas para relacionar dados, descobrir padrões, para descobrir tendências ou prevê-las. Entramos no campo de análise dos dados e estatísticas, mas também de visualização e *reporting* o que é muito importante para o processo de extracção de conhecimento;
- Análise dos dados: extrair conhecimento dos dados faz parte de *Data Mining*. É um processo complexo que usa os resultados obtidos nos passos anteriores, que por vezes precisam de um reprocessamento, para entender os dados, para descobrir conhecimento ainda não adquirido;
- Aprendizagem: o último passo do processo é o mais importante e tem de recordar o que foi útil dos dados/informação/conhecimento sobre o subscritor de modo a alimentar análises futuras e no processo de interpretação.

Figura 3 – Passos que compõem o processo de DCBD [1]

A informação necessária para fornecer conhecimento sobre um subscritor está contida nos dados que temos das suas acções e do seu histórico. Uma vez que não é possível usar todos os dados disponíveis é necessário filtrá-los dependendo do objectivo da aplicação. É muito importante olhar para dados comportamentais assim como demográficos e informação financeira sobre os subscritores. O processo de análise de perfis (*profiling*) é muito complexo pois sejam quais forem os dados sob análise podemos sempre encontrar padrões.

A análise de perfis e comportamentos dos subscritores, para detecção e prevenção de fraude, normalmente recai sobre as mesmas origens de dados presentes no mundo das telecomunicações.

3.1. Aquisição de Dados

As operadoras recebem as actividades dos subscritores através de CDRs produzidos por *switches*. Os *switches* são sistemas de componentes electrónicos que conectam chamadas telefónicas. Independentemente do seu tamanho, os PBX e *Property Management Systems* (PMS) produzem CDRs. Geralmente estes são criados no fim de uma chamada mas em alguns dos sistemas dos telefones os dados são enviados durante a chamada. Estes dados são enviados do sistema do telefone por uma ligação série conhecida por estação de gravação de detalhes de mensagens *Station Message Detail Recording Port* (SMDR). Alguns dos detalhes incluídos nos registos das chamadas são: data e hora, duração da chamada, número para o qual foi efectuada a chamada, identificação do originador da chamada, localização da linha usada, custo e estado da chamada.

Os CDRs são enviados pelos *switches* para as operadoras de telecomunicações com detalhes das chamadas, o mínimo de detalhe existente num CDR é:

- O número que efectua a chamada: *A number*;
- O número que recebe a chamada: *B number*;
- O número a que a chamada é cobrada: *C number*;
- Quando a chamada começou;
- Quanto tempo durou a chamada;
- Tipo de chamada (Voz, SMS, etc..).

Pode também ter incluído alguns dados não usados para propósitos de facturação tais como:

- O identificador do *switch* que registou a chamada;
- Um número sequencial a identificar o registo no *switch*;

- Dígitos adicionais ao *B number* para encaminhar ou cobrar correctamente a chamada;
- O resultado da chamada (se a chamada foi atendida ou se estava interrompido)
- Por onde a chamada foi reencaminhada para o *switch*;
- Alguma falha encontrada na chamada;
- Algum serviço usado pela chamada (chamada em espera ou desvio da chamada).

Os registos das chamadas, tanto locais como de longa distância, podem ser usados para verificação, facturação, gestão da rede e para monitorizar o uso do telefone para determinar o volume de chamadas assim como abusos do sistema de telefone da empresa. Os CDRs permitem gerir custos de chamadas de longa distância e ainda ajudar no planeamento de necessidades futuras de telecomunicações.

Como controlar os custos de telecomunicações com a análise de CDRs:

- Rever todos os CDRs para precisão;
- Verificar custos e utilizações;
- Resolver discrepâncias entre vendedores;
- Desligar serviços que não são usados;
- Terminar contratos de equipamentos não usados;
- Impedir ou detectar fraude em serviços de longa distância;
- Negociar a melhor relação eficácia-custo para o serviço de *roaming*.

As origens de dados mais usadas são também os sistemas de facturação (*billing*) e de apoio ao cliente:

- *Billing*: é neste sistema que está armazenada a informação de utilização tal como informação pessoal dos subscritores. Dependendo da implementação do sistema esta informação pode ser muito precisa ou muito genérica.

- Apoio ao cliente: Os dados armazenados são também informação do cliente, interacções com o operador, e também alguns dados históricos de curto termo.

Existem várias formas de usar os dados e de providenciar resultados operacionais para análise de fraude dependendo dos utilizadores. As pessoas de *Back Office* precisam de soluções que rapidamente lhes forneça relatórios de análise de dados e informação, enquanto que no *Front Office* estão mais preocupados com a interacção com o cliente.

3.2. Armazenamento, Gestão e Processamento de Dados

O processo de DCBD inicia-se com a compreensão do estudo do domínio da aplicação e percepção dos objectivos finais a serem atingidos. De seguida é feito um agrupamento organizado de um volume de dados que vai ser alvo da prospecção. Passa-se então à fase de limpeza dos dados (*data cleaning*), através de um pré-processamento dos dados, visando adequá-los aos algoritmos que posteriormente serão utilizados. Esta etapa inclui a integração de dados heterogéneos, eliminação de dados incompletos e problemas de consistência. O pré-processamento pode consumir até cerca de 80% do tempo necessário para todo o processo devido, sobretudo às conhecidas dificuldades de integração de Bases de Dados (BD) heterogéneas [1].

No processo de DCDB, existem alguns problemas como: representação do conhecimento extraído; complexidade da pesquisa; controlo da operação de descoberta; selecção do objectivo de *Data Mining* mais apropriado; e a escolha dos métodos e técnicas adequadas. Essas decisões dependem essencialmente dos objectivos do negócio, da BD utilizada, do domínio e da aplicação do conhecimento descoberto.

Além destes problemas existem alguns desafios referentes à BD e ao próprio sistema a ser implementado [1]:

- Volume da BD: as BD com muitas tabelas que ocupam muito espaço de armazenamento e possuam grande número de registos podem originar uma enorme variedade de padrões, combinações e hipóteses. A solução passa pela utilização de algoritmos que enumerem todas as regras de associação, ou soluções incluindo a amostragem, os métodos de aproximação e o processamento paralelo;

- Alta dimensionalidade da BD: a alta dimensionalidade é medida pelo número de campos (variáveis e atributos) de uma BD, o que aumenta de forma exponencial o espaço de procura e também as probabilidades do algoritmo encontrar padrões falsos. Uma solução possível é a utilização de pesos para identificar variáveis irrelevantes;
- Dados inconsistentes: atributos com valores nulos na BD e atributos importantes para o processo podem não estar presentes na BD;
- Ruído na BD: este tipo de problema é muito comum, indica quais os atributos importantes que podem estar a ser omitidos ou podem conter valores errados, caracterizando o ruído. A solução é utilizar métodos estatísticos para identificar variáveis ocultas e as suas dependências ou utilizar amostras muito grandes dos dados, tornando o ruído menos significativo;
- Dados irregulares: diferentes BD podem ser usadas e consequentemente, os dados operacionais podem ter diferentes domínios para definir uma mesma informação e variar em termos de qualidade. A solução para este problema é uma análise efectiva de qual a melhor BD para seleccionar os dados, ou, construir o *Data Warehouse* (DW) que apresenta um ambiente estável e integrado dos dados;
- Dados constantemente alterados: a natureza dinâmica dos dados faz com que eles sejam constantemente alterados, o que pode levar a conclusões imponderadas e erradas, pois as variáveis medidas podem ter sido removidas ou modificadas. A solução é a utilização de métodos para actualizar os padrões que sofreram mudanças;
- Interacção com o utilizador: os sistemas devem ser autónomos e extrair apenas hipóteses úteis. Por outro lado, os sistemas devem ser configurados para a aplicação e para as BD de cada utilizador de acordo com as suas necessidades e conhecimento que ele possui;
- Conhecimento prévio: muitos métodos e ferramentas de DCBD não são interactivos e não podem incorporar o conhecimento prévio acerca do problema de modo simples. A utilização do conhecimento, de probabilidades retiradas dos dados

anteriormente e de BD dedutivas tornam-se importantes em todas as etapas do processo de DCBD;

- Representação da informação: a informação descoberta deve ser clara, compreensível e acessível ao utilizador; caso contrário pode-se interpretar o conhecimento erradamente. Uma solução possível seria a inclusão de representações gráficas, linguagem natural e técnicas de visualização de dados.

Estes aspectos são importantes e devem ser considerados durante o processo de desenvolvimento de um projecto de DCBD uma vez que 80% do tempo refere-se às etapas de preparação dos dados e 20% à etapa de *Data Mining* propriamente dita [12].

3.3. Análise dos Dados e Aprendizagem

Descobrir conhecimento significa extrair, de grandes conjuntos de dados, sem nenhuma formulação prévia de hipóteses, informações genéricas, relevantes e previamente desconhecidas, que podem ser usadas para a tomada de decisões. A principal característica é a extracção não trivial de informações a partir de um conjunto de dados de grande porte. Essas informações são necessariamente implícitas, previamente desconhecidas, válidas e potencialmente úteis.

O processo de aquisição de conhecimento é essencialmente composto por três etapas fundamentais: o pré-processamento, o *Data Mining* e o pós-processamento, sendo cada uma destas etapas constituída por várias subtarefas.

A saída desta fase não é apenas informação (valores e tendências) mas também modelos para previsão ou classificação. O objectivo desta fase é obter conhecimento dos subscritores. As técnicas envolvidas são baseadas em inteligência artificial para modelar os padrões de comportamento dos subscritores. O problema principal destas técnicas é transmitir aos utilizadores o que precisam de saber de modo a obter conhecimento útil.

A aquisição de conhecimento nos seres humanos é feita através de processos de aprendizagem. Do ponto de vista de um sistema, esta pode ser definida como as alterações do sistema, que lhe permitem refazer as mesmas tarefas de uma forma mais eficiente e

eficaz. Do ponto de vista da matemática, a aprendizagem pode ser vista como a percepção de conjuntos de dados [12].

3.4. Modelos e Algoritmos

Por modelo entende-se a representação matemática de um sistema (ou processo) com o objectivo de o estudar, sendo a descrição formal e simplificada de um sistema, ou seja, a estruturação da representação do conhecimento com vista a atingir certo resultado [13]. Os modelos podem ser dinâmicos, quando o representam ao longo do tempo, acompanhando as alterações inerentes ao seu funcionamento.

Segundo Rezende [14] as várias técnicas de Aprendizagem Automática (AA) que podem ser utilizados no contexto do *Data Mining*, agrupam-se em cinco paradigmas:

- Simbólico: Procuram aprender, através de exemplos e contra-exemplos, representações simbólicas de conceitos (e.g. Árvores de decisão e regras);
- Estatístico: Utilização de modelos estatísticos para encontrar uma boa aproximação ao conceito induzido. Entre os métodos estatísticos destacam-se os métodos paramétricos e os modelos de Bayes [15]. Alguns autores têm considerado as redes neuronais como métodos estatísticos paramétricos;
- Baseado em exemplos: Classificam exemplos nunca analisados através de exemplos similares conhecidos. As técnicas mais conhecidas desta classe são o Raciocínio Baseado em Casos (RBC) e uma técnica designada por *Nearest Neighbours* (NN);
- Conexionistas: As redes neuronais correspondem a construções matemáticas simplificadas inspiradas no modelo biológico do sistema nervoso;
- Evolutivo: Este paradigma deriva do modelo biológico associado à evolução. Nesta classe incluem-se os Algoritmos Genéticos (AG) e os sistemas de classificação.

O tipo de AA utilizado em *Data Mining* corresponde à aprendizagem indutiva, i.e. a partir de um conjunto de exemplos (que representam o universo em estudo), por aplicação de um algoritmo (técnica ou indutor), obtém-se um padrão (modelo ou hipótese), que é aplicável a novos casos. Na construção de um modelo, definem-se as principais características do

sistema, que devem ser o mais próximo do real, recolhem-se os dados necessários para a construção do modelo para a posterior validação. É vulgar, utilizando técnicas de amostragem, separar-se os dados em dois conjuntos: um de treino para ajustar ou induzir o modelo e outro de teste para efeitos de validação.

Ao modelo estão associados algoritmos, de forma a identificar padrões e relacionamentos (um modelo é a generalização de um padrão, i.e. a instanciação das variáveis do modelo). Após a construção do modelo, aplica-se o conjunto de dados de validação de forma a validar o modelo criado. Em problemas de classificação, o modelo induzido designa-se por classificador.

A fase de modelação e escolha dos modelos e técnicas é muito importante, porque nem todas as técnicas respondem adequadamente às necessidades, nem vão de encontro aos objectivos que se pretende. A Tabela 1 faz a correspondência entre os objectivos de *Data Mining* e os modelos e as técnicas. Existem dois modelos de *Data Mining*, o de verificação (usada quando se sabe o que pesquisar, e.g., Árvores de Decisão, Indução de Regras, Redes de Bayes) e o de descoberta (quando não existe indicação do objectivo da pesquisa, e.g. RNA, AG, Sistemas de Classificação) [16]. São duas abordagens diferentes, logo, requerem técnicas diferentes e consequentemente, exigem ferramentas diferentes. A descoberta de padrões nos dados para que seja possível prever quais os subscritores com comportamentos fraudulentos enquadra-se no objectivo de *Data Mining* de previsão. Para o objectivo de previsão, os problemas são tratados como pertencentes a uma das seguintes classes:

- Classificação**:** encontrar uma função que faça o mapeamento dos dados em classes pré-definidas (e.g. diagnóstico de uma dada doença a partir de um conjunto de sintomas);
- Regressão**:** encontrar uma função desconhecida cuja saída (ou variável dependente) tem um domínio de valores reais (e.g. previsão do valor de uma acção da bolsa com base em indicadores financeiros).

Objectivos de *Data Mining*	Modelos e técnicas
Segmentação	Árvores de Decisão Redes Neuronais Algoritmos Genéticos Indução de Regras Redes de Bayes
Classificação	Árvores de Decisão Redes Neuronais Indução de Regras Conjuntos Difusos Conjuntos Aproximados Sistemas de Classificação Redes de Bayes Algoritmos Genéticos
Previsão	Árvores de Decisão Redes Neuronais Algoritmos Genéticos Indução de Regras
Associação	Redes Neuronais Indução de Regras Redes de Bayes
Sumarização	Redes de Bayes
Visualização	Árvores de Decisão Redes de Bayes

Tabela 1 – Objectivos de *Data Mining* vs. Modelos e Técnicas [16]

De seguida vão ser estudados os modelos e algoritmos de *Data Mining* de previsão que irão ser usados para análise de melhores resultados de reconhecimento de similaridade de padrões de comportamentos de subscritores fraudulentos.

3.4.1. Raciocínio Baseado em Casos

O Raciocínio Baseado em Casos (RBC) ou Indução de Regras é apresentado como sendo um processo susceptível de utilizar conhecimento duma experiência passada, de um problema ou de um caso similar, para resolver um novo problema. Por outro lado, o RBC dá corpo a uma forma de aprendizagem que se quer incremental e sustentada [17].

O RBC procura a partir da análise do problema encontrar na memória de casos um caso(s) similar(es), usa-o(s) para sugerir uma solução para o problema, avalia a solução proposta e actualiza a memória de casos, aprendendo com esta experiência [17].

Na Figura 4 podemos ver o modelo do raciocínio baseado em casos que inicia o problema num caso, o qual por sua vez é usado na recuperação de um ou mais casos a partir do conjunto de casos passados. Esta recuperação tanto pode ser baseada em similaridades sintácticas entre os casos (abordagem mais comum), como feita a partir de similaridades semânticas; i.e., pela forma ou pelo contexto. Uma solução para o caso proposto surge da combinação do caso recuperado com o caso proposto, através de um processo de adaptação. O processo de revisão testa a solução, aplicando a solução proposta à situação concreta, ou evoluindo com base num controlo exterior, corrigindo-se a solução se necessário (reparação) [17]. Durante o processo de aprendizagem o caso é assimilado para futuras utilizações e a base de conhecimento é actualizada com o novo caso.

Um caso descreve uma experiência de diagnóstico e armazena as características mais significativas e os seus respectivos valores [17].

Para Kolodner e Leake [18] um caso pode ter diferentes formas e tamanhos, associando soluções com problemas, resultados com situações ou vice-versa: "se o que for diferente numa nova situação ensinar algo que não possa ser facilmente inferido do novo caso já gravado, então é útil gravá-lo como um novo caso". Para Watson [19] um caso é feito de dois componentes: a descrição do problema e a descrição da solução.

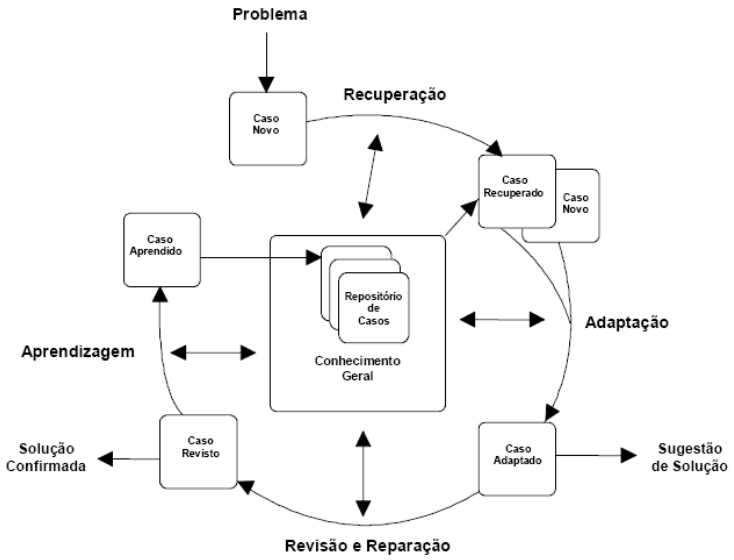

Figura 4 – Modelo do Raciocínio Baseado em Casos [17]

No contexto de prevenção e detecção de fraude um caso contém os atributos mais relevantes dos padrões do comportamento de um subscritor fraudulento. Cada vez que um novo subscritor se regista na rede ao comparar o seu comportamento com os atributos dos casos existentes procuramos similaridades para verificar se é, ou não, um subscritor fraudulento.

A representação dos casos é uma tarefa complexa e importante para o sucesso do sistema RBC. O problema é decidir o que será armazenado em um caso e encontrar a estrutura mais apropriada para descrever seu conteúdo [20]. A escolha adequada depende da consideração de tópicos pertinentes à aquisição e às demais etapas do processo de desenvolvimento, tais como recuperação, adaptação e aprendizagem [21].

3.4.1.1. Recuperação

O objectivo desta etapa é recuperar os casos para que possam auxiliar o raciocínio que se produz nos passos seguintes. A recuperação é feita com as características do novo caso que são relevantes na solução de um problema.

Leake [22] coloca que uma característica importante dos sistemas de RBC é possuir alternativas para identificar os casos de forma a conseguir representá-los e indexá-los, garantindo que os mais úteis são recuperados para resolver o problema. Apenas se consegue

alternativas para identificar os casos através de procedimentos de comparação e medição de similaridades.

As tarefas envolvidas na etapa de recuperação de casos são:

- Avaliação e Métrica da Similaridade;
- Recuperação;
- Selecção.

O estabelecimento de métricas de similaridade num RBC é uma das etapas mais importantes e cruciais para a eficiência da metodologia como um todo. A determinação da medida de similaridade é um importante componente para avaliar a utilidade do caso. Deve-se considerar também, que o grau de utilidade de um caso depende dos propósitos a que ele se destina e quais os atributos que foram relevantes no passado.

As técnicas mais comuns de avaliação, recuperação e selecção de casos são:

- Vizinho Mais Próximo (*Nearest Neighbor Retrieval*): O processo de definição e identificação dos atributos é fundamental para uma recuperação de sucesso pois permite indicar em que região do espaço o problema em questão está inserido. É a técnica mais indicada para problemas com bases de conhecimentos de casos pequenas e com poucos atributos indexados, devido ao volume de cálculos necessários para determinar cada um dos atributos indexados e cada um dos casos. A similaridade entre o caso alvo e um caso na base de conhecimento é determinada para cada atributo. Esta medida deve ser multiplicada por um factor peso. A soma de todos os atributos é calculada e permite estabelecer a medida de similaridade entre os casos da biblioteca e o alvo.

$$similaridade(A,B) = \sum_{i=1}^{N} f(A_i, B_i) \times w_i$$

  Em que:

  - A é o caso alvo

- *B* é o caso fonte
- *N* é o número de atributos em cada caso
- *i* é cada atributo individual variando de 1 a *N*
- *f* é a função de similaridade para o atributo *i* no caso *A* e *B*
- *w* é peso relativo ao atributo *i*

Este cálculo é repetido para cada caso da biblioteca para se obter um *ranking* dos mesmos. As similaridades são usualmente normalizadas num intervalo entre zero e um (zero quando não existe similaridade, um quando a similaridade é total). A grande dificuldade é a determinação dos pesos relativos dos atributos. A limitação desta abordagem é a dificuldade na convergência para a solução adequada e o número de recuperações. Em geral, o tempo de recuperação aumenta linearmente com o número de casos.

- Indução (*Inductive Retrieval*): Estes algoritmos identificam padrões entre os casos e particionam os mesmos em conjuntos (*clusters*). Cada conjunto contém casos que são similares. Um requisito da indução é a definição dos atributos do caso alvo. Casos com descrição de problema similares fazem referência a problemas similares e soluções similares. Na pesquisa indutiva constroem-se árvores de decisão baseadas em dados de problemas passados. Para a construção da árvore a partir dos casos da base de conhecimento, é necessário passar-lhe os atributos que melhor identificam os casos. Encontrado o primeiro atributo é montado o 1º nó da árvore. O passo seguinte é encontrar dois novos atributos que formem os próximos nós e assim por diante. Montada a árvore a partir da base de casos, o próximo passo é percorrer a árvore com o caso em questão. O último nó da árvore contém os casos mais similares.

3.4.1.2. Adaptação

Kolodner [23] diz que pelo facto de nenhum problema passado ser exactamente igual a um problema actual, as soluções passadas terão de ser adaptadas de modo a solucionarem

novos problemas. A adaptação pode ser uma simples substituição de um atributo da solução por outro ou uma complexa e total modificação na estrutura da solução.

Este processo geralmente ocorre fora do sistema RBC e os resultados da adaptação de uma solução podem demorar a aparecer, dependendo do tipo da aplicação.

Vergara [24] afirma no seu trabalho que existem dois tipos de adaptação generalizados descritos na literatura:

- Estrutural: neste processo a adaptação de regras é aplicada directamente à solução armazenada no caso;
- Derivacional: neste processo as regras geradas para a solução original são executadas novamente para gerar uma solução nova. Quando um caso é recuperado, o sistema verifica se as diferenças entre o caso proposto e o caso passado afectam algumas decisões básicas à solução armazenada no caso. A solução armazenada é adaptada pela re-execução das partes do processo da solução original e não a mudando directamente.

Watson [19] conclui que, apesar da adaptação poder ser usada de várias formas e em várias situações, ela não é essencial e muitos sistemas comerciais de RBC não implementam a adaptação. Eles simplesmente recuperam o caso mais similar e disponibilizam a solução.

3.4.1.3. Revisão e Reparação

A etapa de revisão e reparação é necessária se o novo caso alvo não estiver representado da mesma forma que os casos existentes na base de conhecimento, o objectivo é capturar a nova situação e modelá-la para ter a mesma forma.

As diferenças entre a representação do novo caso alvo e os casos da base de conhecimento não se referem apenas à modelação, como também à possibilidade do caso estar incompleto, impreciso ou simplesmente por não incorporar o mesmo conjunto de atributos que possam ter sido inferidos durante a representação do caso.

O processo de revisão e reparação é incremental e pode ser desenvolvido antes ou durante a recuperação, em ciclos que refinam a indexação até que a recuperação de um caso similar

seja eficientemente concluída. Parte do processo ocorre antes da pesquisa, outra durante a avaliação inicial se for pouco acertada ou incompleta para permitir a recuperação de casos úteis, e outra parte ocorre após a pesquisa.

3.4.1.4. Aprendizagem

Após realizada a adaptação e a revisão e reparação a solução do caso seleccionado pode ser reutilizada para resolver o problema de entrada. Um sistema de RBC somente se tornará eficiente quando estiver preparado para aprender a partir das experiências passadas e da correcta indexação dos problemas [23].

Os casos passados conduzem um sistema RBC a tomar decisões e a apreender das suas experiências de três formas:

- Generalização e especialização;
- Pesquisa restringida;
- Avaliação comparativa.

Leake [22] comenta que na medida em que os casos vão sendo utilizados, pode-se colocar alguns atributos que apresentem o resultado da reutilização daquele caso. O processo de aprendizagem num sistema RBC não deve ser dirigido apenas pelo sucesso da aplicação de um caso recuperado num caso de entrada. É importante também aprender com os fracassos, por dois motivos fundamentais: soluções falhas revelam a necessidade do aprendizado e revelam ao sistema o que deve ser aprendido.

3.4.2. Redes Neuronais

As Redes Neuronais (RNs) são uma das mais conhecidas e usadas técnicas em *Data Mining* [25].

A multiplicidade e heterogeneidade dos cenários de fraude requerem o uso de sistemas de detecção inteligentes. Os motores de detecção de fraude têm de ser flexíveis o suficiente para lidar com a diversidade de tipos de fraude. Deve ser suficientemente adaptativa para assimilar novos cenários de fraude, uma vez que os fraudulentos estão sempre a

desenvolver novas formas de cometer fraude uma vez que os ataques previamente cometidos se tornam impraticáveis [26]. A fraude pode ser detectada através de padrões dos registos de comportamentos dos subscritores. A função dos motores de detecção de fraude é reconhecer esses padrões e produzir os alarmes necessários. Alta flexibilidade e adaptabilidade em problemas de reconhecimento de padrões apontam para redes neuronais como uma potencial solução.

As redes neuronais são unidades de sistemas de decisão elementares que podem ser adaptadas através de treino para reconhecer e classificar padrões arbitrários. A interacção de um grande número de unidades elementares torna possível aprender tarefas complexas e arbitrárias.

Existem dois paradigmas de aprendizagem das redes neuronais a aplicar no reconhecimento de padrões [26]:

- Não supervisionada: Na aprendizagem não supervisionada, a rede agrupa padrões de treino similares em *clusters*. Cabe ao utilizador reconhecer a classe ou comportamento associado a cada *cluster*. Quando os padrões são apresentados à rede após o treino podem ser associados ao *cluster* a que se assemelham mais e reconhecidos como pertencendo à classe correspondente desse *cluster*.
- Supervisionada: Na aprendizagem supervisionada os padrões são etiquetados como pertencendo a uma classe. Durante a aprendizagem a rede tenta adaptar as suas unidades de forma a produzir a etiqueta correcta no output de cada padrão de treino. Uma vez que o treino está completo as unidades são congeladas e quando um novo padrão é apresentado é classificado de acordo com o output produzido pela rede.

As Redes Neuronais Artificiais (RNAs) são simples modelos matemáticos desenhados numa tentativa de emular algumas funções humanas. As Redes Neuronais Recorrentes (RNRs) são ideais para a modelação de sistemas dinâmicos com estados escondidos. Os outputs desses processos são tipicamente guardados em forma de séries temporais.

A escolha da arquitectura e do método de aprendizagem das redes neuronais é influenciada pela tarefa de aprendizagem a ser desempenhada pela RNA. No contexto de sistemas de detecção e prevenção de fraude as categorias principais que se aplicam são [27]:

- Reconhecimento de padrões: Formalmente, esta tarefa define-se como o processo pelo qual um sinal/padrão recebido é atribuído a uma de diversas categorias possíveis. Primeiro, é necessário treinar uma rede, onde os padrões, associados à respectiva categoria, são alimentados à rede de forma repetitiva. Mais tarde, um padrão novo é fornecido à rede, que terá de ser capaz de identificar a categoria correcta, de acordo com a informação assimilada. De uma forma geral, o reconhecimento de padrões pode tomar duas formas:

- Na forma mais simples, utiliza-se uma única Rede *Feedforward* Multi-camada (RFMC) com um algoritmo de aprendizagem supervisionado, tendo os nodos intermédios a função de extracção de características; i.e., uma transformação da entrada (x) num ponto intermédio (y) pertence a uma dimensão inferior. Esta redução facilita a tarefa de classificação, descrita como uma transformação do ponto intermédio (y) em uma das classes possíveis num espaço de decisão r-dimensional, para r classes distintas.

- Na segunda forma, a máquina de aprendizagem é composta por duas partes, a primeira para a extracção de características, via uma rede não supervisionada, e a restante para a classificação, via uma rede supervisionada.

- Regressão/Previsão: A ideia é conceber uma RNA capaz de modelar uma função desconhecida $f(.)$ que se aproxime da função $F(.)$ dada por um conjunto de vectores etiquetados; i.e., composta por um par de entrada-saída ($x \rightarrow y$), para que distância euclidiana seja pequena para todas as entradas; i.e.:

$$\forall x, \left|\left|F(x) - f(x)\right|\right| < \rho$$

- onde ρ representa um valor pequeno. A regressão é uma tarefa perfeita para a aprendizagem supervisionada. As RNAs têm-se mostrado eficazes como ferramentas de previsão de variadas formas, desde a ocorrência (ou não ocorrência) de eventos assim como o tempo e nível de intensidade destes.

As Redes *Feedforward* Multi-camada (RFMC), também conhecidas por Redes Perceptrão Multi-camada (MLP) (Figura 5), constituem uma das mais importantes e populares arquitecturas de RNAs, com um vasto leque de aplicabilidade, utilizáveis em problemas de

reconhecimento de padrões e regressão [28]. Os neurónios são compostos por três ou mais camadas que se encontram interligadas e normalmente utiliza a função de activação *Sigmoide*. Esta função é do tipo não linear, e é a característica que permite fazer representações complexas [29]: $\frac{1}{1+e^{-x}}$

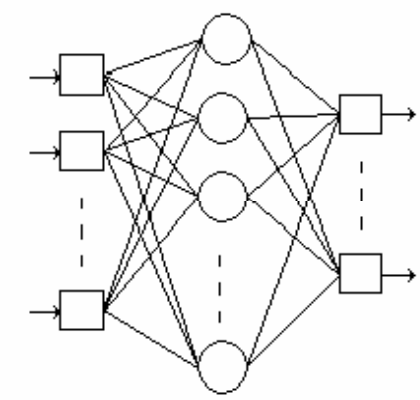

Figura 5 – Rede Perceptrão Multi-camada

A grande vantagem destas redes é a capacidade de abranger as mais variadas classificações de padrões. Por outro lado, os problemas são muitos. O tempo de treino é extremamente longo; à medida que o processo de treino vai decorrendo, os pesos podem atingir valores muito elevados, a soma ponderada torna-se muito grande e a rede não consegue aprender. Para evitar este problema deve-se escolher valores para os pesos dentro de um intervalo pequeno e o número de neurónios nas camadas escondidas deve ser também pequeno [29].

À primeira vista pode ter-se a ilusão de que quantas mais iterações forem executadas melhor será a aprendizagem. Tal pode ser falso porque, nesse caso, poder-se-á cair numa situação em que a rede adapta-se muito bem aos casos de aprendizagem, mas responderá mal a outros casos. Está-se perante uma situação de sobre-ajustamento ou *overfitting*. Portanto, os critérios de paragem têm que conseguir um equilíbrio entre a precisão e a generalização. Nas RNAs, o paradigma de aprendizagem mais comum é o de aprendizagem supervisionada. Assim, o processo de aprendizagem ou treino da rede consiste em ajustar os valores dos pesos das sinapses de modo que as entradas na rede produzam as saídas correctas. Para tal, são utilizados algoritmos de treino dos quais o mais conhecido é o *Backpropagation*.

3.4.2.1. Backpropagation

No algoritmo *Backpropagation*, em cada ciclo de aprendizagem, os erros obtidos pelas diferenças entre as saídas da rede e os valores de treino são propagados para trás, desde os neurónios de saída até aos de entrada, ocorrendo depois um reajuste dos pesos das conexões. O treino termina usualmente quando se obtém o mínimo erro à saída, no caso de regressão, ou o mínimo de classificações erradas. O algoritmo *Backpropagation* é computacionalmente pesado e de convergência lenta pelo que surgiram melhorias a este, tais como o *QuickPropagation* desenvolvido por Fahlman em 1998 [30], ou o *Resilient Backpropagation* (RPROP) proposto por Riedmiller [31].

O algoritmo *Backpropagation*, embora não seja perfeito, permitiu uma forma automática de ajustamento dos pesos das sinapses. Os outros algoritmos mencionados são melhorias, do *Backpropagation*, pelo que a compreensão deste é relevante para a compreensão do processo de aprendizagem. Neste algoritmo é fundamental o cálculo do erro à saída de um neurónio, erro este que, para um neurónio *i*, é dado por:

$$e_i(k) = y_i(k) - d_i(k)$$

em que:

- e_i : erro do neurónio *i*
- y_i : saída do neurónio *i*
- d_i : saída desejada do neurónio *i*
- *k*: entrada em causa

O erro total é $e(k) = \sum_{i=1}^{N} \frac{1}{2} e_i^2(k)$, em que *N* é o número de neurónios.

Após o cálculo do erro, é necessário actualizar os pesos das várias sinapses, recorrendo-se à regra Δ [30]:

$$\Delta w_{ij} = w_{ij}(k+1) + w_{ij}(k) = -\eta \frac{\partial \mathrm{e(k)}}{\partial \mathrm{w_{ij}}}$$

em que:

- w_{ij} : peso da sinapse *ij*;

- η: controlo da aprendizagem (de 0 a 1).

Os pesos são então ajustados de acordo com $w_{ij}(k+1) = w_{ij}(k) - \eta.\mathrm{y(k)}.\mathrm{e_j}(k)$, onde k designa a iteração actual do algoritmo. Isto quer dizer que o novo peso de uma sinapse w_{ij} será resultado da diferença do peso anterior e do produto do factor de controlo da aprendizagem pela saída do neurónio *i* (início da sinapse) e pelo erro do neurónio *j* (fim da sinapse). O efeito do factor de controlo da aprendizagem é o de aumentar ou diminuir o efeito do erro no peso da sinapse, pelo que pode levar a uma convergência mais ou menos rápida, conforme o seu valor for mais próximo de 1 ou de 0. Se a convergência for mais rápida, poderá cair num mínimo local ou nunca convergir devido à variabilidade dos pesos. Uma busca mais lenta poderá levar ao mínimo global mas consumirá bastante mais tempo. A escolha deste valor é de importância fundamental, já que permite o controlo directo da capacidade de generalização da rede. Numa versão que permite um maior controlo do processo, é acrescentado um novo termo, designado por *momentum* [32], onde o cálculo do novo peso é dado por:

$$w_{ij}\,(k\,+1) = \,w_{ij}\,(k) - \,\eta.\,y_i\,(k).\,e_j\,(k) + \mu\Delta w_{ij}(k\,-1)$$

A diferença está na introdução de um termo que corresponde ao ajuste da iteração anterior aferida de um factor μ, chamado de inércia ou *momentum*, que ajuda à convergência para um mínimo global, pois permite que parte do ajuste da iteração anterior vá reflectir-se no ajuste actual.

3.4.3. Árvores de Decisão

Uma árvore de decisão é uma forma de representação de um conjunto de regras que seguem uma hierarquia de classes ou valores desde o nó da raiz até aos nós terminais (folhas) (Figura 6). Cada nó da árvore especifica um teste para os atributos da instância (variáveis), e cada ramo descendente desse nó corresponde a um dos valores possíveis para esse atributo [16].

Os algoritmos de indução de Árvores de Decisão constroem árvores a partir dos dados de treino, de uma forma recursiva, subdividindo este conjunto de dados até que seja formada apenas por nós que representem apenas uma única classe ou a satisfação de um critério

[33]. A construção de uma árvore de decisão é guiada pelo objectivo de diminuir a entropia ou seja a aleatoriedade, dificuldade de previsão, da variável objectivo.

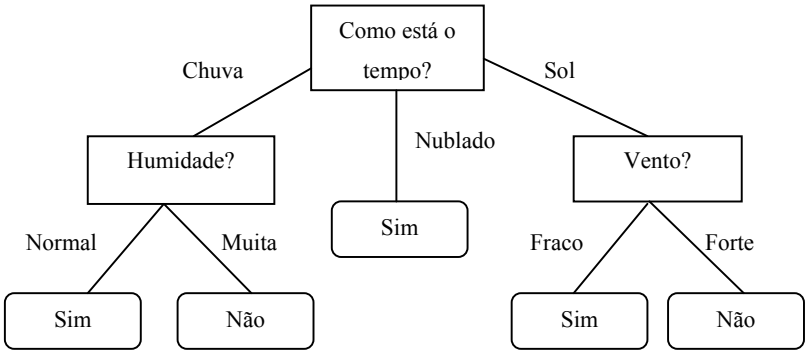

Figura 6 – Árvore de Decisão sobre Jogar Ténis

As árvores geradas têm a seguinte estrutura [34]:

- Folhas: correspondem às classes/objectos;
- Nós internos: correspondem aos atributos; especificam algum teste efectuado num único atributo, com duas ou mais sub árvores que representam saídas possíveis;
- Ramos: correspondem aos valores dos atributos.

Existem dois tipos de árvores de decisão [16]:

- Classificação: qualificam os registos e associam-nos à classe determinada e garantem que essa mesma classificação esteja correcta;
- Regressão: realizam a estimativa do valor de uma determinada variável.

Os algoritmos de indução de Árvores de Decisão/Regressão muitas vezes constroem estruturas com mais ramificações que o necessário, pelo que se torna imperativo "podar" a estrutura. A poda pode ser feita durante a aprendizagem, o que torna o processo mais complexo. Além disso, determinadas podas só podem ser decididas após a construção da árvore, pelo que a maior parte dos algoritmos faz a poda no final da construção da árvore. Após a poda, surge o problema de determinar taxas de erro da árvore. Se houver dados em elevado número, pode-se reservar parte deles para teste após a construção da árvore (técnica de *reduced-error pruning*). Caso os dados sejam escassos, pode-se utilizar um

esquema de validação cruzada (*Cross-Validation*). Neste caso, os dados são divididos em *N* blocos de dimensão semelhante. A aprendizagem faz-se com recurso a *N* iterações, em que a cada iteração são utilizados *N*-1 blocos para aprendizagem e o outro para teste, sendo este diferente a cada iteração.

Em 1963, Morgan e Sonquist [35] propõem um método de criação de árvores através da separação dos dados em grupos sucessivos com base nos seus valores. Este método foi denominado de *Automatic Interaction Detection* (AID). Hartigan [36] em 1975 introduz o teste do chi-quadrado para identificar variáveis independentes ao método AID dando origem ao CHAID. Este novo método permitiu diminuir o crescimento inicial das árvores facilitando a poda. Breiman et al. [37] desenvolveram em 1984 o algoritmo *Classification and Regression Trees* (CART). Este algoritmo cria árvores binárias pelo que, se o teste de uma variável aponta para mais que dois valores, os ramos do nó de teste terão que incluir sub-testes. Tal facto leva a árvores de grande extensão. Ross Quinlan [38] desenvolveu o algoritmo *Iterative Dichtomizer* (ID) que teve várias versões e serviu de base ao C4.5 que, por sua vez, também tem tido várias versões, entre as quais algoritmos para regressão. Tal como o CHAID, estes algoritmos identificam as variáveis independentes, mas recorrendo aos conceitos de "entropia" e "ganho de informação". A entropia está relacionada com a distribuição dos valores de uma variável, ou seja, entropia elevada quer dizer que há uma distribuição mais uniforme, ao invés uma entropia pequena quer dizer que há um valor predominante [36]:

$$Entropia\ (Variavel) = \sum_i [P(Variavel_i \times \log_2(P(Variavel_i)))]$$

O ganho de informação indica a capacidade de uma variável em separar os casos de treino [36]:

$$Ganho\ (Casos, Variavel) = Entropia(Casos) - \sum_i [P(Variavel_i \times Entropia(Variavel_i)]$$

Num sistema de prevenção e detecção de fraude nas telecomunicações as árvores de decisão são treinadas com dados de comportamentos de subscritores fraudulentos agrupando os atributos coincidentes dos subscritores de modo a encontrar os critérios de atributos ideais a aplicar em cada nó. A raiz é o critério presente num maior número de

subscritores e cada folha corresponde a critérios de atributos em menos subscritores fraudulentos. Ao percorrer a árvore com dados de um novo subscritor podemos verificar se o seu comportamento pode ser classificado como sendo um dos subscritores fraudulentos usado no treino da árvore.

As vantagens da utilização das árvores de decisão são as seguintes:

- É um método não-paramétrico: não assume nenhuma distribuição particular para os dados e pode construir modelos para qualquer função desde que o número de exemplos de treino seja suficiente;
- A estrutura da árvore de decisão é independente da escala das variáveis: as transformações monótonas das variáveis (*log x, 2*x, ...*) não alteram a estrutura da árvore;
- Elevado grau de interpretabilidade: uma decisão complexa (prever o valor da classe) é decomposta numa sucessão de decisões elementares;
- É eficiente na construção de modelos;
- Robusto á presença de pontos extremos e atributos redundantes ou irrelevantes pois tem como base um mecanismo de selecção de atributos.

Os inconvenientes da utilização de algoritmos baseados em Árvores de decisão são:

- Instabilidade: Pequenas perturbações do conjunto de treino podem provocar grandes alterações no modelo aprendido;
- Presença de valores desconhecidos;
- Fragmentação de conceitos que podem resultar em replicação de sub-árvores.

3.4.4. Modelos Bayesianos

Não existem regras determinísticas que nos permitam identificar um subscritor como fraudulento. Na melhor das hipóteses é possível formular um grau de confiança no comportamento fraudulento. Modelos gráficos como as redes de Bayes fornecem um

quadro geral para lidar com definições probabilísticas [39] e estão adaptadas a resolver problemas de detecção e prevenção de fraude.

O teorema de Bayes é fundamentado na teoria das probabilidades, e permite representar numericamente o grau de certeza de um dado evento e manipulá-lo de acordo com as regras definidas na teoria da probabilidade [40]. A teoria de Bayes é descrita como uma forma de organizar as regras de probabilidade para actualizar a probabilidade a priori, ou confiança, dada a nova informação, resultando na confiança *a posteriori*. O teorema de Bayes é definido como [41], [42]:

$$P(H|X) = \frac{\mathrm{P(X|H)P(H)}}{P(X)}$$

Em que *P(H|X)* é a probabilidade a posteriori, onde *H* condiciona *X*, i.e., *P(H|X)* reflecte a confiança em *X*, representa a probabilidade do evento *X* condicionado à ocorrência de algum evento *H* (evidência). *P(H)* é a probabilidade a priori de *H* [41].

As redes de Bayes têm base na teoria de Bayes e são uma técnica fundamental no reconhecimento de padrões e classificação[43].

Uma rede de Bayes é construída pela aquisição de conhecimento de um modelo qualitativo do domínio de interesse, representando, portanto, o conhecimento genérico, sendo essa rede composta por duas partes, uma qualitativa e outra quantitativa [40]. A parte qualitativa é representada por um modelo gráfico acíclico, onde as variáveis são os nós e os arcos significam dependências directas entre as variáveis ligadas. Associada à parte qualitativa de uma rede de Bayes está um grupo de funções que representam valores numéricos, os quais compõem a parte quantitativa. Para cada vértice, contendo nós e arcos do grafo, é associada uma função de probabilidade, a qual basicamente é um conjunto de probabilidades condicionais [42], ou seja, é definida por um grafo direccionado, onde os nós representam as variáveis e os arcos representam a dependência condicional ou informativa entre as variáveis. A força da dependência é representada por probabilidades condicionais que são associadas a cada grupo de nós na rede, em que cada variável deve ser independente de todos os nós que não são descendentes excepto, dos pais.

Em suma cada grafo de uma rede Bayesiana codifica uma classe de distribuições probabilísticas. Os nodos desse grafo representam as variáveis do domínio do problema. As setas entre os nodos denotam as relações permitidas entre as variáveis. Estas dependências são quantificadas por distribuições condicionadas para cada nodo dado os seus pais. A principal preocupação, ao utilizar este tipo de estrutura, em forma de rede, é a representação das dependências e independência entre os eventos. Quando se utilizam grafos, somente as variáveis que estão ligadas por arcos é que possuem uma relação de dependência. Isto possibilita a redução dos parâmetros numéricos das condicionadas que fazem parte da distribuição em questão.

As redes Bayesianas podem ser usadas como um sistema especialista. Isto significa que um especialista do domínio do problema desenha um grafo de acordo com os impactos causais assumidos entre as variáveis. As distribuições condicionadas correspondentes podem também ser injectadas pelos especialistas que fazem a tomada de decisão sobre as relações casuais ou são estimadas dos dados usando métodos tradicionais de estimação. Uma vez que uma rede Bayesiana é definida podemos inferir probabilidades de variáveis desconhecidas através da introdução de dados na rede e propagando-os na rede usando regras de propagação [44].

As redes de Bayes podem ser consideradas como diagramas que organizam o conhecimento numa dada área, através de um mapeamento entre causas e efeitos, i.e., pode-se calcular a probabilidade de um evento ocorrer, condicionado pela ocorrência de outro [42]

No contexto de detecção de fraude nas telecomunicações podemos modelar redes Bayesianas para descrever o comportamento dos subscritores de redes móveis.

Para o propósito de modelar o método de *fingerprinting* as redes Bayesianas teriam de ser definidas para cada subscritor fraudulento com os dados do seu comportamento na rede. De modo a perceber se cada novo subscritor tem o mesmo comportamento propaga-se os dados nessas redes e obtemos a probabilidade de ser um subscritor com o mesmo comportamento fraudulento.

3.5. Avaliação de Modelos

O resultado da etapa da utilização de um algoritmo de *Data Mining* é um modelo. É comum utilizar mais do que um algoritmo sobre os mesmos dados, cada um produzindo o respectivo modelo. A ideia é escolher o modelo que melhores resultados obtém. Quer se usem vários algoritmos ou só um, coloca-se sempre a questão da eficácia do modelo. Torna-se, assim, necessária a utilização de métodos de avaliação dos modelos que nos permitam aferir o grau de eficácia dos mesmos.

3.5.1. Matriz de Confusão

Utilizada em classificação, a Matriz de Confusão permite uma visualização inequívoca dos resultados de um modelo [45]. Os resultados são apresentados sob a forma de tabela de duas entradas: uma das entradas é constituída pelas classes desejadas, a outra pelas classes previstas pelo modelo. As células são preenchidas com o número de instâncias que correspondem ao cruzamento das entradas. Na Tabela 2 ilustra-se um exemplo de uma matriz de confusão, em que a entrada vertical são as classificações obtidas pelo modelo, e a entrada horizontal são as classificações originais dos dados. Pode-se ver que no caso da classe B, foram classificados correctamente 46 instâncias, e incorrectamente 4. Já no caso da classe A, todas as instâncias foram correctamente classificadas.

	A	**B**	**C**
A	50	0	0
B	0	46	4
C	0	1	49

Tabela 2 – Matriz de confusão

Através da matriz de confusão é possível obtermos os valores de classificações que se traduzem em positivos verdadeiros, falsos positivos, falsos negativos e negativos verdadeiros, tal como é possível interpretar da Tabela 3.

	A	**B**
A	Positivos Verdadeiros (TP)	Falsos Negativos (FN)
B	Falsos Positivos (FP)	Negativos Verdadeiros (TN)

Tabela 3 – Valores de classificação obtidos da Matriz de Confusão

3.5.2. Curva *Receiver-Operating Characteristic*

A curva *Receiver-Operating Characteristic* (ROC), também conhecida como teoria de detecção de sinais, foi desenvolvida durante a segunda guerra mundial para detectar objectos inimigos em campos de batalha [46].

O resultado dos modelos de classificação em classes são geralmente contínuos, ou seja, produzem um valor situado dentro de um determinado intervalo contínuo, como [0,1] o que implica ser necessário definir um limiar de decisão, para se classificar e contabilizar o número de classificações positivas e negativas. Para cada limiar de decisão são calculados valores de sensibilidade e especificidade, que podem ser dispostos num gráfico denominado curva ROC, que apresenta no eixo das ordenadas os valores de sensibilidade e nas abcissas o complemento da especificidade, ou seja, o valor (1-especificidade). A sensibilidade e a especificidade baseiam-se nas seguintes métricas [47]:

- Sensibilidade = Taxa de Positivos Verdadeiros (TPR) = $\frac{TP}{TP+FN}$
- (1-Especificidade) = Taxa de Falsos Positivos (TFP) = $\frac{FP}{TN+FP}$

Um classificador perfeito corresponde a uma linha horizontal no topo do gráfico, a que é difícil de obter. Na prática, as curvas consideradas boas estão entre a linha diagonal e a linha perfeita, onde quanto maior a distância da linha diagonal, melhor o modelo (Figura 7). A linha diagonal indica uma classificação aleatória, ou seja, um modelo que aleatoriamente selecciona saídas como positivas ou negativas, como atirar uma moeda ao ar e obter cara ou coroa. Definitivamente, não é o tipo de modelo mais confiável possível. No entanto, um modelo em que a curva ROC esteja localizada abaixo da diagonal pode ser convertido num bom modelo – basta inverter as saídas e a curva também é invertida.

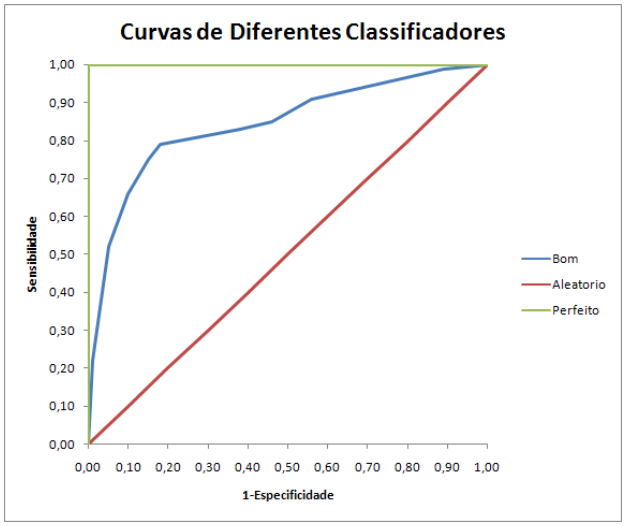

Figura 7 – Exemplos de curvas ROC de diferentes classificadores

Para comparar modelos de classificação é necessário reduzir a curva ROC a um valor escalar. Uma medida padrão para a comparação de modelos é a Área Abaixo da Curva (AUC), que varia entre valores [0,1], em que 0,5 corresponde à área ROC de um modelo de classificação aleatório. Teoricamente, quanto maior a AUC, i.e., quanto mais perto do valor 1, melhor o modelo.

3.5.3. Regressão

Nos modelos de regressão pretende-se escolher aquele que produz valores mais próximos dos dados. A diferença entre o valor real (y) e o previsto ($\hat{y}$) é designada por erro ou resíduo (e_i), e pode-se calcular um erro global, ou seja, de todos os valores previstos, usando as seguintes medidas [48]:

- Mean Absolute Deviation (MAD): $MAD = \frac{1}{2}\sum_i^N |e_i^2|$
- Sum Squared Error (SSE): $SSE = \sum_i^N e_i^2$
- Mean Squared Error (MSE): $MSE = \frac{SSE}{N}$
- Root Mean Squared Error (RMSE): $RMSE = \sqrt{MSE}$
- Root Relative Squared Error (RRSE): $RRSE = \frac{RMSE}{RMSE_{\bar{y}}}$

onde $RMSE_{\bar{y}}$ denota o valor de RMSE calculado para o método simples de prever o valor de *y* com o valor da sua média ($\bar{y}$). De notar que o RRSE é uma medida que é independente da escala dos valores de *y*, sendo que um valor abaixo de 100% significa que o método de previsão avaliado é melhor do que o método simples da média.

3.6. Sumário

O processo de perceber os dados segue uma cadeia complexa desde a aquisição até à aprendizagem e geração de conhecimento. A descoberta de conhecimento em bases de dados concretiza-se num processo com várias etapas ou sub-processos. Fayyad, e al. [1] resume o processo de KDD nas seguintes etapas: Selecção, Pré-processamento, Transformação, *Data Mining*, e Interpretação. A etapa de *Data Mining* distingue-se por ser responsável pela efectiva descoberta de padrões nos dados ou pela verificação de hipóteses.

A fase de modelação e escolha dos modelos e técnicas é muito importante, porque nem todas as técnicas respondem adequadamente às necessidades, nem vão de encontro aos objectivos que se pretende. Foram escolhidos e descritos os modelos e algoritmos com o objectivo de *Data Mining* de previsão:

- Raciocínio Baseado em Casos;
- Redes Neuronais;
- Árvores de Decisão;
- Modelos Bayesianos.

A previsão pode ser conseguida com métodos de classificação ou com métodos de regressão. Estes dois métodos serão abordados no próximo capítulo para realizar a fase de experiências.

Capítulo 4

Caso de Estudo

Neste capítulo é apresentada a solução para geração dos dados necessários, de casos de subscritores fraudulentos e normais, para testar as metodologias descritas no capítulo anterior. A aplicação das metodologias ao conjunto de dados gerados é efectuada através do uso de uma ferramenta que implementa os algoritmos de *Data Mining*. Por fim são apresentadas as experiências a realizar com a ferramenta para avaliar o modelo com mais sucesso de detecção de similaridades nos padrões dos subscritores novos na rede para com os detectados como fraudulentos.

4.1. Geração de Dados

Antes de efectuar a modelação um passo essencial é gerar dados com atributos relevantes, de acordo com padrões de comportamentos de subscritores, organizados em grupos de fraudulentos e não fraudulentos dividindo-os em conjuntos de dados de treino, validação e teste. Esta divisão é necessária pois existem outras necessidades para além do treino dos modelos. O conjunto de validação é necessário para evitar que os modelos fiquem viciados, ou seja decorem os dados e percam capacidades de generalização.

Após o treino e a validação é necessário testar os modelos para medir o desempenho dos mesmos.

Para efectuar a geração de dados foi necessário desenvolver um gerador para simular casos de fraude bem como comportamentos normais dos subscritores. O gerador deve cumprir requisitos como ser capaz de gerar dados em grande volume, ter em conta os atributos dos

registos de chamadas das operadoras de telecomunicações bem como ter em conta a estrutura dos dados da ferramenta que implementa algoritmos de *Data Mining*.

Os comportamentos a simular devem seguir algumas restrições. Nem todos os tipos de fraude têm associado um padrão de actividades que resultam em fraude. Para a modelação é necessário estudar padrões de chamadas de fraudulentos que resultam em fraude e gerar dados de acordo com esses padrões.

A tecnologia usada para o desenvolvimento da ferramenta de geração de CDRs é J2EE.

De modo a simular dados de actividades de subscritores na rede através do gerador é necessário efectuar os seguintes passos:

- Definir um tipo de agente: Onde são definidos os atributos que irão compor os tipos de dados a gerar (A *number*, B *number*, *International Mobile Subscriber Identity* (IMSI), *International Mobile Equipment Identity* (IMEI),…) e o seu tipo de dados (Figura 8);
- Definir agentes: onde são criados e editados os agentes. Um agente contém os valores possíveis para cada atributo definido no tipo de agente e a forma de geração de cada atributo (Figura 9):
    - Valor definido: a geração de dados assume apenas esse valor para o atributo;
    - Escolha aleatória de um valor ou outro: o atributo irá tomar um valor ou outro;
    - Um intervalo de valores sequencial: irá ser gerado um CDR por cada valor do intervalo de valores para este atributo;
    - Um valor aleatório de um intervalo de valores: este atributo terá um valor aleatório no CDR.
- Gerar os dados: permite gerar ficheiros nos formatos *Comma Separated Values* (CSV) e *Attribute Relationship File Format* (ARFF) com os dados definidos para os agentes existentes (Figura 10).

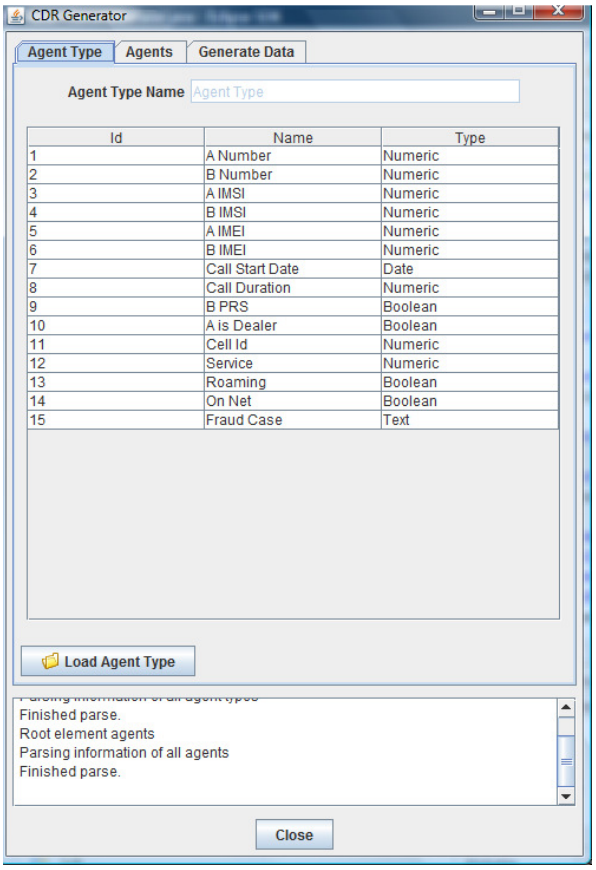

Figura 8 – CDR *Generator* GUI – Definição do tipo de agente

Os atributos usados para os dados, definidos na ferramenta que constituem um tipo de agente são:

- A *Number*: número originador da chamada;
- B *Number*: número para o qual foi efectuada a chamada;
- A IMSI: identificação internacional móvel do subscritor originador da chamada;
- B IMSI: identificação internacional móvel do subscritor destinatário da chamada;
- A IMEI: identificação internacional do equipamento móvel do originador da chamada;
- B IMEI: identificação internacional do equipamento móvel do destinatário da chamada;

- *Call Start Date:* data do início da chamada;
- *Call Duration:* duração da chamada;
- B PRS: indicador se a o destinatário é um serviço *Premium;*
- A *is Dealer*: indicador se o originador da chamada é um revendedor;
- *Cell Id*: identificador da célula de onde foi efectuada a chamada;
- *Service*: identificador do serviço usado na chamada: 1 – Dados, 2 – Voz;
- *Roaming*: indicador se a chamada é efectuada para um número internacional;
- *On Net*: indicador se a chamada é efectuada para um número do mesmo operador;
- *Fraud Case*: identificação do tipo de fraude associado ao registo da chamada.

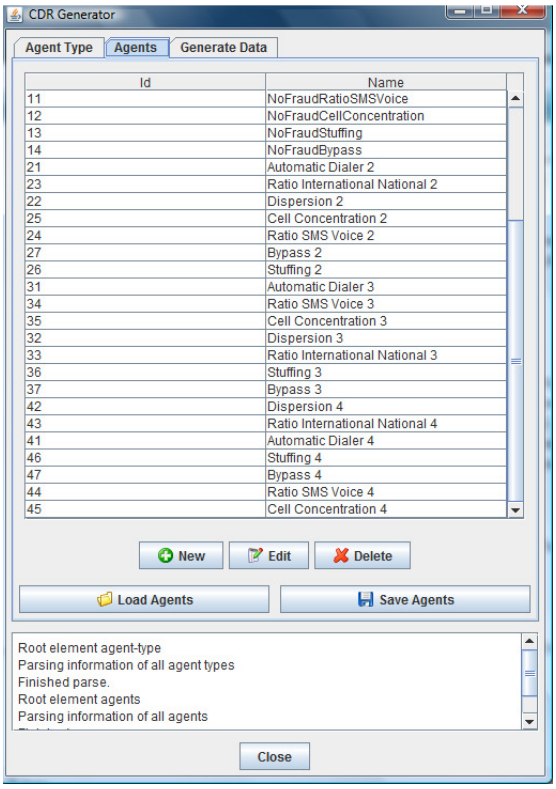

Figura 9 – CDR *Generator* GUI – Definição dos agentes

Após a definição do tipo de agente (Figura 8), atributos e seus tipos de dados, é necessário definir os agentes que vão ser usados para geração de dados (Figura 9). Cada agente

representa um caso de fraude distinto, definindo quais os valores possíveis que cada atributo do tipo de agente pode ter de modo a efectuar a geração dos CDRs com esses valores.

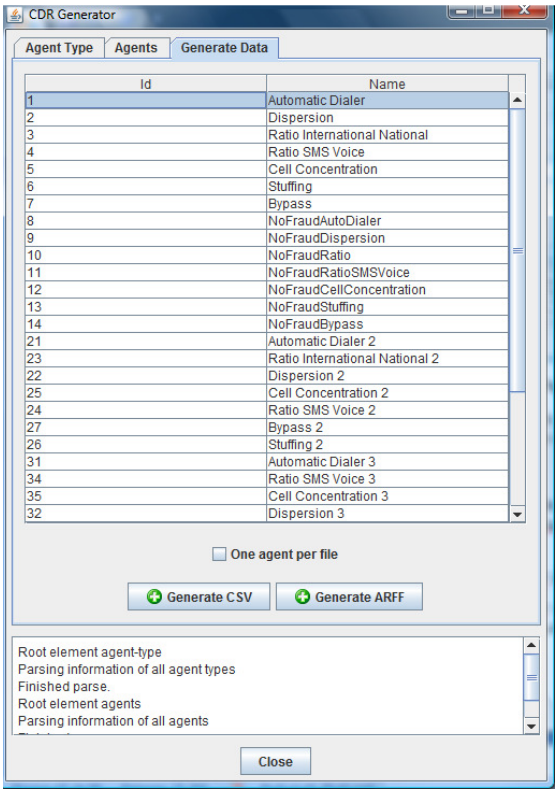

Figura 10 – CDR *Generator* GUI – Gerador de dados

Para a geração dos dados existem algumas considerações a ter em conta:

- Uma vez que os dados são fictícios e uma operadora apenas tem acesso aos registos em que as chamadas são originadas na sua rede os dois primeiros números do campo A *Number* serão o prefixo de uma operadora fictícia, 98.
- O IMSI tem normalmente 15 dígitos. Os primeiros três correspondem ao Código Móvel do País (MCC), o de Portugal é o 268, seguido pelo Código da Rede Móvel (MNC) de dois dígitos na Europa, assumiremos o valor 08. Os dígitos restantes correspondem ao Número de Identificação na Estação Móvel (MSIN). Iremos assumir o valor zero quando não se trata de um subscritor móvel.

- O IMEI é único para cada equipamento móvel e tem 15 dígitos e inclui informação sobre a origem, modelo e número de série do equipamento. Iremos assumir o valor zero quando não se trata de um equipamento móvel.

A geração dos CDRs é efectuada no painel de Geração de dados (Figura 10), para os agentes seleccionados, agrupados por ficheiro ou num único ficheiro e pode ser efectuada para ARFF ou CSV.

Os padrões de fraude, que vão ser usados para definição dos agentes para gerar os conjuntos de dados (*datasets*) de treino dos modelos são descritos nas secções seguintes.

4.1.1. *Automatic Dialer*

Um *Automatic Dialer* é um sistema automático que realiza chamadas sequencialmente, de curta duração. Esta situação ocorre tipicamente em concursos, onde existe um número *premium* para onde as pessoas devem ligar. Por exemplo quando o prémio é atribuído à enésima chamada (e.g. 100ª). Neste processo, as chamadas são recebidas, e aquelas que se encontrem na ordem 1 a 99 são direccionadas para um atendimento automático indicando que não ganharam o prémio. A cada centésima chamada, esta é efectivamente atendida para que a pessoa possa recolher o seu prémio. Em face ao prémio em jogo, pode ser compensatório ter um mecanismo que realize 100 chamadas consecutivas até conseguir ganhar o prémio.

O padrão de fraude associado é reconhecido quando o número originador da chamada é sempre o mesmo, o número destinatário é um número *premium* (ou um conjunto de números) bem identificado para o qual são realizadas chamadas. Um determinado subscritor (fraudulento) realiza chamadas de curta duração (tentativas) e muito pouco espaçadas para o número *premium* até conseguir ganhar o prémio em jogo.

Como técnica de despistagem de detecção deste padrão de fraude um fraudulento mais sofisticado e que tenha noção dos mecanismos mais simples de detecção poderá realizar chamadas para outros números pelo meio da sequência.

4.1.2. Dispersão de Chamadas

O padrão de fraude associado à dispersão de chamadas consiste num IMSI ou IMEI que realiza chamadas para um conjunto de destinos muito diversificado. Está associado a vendas ou desvios de chamadas nos quais o subscritor tira proveito da rede de telecomunicações sem intenção de pagar pelos serviços de que usufrui (fraude de subscrição). Por exemplo, em cada 100 chamadas, existe uma diversidade de 80% destinos.

Existem porém algumas situações muito próximas que não podem ser consideradas fraude, como por exemplo quando as chamadas são efectuadas por vendedores ou equipamentos em lojas de agentes.

4.1.3. Rácios de Chamadas

A fraude associada aos ratios de chamadas é verificada quando a proporção de chamadas de determinado tipo: internacional/nacional, originadas/recebidas, volume excepcional de chamadas de um tipo de serviço, etc.

Existem porém situações de não-fraude com padrão próximo, como por exemplo pessoas que usam o telemóvel apenas para serem contactadas, pessoas que têm pacotes de SMS, e que pertencem a determinada categoria etária (jovens, adolescentes, …) ou mesmo cartões e equipamentos em lojas de agentes para demonstrações.

Nesta situação verifica-se que não só o comportamento tem preponderância, mas existem outros factores que devem ser incorporados (como por exemplo se é um agente).

4.1.4. Concentração de Chamadas numa Célula

A concentração de chamadas efectuadas numa só célula é também um indicador de um padrão de fraude, em que existe um grande volume de chamadas e todas as chamadas efectuadas por um cartão/equipamento são realizadas numa única célula. Por vezes os fraudulentos utilizam mais do que um cartão ou equipamento para realizar esta fraude, atenuando o efeito de volume gerado. Existem porém algumas situações de não-fraude com padrão próximo, como por exemplo pessoas que não saiam muito de casa (por motivos de

saúde ou modo de vida) e utilizam o equipamento para contactar outras pessoas, ou cartões e equipamentos em lojas de agentes para demonstrações.

4.1.5. *Stuffing*

O padrão de *stuffing* corresponde a uma tentativa de despistar sistemas de detecção de fraude que estejam a monitorizar altos consumos por subscritor ou por equipamento em que o fraudulento troca frequentemente o cartão, utilizando vários no mesmo equipamento. A detecção é efectuada através das trocas frequentes de cartão/equipamento (números de pares). Existem porém algumas situações de não-fraude com padrão próximo como por exemplo equipamentos *DualSim,* que suportam dois cartões (para intuitos profissionais e pessoais).

4.1.6. *Bypass*

É denominado de *bypass* a utilização de um equipamento para desvio de chamadas recebidas internacionais, em que a chamada é realizada entre uma origem internacional e um destino local. A parte internacional da chamada é realizada por Voz sobre IP (VoiP), por exemplo, ou por um outro qualquer operador que tenha tarifas mais baixas. A chamada é completada para o destino local encaminhando desde o conector VoiP para uma máquina que contem cartões. Estes cartões fazem uma chamada local e completam o circuito.

Para a operadora, não vai receber o pagamento de interligação de uma chamada internacional terminada na sua rede: apenas cobrará uma chamada local.

O fraudulento pode refinar o processo, tendo cartões de várias operadoras, encaminhando as chamadas para as respectivas operadoras, e neste caso a operadora só cobrará uma chamada dentro da rede.

Se a esta situação juntarmos fraude de subscrição (roubo ou compra de identidade), o fraudulento pode adquirir cartões com planos tarifários nos quais as chamadas dentro da rede não são cobradas.

Esta operação pode envolver várias entidades de vários países (uma rede de fraudulentos). Nomeadamente, uma entidade na origem a cobrar o preço internacional (ainda que possa ser abaixo de custo típico por operadoras), uma entidade (não necessariamente envolvida: *Skype*, *Gtalk*, etc são alguns exemplos; ou operadoras que só servem de intermediárias e não estão implicadas) no meio para transportar a chamada entre a origem e o destino intermédio; a entidade que transporta a chamada entre o destino intermédio e a SIMBOX para realizar chamadas *on-net*.

4.2. Análise de Ferramentas de *Data Mining*

A Descoberta de Conhecimento em Bases de Dados (DCDB), tem como objectivo desenvolver métodos e técnicas de extracção de conhecimento de alto nível a partir de informação guardada em bases de dados [49]. Para atingir este objectivo usa computadores bem como bases de dados e/ou *Data warehouses*. É, portanto, incontornável o desenvolvimento de aplicações de *software* que implementem técnicas de *Data Mining* ou acompanhem mesmo todo o processo de DCDB.

Com este capítulo pretende-se fazer um levantamento e a caracterização de ferramentas de *Data Mining,* com um particular destaque para com as que implementam os modelos analisados. Visto que actualmente existe um enorme número de aplicações nesta área, trata-se de uma tarefa morosa.

Como uma análise exaustiva de todas estas ferramentas se encontra fora do âmbito desta dissertação e haver uma enorme oferta de ferramentas de *Data Mining*, naturalmente tem que se proceder a uma selecção de modo a tornar viável uma comparação de técnicas. Os critérios de escolha da ferramenta passam pelo facto da mesma ter que implementar RNAs, Árvores de Decisão e Redes de Bayes. Por outro lado, a ferramenta deve ter uma interface simples, passível de ser experimentada por utilizadores não especializados.

A ferramenta Weka[1] foi a ferramenta recomendada para a fase de implementação dos modelos pois implementa as técnicas das RNAs, Árvores de Decisão/Regressão podendo estas ser aplicadas tanto à classificação como à regressão e Redes de Bayes. Como interface

[1] Weka na WEB: http://www.cs.waikato.ac.nz/ml/weka/

utiliza um *Graphic User Interface* (GUI), tendo também uma licença não comercial (*freeware*). Criado pela Universidade de Waikato na Nova Zelândia, o Weka foi desenvolvido na linguagem de programação Java (orientada aos objectos) e implementa uma grande variedade de técnicas [47]. Disponibiliza também, diversos algoritmos de reprocessamento de dados, bem como de análise de resultados. O leque de técnicas que implementa permite a utilização da ferramenta em problemas de classificação, regressão e segmentação.

O menu inicial (Figura 11) confronta o utilizador com a escolha de uma interface, de entre quatro possíveis, cada uma com as suas características específicas:

Figura 11 – Weka GUI *Chooser*

- *Explorer:* proporciona um ambiente gráfico de manipulação de dados pela utilização de diversos algoritmos. É a interface mais fácil de utilizar, guiando o utilizador através de menus e formulários, impedindo-o de fazer escolhas não aplicáveis, e apresentando *pop ups* de informação sobre o preenchimento dos vários campos.

- *Experimenter*: também em ambiente gráfico, permite testar técnicas diferentes em classificação ou regressão, de forma a compará-las. Embora tal também seja possível no *Explorer* e no *Knowledge Flow*, no *Experimenter* é possível escolher diversos conjuntos de dados a serem utilizados numa só experiência, várias técnicas a serem experimentadas, o número de repetições (*runs*) do teste, entre outras escolhas. Depois a experiência é executada sem ser necessária a supervisão do utilizador. Os resultados são guardados em ficheiro para posterior análise. É

possível fazer a experiência com computação distribuída através de RMI (*Remote Method Invocation*). Esta é a interface ideal para experiências pelo que será utilizada para levar a cabo as experiências desta dissertação.

- *Knowledge Flow:* permite o desenvolvimento de projectos de *Data Mining* num ambiente gráfico com fluxos de informação (Figura 12). Por outro lado, entre as várias vantagens que possui, é de realçar o *layout* intuitivo, e o facto de permitir o processamento de dados em *batch* ou incrementalmente, o que lhe confere a possibilidade de aplicação a conjuntos de dados de elevada dimensão. Permite, também, o processamento paralelo, em que cada fluxo de dados distinto é processado no seu *thread*.

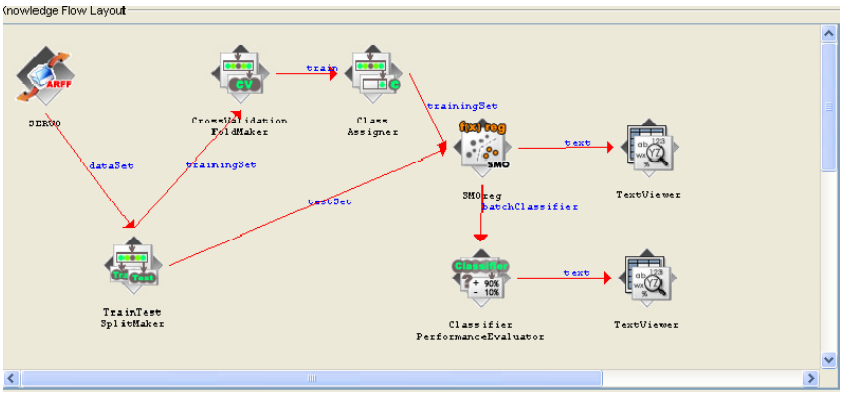

Figura 12 – Exemplo do ambiente *Knowledge Flow* no Weka

- *Simple Command Line Interface* (CLI): proporciona uma interface de linha de comandos onde se podem executar directamente comandos do Weka. Embora disponibilize todas as funcionalidades, requer um elevado grau de conhecimento dos comandos que podem ser utilizados.

A ferramenta Weka, reconhece ficheiros em formato ARFF e CSV, recomendando a utilização de ARFF para um processamento mais rápido dos dados. Para formatar os dados como ARFF, pode utilizar-se um simples editor de texto como o *MS Word*.

Os ficheiros ARFF são compostos por duas secções distintas:

- Cabeçalho: O cabeçalho do ficheiro ARFF contém a declaração da relação e dos atributos.

A declaração da relação é a primeira linha do ficheiro e tem a seguinte formato:

@relation <nome-relação>

As declarações dos atributos definem o nome e o tipo de dados. A ordem em que os atributos são declarados indica a posição da coluna dos dados em cada secção do ficheiro. A declaração dos atributos tem a seguinte sintaxe:

@attribute <nome-atributo> <tipo-dados>

- Os tipos de dados suportados pelo Weka são:
  - *Numeric*: Podem ser números decimais ou inteiros;
  - *String*: Dados de valores em texto;
  - *Date*: Dados de valores que representam datas;
  - *Nominal*: Dados que estão dentro de uma lista de valores possíveis.
- Dados: A secção dos dados inicia com uma linha com o texto @data seguindo-se dos dados separados por vírgula ou *tab*.

Um exemplo de um ficheiro ARFF pode ser visualizado na Figura 13.

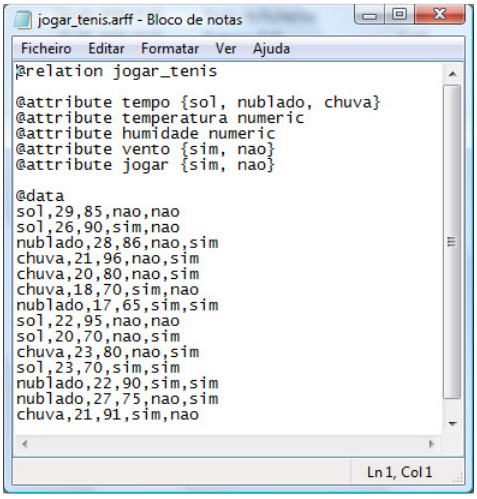

Figura 13 – Exemplo de um ficheiro ARFF

4.3. Experiências

O Weka disponibiliza uma vasta variedade de algoritmos de *Data Mining*, entre os quais algoritmos de RBCs, RNAs, Árvores de Decisão/Regressão e Redes de Bayes.

As experiências realizadas serão de classificação e regressão, para as quais serão usados dois interfaces do Weka: *Experimenter* e *Explorer*.

O *Experimenter* é a interface mais adequada a experiências já que é possível escolher várias tarefas e técnicas a serem testadas numa única experiência de modo a comparar os resultados das técnicas entre si. Depois, a experiência é executada sem ser necessária intervenção do utilizador, tendo posteriormente acesso aos resultados guardados em ficheiro. É ainda possível fazer a experiência com computação distribuída através de *Remote Method Invocation* (RMI), o que pode acelerar consideravelmente as experiências. O *Explorer* permite fazer uma análise mais detalhada dos resultados dos algoritmos de classificação.

A interface *Experimenter* fornece ao utilizador três painéis: *Setup*, *Run* e *Analyse*. A experiência tem início no painel *Setup* que disponibiliza dois modos de configuração: simples (*Simple*) ou avançado (*Advanced*). Ao escolher o modo simples começa-se por configurar uma nova experiência (*New*) definindo o ficheiro de destino dos resultados para posterior análise em *Results Destination*. Em *Experiment Type*, escolhe-se entre *cross-validation* ou *train/test split*, pode-se escolher, também, o método de classificação/regressão.

Na primeira experiência será escolhido o *cross-validation* com o valor por omissão de 10-*fold*, ou seja, são utilizados 10 desdobramentos. O método de validação cruzada reordena aleatoriamente os dados de treino e divide-os em *n folds* de tamanho igual. Em cada iteração, um dos *folds* é usado para teste e *n*-1 são usados para treinar o classificador. Os resultados de teste são coleccionados e calculados para todos os *folds*. Desta forma obtém-se a estimativa da precisão *cross-validation*. *Leave-one-out* (LOO) *cross-validation* significa que *n* é igual ao número de exemplos. Em *Datasets* adicionam-se os conjuntos de dados: *Automatic Dialer*, *Dispersion*, *Ratio International/National, Ratio SMS/Voice*, *Cell concentration*, *Stuffing*, *Bypass* e *No Fraud*. Em *Iteration Control* define-se o número de

vezes que cada técnica será testada, sendo possível mudar a ordem da iteração entre *Datasets first* ou *Algorithms first*. Neste trabalho optou-se por 20 repetições (*runs*) e *Algorithms first*.

Na interface *Explorer* serão usados os painéis de *Preprocess* e de *Classify*. Em *Preprocess* é seleccionado o ficheiro de dados ARFF com a opção *Open File* (Figura 14). Os dados são carregados e podemos passar para o painel de *Classify*, onde são seleccionados e configurados os algoritmos de classificação. Nas opções de teste é seleccionado o *cross-validation* com o valor por omissão de 10-*fold.* O atributo nominal escolhido é o FRAUD_CASE.

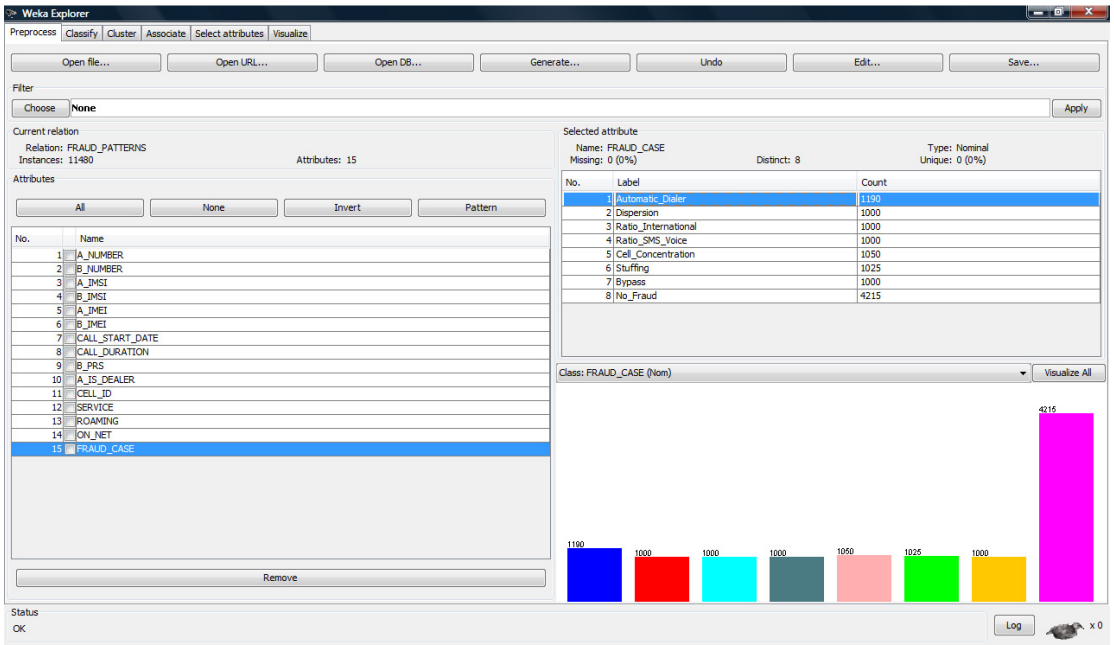

Figura 14 – Weka Explorer – Painel de *Preprocess*

A parametrização dos algoritmos de *Data Mining* (RBCs, RNAs, Árvores de Decisão/Regressão e redes de Bayes) usadas nesta experiência mantém todas as definições (parâmetros do algoritmo) sugeridas inicialmente pelo Weka. O algoritmo de RBC utilizado para realizar as experiências é o de *nearest neighbour* IBk, que procura o valor de distância (*k*) com base na validação cruzada, para encontrar a instância de treino mais próxima da de teste fornecida prevendo assim o valor de classe correspondente à instância de teste. Os valores usados nos parâmetros do algoritmo IBk no Weka são apresentados na Tabela 4.

Parâmetro	Descrição	Valor
KNN	Número de vizinhos a utilizar.	1
crossValidate	Utilização de *leave-one-out cross-validation* para seleccionar o melhor valor de k.	Falso
Debug	Apresentação de informação adicional.	Falso
distanceWeighting	Método de medida de ponderação da distância usado	Sem ponderação da distância
meanSquared	Utilização de mean squared error em vez de mean absolute error durante o *cross-validation* para problemas de regressão	Falso
nearestNeighbourSearchAlgorithm	Algoritmo de procura do vizinho mais próximo.	LinearNNSearch
windowSize	Número de instâncias de treino permitidas. Acima do valor zero implica que instâncias antigas sejam removidas. O valor zero significa que não há limite para o número de instâncias de treino.	0

Tabela 4 – Parametrização do algoritmo IBk

O algoritmo de RNAs utilizado nestas experiências implementa uma rede do tipo *MLP*, chamando-se por isso, *MultiLayerPerceptron*. Pode ser utilizado para classificação ou para regressão. Os valores dos parâmetros usados para correr o algoritmo no Weka são descritos na Tabela 5.

Parâmetro	Descrição	Valor
GUI	Permite visualizar uma interface GUI para configuração da topologia da rede.	Falso
autoBuild	Constrói as camadas intermédias da rede.	Verdadeiro
Debug	Apresentação de informação adicional.	Falso
Decay	Diminui a taxa de aprendizagem: a taxa de aprendizagem de cada iteração é obtida dividindo-a pelo número da iteração.	Falso
hiddenLayers	Define as camadas intermédias.	(Atributos+classes)/2
learningRate	Taxa de aprendizagem.	0,3
Momentum	*Momentum* aplicado aos pesos durante a aprendizagem	0,2
nominalToBinaryFilter	Converte os atributos nominais em binários.	Verdadeiro
normalizeAttributes	Normaliza os atributos numéricos.	Verdadeiro

normalizeNumericClass	Normaliza o atribuído a prever caso seja numérico.	Verdadeiro
randomSeed	Semente (*seed*) para a geração aleatória dos pesos iniciais das sinapses.	0
Reset	Permite que se reinicie a aprendizagem com uma taxa de aprendizagem menor caso o algoritmo esteja a divergir.	Verdadeiro
trainingTime	Número de iterações de treino.	500
validationSetSize	Percentagem dos dados a serem utilizados para validação.	0
validationThreshold	Número de vezes que o erro pode piorar nos dados de validação até terminar o treino.	20

Tabela 5 – Parametrização do algoritmo MLP

No que diz respeito a Árvores de Decisão/Regressão, é utilizado o *J48* para classificação, que mais não é do que uma implementação do famoso algoritmo *C4.5* (criado por J. Quinlan) para o Weka. Para regressão será utilizado o *REPtree*, que constrói Àrvores de Decisão ou de Regressão, fazendo a poda com recurso à técnica de *Reduced Error Pruning (REP)*. Os valores dos parâmetros usados nos algoritmos J48 e *REPtree* podem ser visualizados nas tabelas Tabela 6 e Tabela 7 respectivamente.

Parâmetro	Descrição	Valor
binarySplits	Divisão binária em atributos nominais.	Falso
confidenceFactor	Factor de confiança utilizado na poda.	0,25
Debug	Apresentação de informação adicional.	Falso
minNumObj	Número mínimo de instâncias por folha.	2
numFolds	Determina os dados utilizados para a poda.	3
reducedErrorPruning	Permite optar por *reduced error pruning* ou poda do C.4.5.	Falso
saveInstanceData	Opção para guardar informação dos dados.	Falso
Seed	Semente (*seed*) para gerar aleatoriamente dos índices quando se usa *reduced errror pruning*.	1
subtreeRaising	Permite a *subtree raising* na poda.	Verdadeiro
Unpruned	Define se a poda é realizada.	Falso
useLaplace	Método Laplace na contagem das folhas.	Falso

Tabela 6 – Parametrização do algoritmo J48

Parâmetro	Descrição	Valor
Debug	Apresentação de informação adicional.	Falso
maxDepth	Máxima dimensão da árvore.	Sem limite
minNum	Peso mínimo total por folha.	2
minVarianceProp	Proporção mínima da variância num nó para que seja efectuada a divisão no caso de Árvores de Regressão.	0,001
NoPruning	Define se a poda é realizada.	Falso
numFolds	Determina a quantidade de dados utilizados para a poda.	3
Seed	Semente (*seed*) para gerar aleatoriamente dos índices quando se usa *reduced errror pruning*.	1

Tabela 7 – Parametrização do algoritmo *REPtree*

O algoritmo que implementa redes de Bayes usado nas experiências é o de classificação NaïveBayes em que os valores de precisão do estimador numérico são escolhidos baseados na análise dos dados de treino. Os valores dos parâmetros usados podem ser consultados na Tabela 8.

Parâmetro	Descrição	Valor
Debug	Apresentação de informação adicional.	Falso
displayModelInOldFormat	Utiliza um formato antigo para output. O formato antigo é melhor quando existem muitos valores nominais.	Falso
useKernelEstimator	Utilização de um estimador *kernel* de atributos numéricos em vez de uma distribuição normal.	Falso
useSupervisedDiscretization	Converter atributos numéricos em nominais.	Falso

Tabela 8 – Parametrização do algoritmo NaïveBayes

A configuração do Weka para a experiência de classificação pode ser visualizada na Figura 15.

Na Figura 16 é possível visualizar o GUI do Weka configurado para a experiência de regressão.

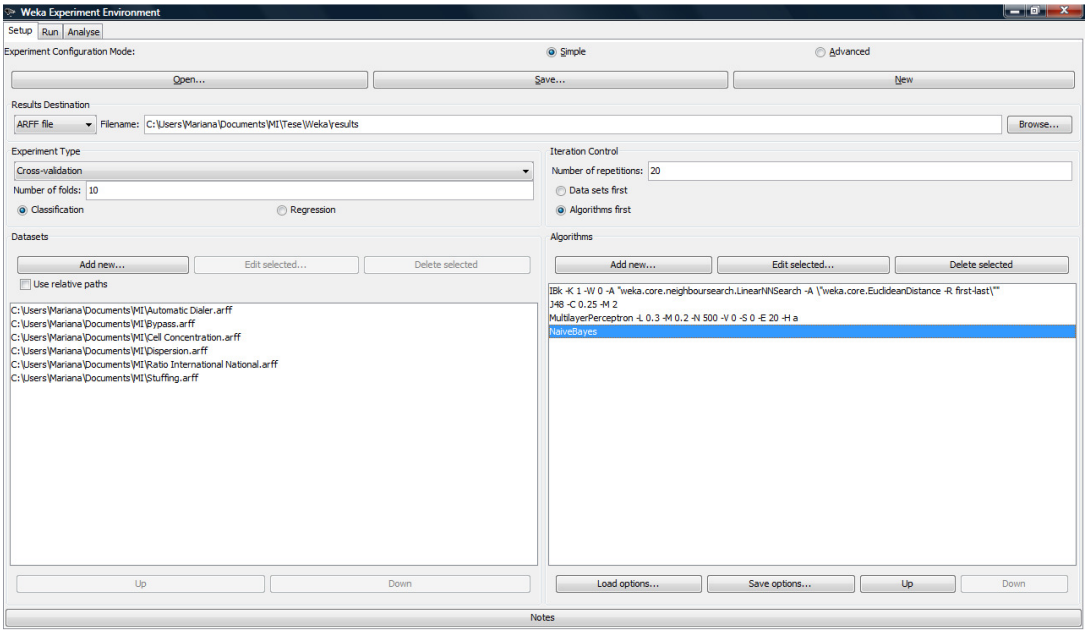

Figura 15 – Weka *Experimenter* GUI configurado para a experiência de classificação

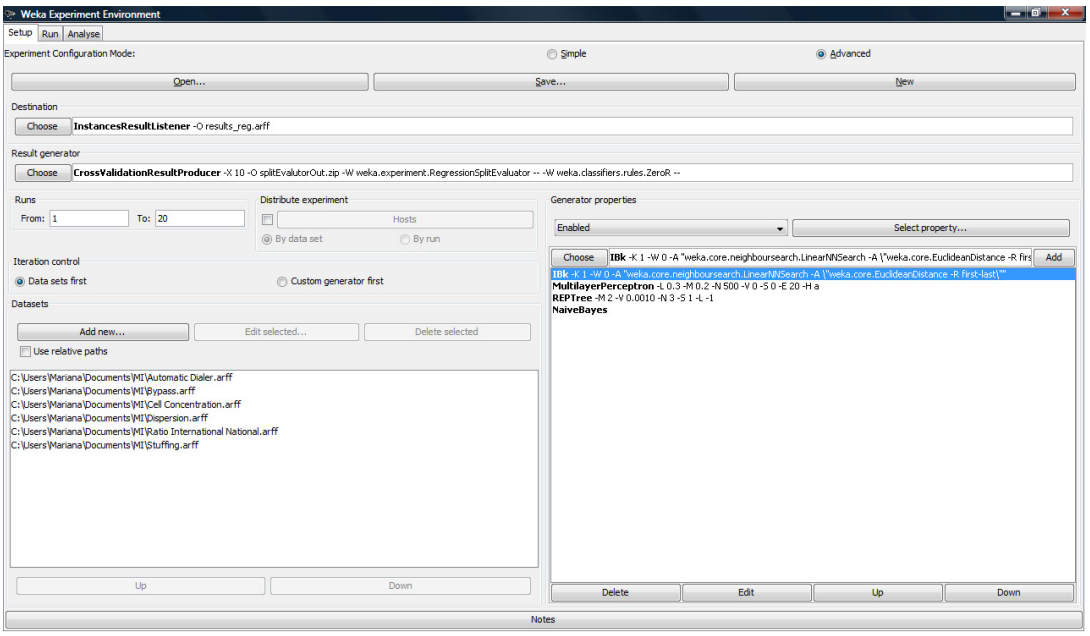

Figura 16 – Weka *Experimenter* GUI configurado para a experiência de regressão

4.4. Sumário

A combinação entre uma boa preparação dos dados e a utilização do algoritmo é essencial para que este processe a informação da melhor forma. Para a geração dos dados foi desenvolvido um gerador de casos de fraude que representam alguns dos padrões de fraude mais comuns.

Para efectuar a modelação o Weka providência um conjunto de ferramentas de *Data Mining*. Fornece uma forma rápida e fácil de explorar e analisar dados. Com os algoritmos implementados pelo Weka foi possível analisar e comparar vários modelos através de métodos de classificação aplicados aos dados gerados pela ferramenta de geração de casos de fraude.

Capítulo 5

Análise de Resultados

Os resultados obtidos nas experiências realizadas no *Experimenter* do Weka são baseados na percentagem de casos de teste que cada modelo classificou correctamente. Podemos assim verificar se existiu alguma diferença estatisticamente significativa na performance de cada modelo através dos resultados de classificação, tais como tempos de execução, percentagens de erro, matrizes de confusão e curvas ROC, bem como os resultados obtidos na experiência de regressão.

Caso de Fraude	**Número de Instâncias**
Automatic Dialer	1190
Dispersion	1000
Ratio International/National	1000
Ratio SMS/Voice	1000
Cell Concentration	1050
Stuffing	1025
No Fraud	4215
Número total de instâncias	11480

Tabela 9 – Número de instâncias por caso de fraude

Podemos ver na Tabela 9 o número de instâncias, para cada caso de fraude, usados no *dataset* de dados gerados para análise dos resultados de classificação e regressão dos modelos.

5.1. Classificação

Para começar a análise carrega-se os resultados da experiência através do botão *Experiment*. O campo de comparação é definido como *Percentage_correct* e o modelo base de comparação é o IBK. Selecciona-se a apresentação dos desvios padrão (*Show Std. Deviations*) e ao carregar em *Perform Test* são apresentados os resultados do teste (Figura 17).

As primeiras cinco linhas correspondem a informação básica sobre o teste. A tabela apresentada (em mais detalhe na Tabela 10) contém a informação principal do teste. As linhas e colunas apresentadas são definidas nas opções *Row* e *Column*. Por omissão cada linha apresentada corresponde aos *datasets* usados e cada coluna corresponde aos resultados de teste de cada algoritmo. Na Figura 17 é apresentada uma tabela 3x5 em que a primeira linha serve para identificar os diferentes esquemas, a primeira coluna mostra o *dataset,* de onde são obtidos os resultados, e o número de instâncias entre parênteses. Uma vez que na experiência é usado um *dataset* e quatro algoritmos, são apresentados quatro resultados de teste são apresentados para um *dataset*. A experiência é realizada com validação cruzada de 10-*fold* (os dados são divididos em 10 partes iguais em que um caso é usado para teste e *n*-1 para treino), o que produziu dez resultados por algoritmo e uma vez que a experiência foi repetida 20 vezes resultou em 200 instâncias, a serem classificadas, por algoritmo.

O primeiro algoritmo é a base do teste, neste caso é o IBK. O *Experimenter* usa uma série de testes emparelhados para comparar os diferentes algoritmos e usa uma notação especial para descrever as comparações. Cada esquema produz três números entre parênteses (#,#,#). Estes números são uma contagem do número de vezes que o esquema foi estatisticamente melhor, igual e pior que o modelo base, respectivamente. Uma vez que o IBK é o modelo base é apresentado com (v//*).

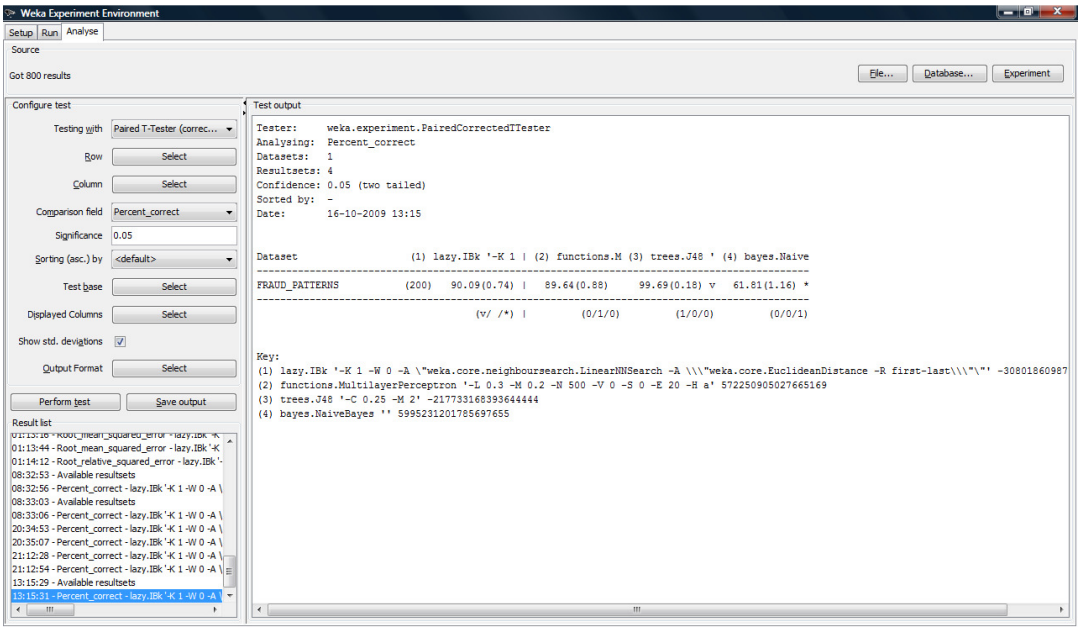

Figura 17 – Resultado da experiência de classificação no Weka Experimenter

Na Tabela 10 são mostrados os resultados das percentagens obtidas assim como o desvio padrão correspondente. O algoritmo com maior número de instâncias classificadas correctamente é o J48 com 99,69% das instâncias classificadas correctamente, com um DP de 0,18%, e o algoritmo com piores resultados é o NaïveBayes com apenas 61,81% e um DP de 1,16%. Estatisticamente o algoritmo MLP foi igual ao algoritmo IBK, pois foi melhor que um algoritmo e pior que outro em termos de classificações correctas.

IBK (DP)	MLP (DP)	J48 (DP)	NaïveBayes (DP)
90,09% (0,74%)	89,64% (0,88%)	99,69% (0,18%)	61,81% (1,16%)
(v/ /*)	(0/1/0)	(1/0/0)	(0/0/1)

Tabela 10 – Percentagem correcta obtida dos casos de teste em classificação

Por vezes é mais conveniente ver os resultados da experiência através de um ranking de acordo com um critério de performance. No *Experimenter*, no painel de análise ao escolher como base de teste em *Select Base* a opção de *ranking*. Os resultados são apresentados na

Tabela 11. As colunas > e < mostram a quantidade de algoritmos que obtiveram melhor e pior, respectivamente, resultados que o algoritmo da linha. Os números da coluna >-< mostram a diferença entre as colunas > e <. Cada linha apresenta os resultados de cada algoritmo. Podemos constatar que o algoritmo J48 foi o melhor do *ranking*, pois classificou melhor do que os restantes 3 algoritmos. O algoritmo NaïveBayes foi o que obteve classificações piores que os outros 3. O MLP e o IBK têm uma posição igual por serem piores e melhores que um algoritmo (NaïveBayes).

Algoritmo	**>**	**<**	**>-<**
J48	3	0	3
MLP	1	1	0
IBK	1	1	0
NaïveBayes	0	3	-3

Tabela 11 – *Ranking* obtido dos casos de teste em classificação

A nível de tempo de modelação o algoritmo MLP demorou significativamente mais tempo que os restantes algoritmos, pois todos os outros foram quase imediatos. A comparação do tempo usado em modelação pelos algoritmos é apresentada na Tabela 12 em segundos.

IBK	**MLP**	**J48**	**NaïveBayes**
0,00s	129,89s	1,42s	0,12s

Tabela 12 – Tempo usado em modelação dos casos de teste em classificação

Podemos constatar através da Tabela 13 que o J48 foi o algoritmo com melhor performance a nível de número de instâncias classificadas correctamente. Os valores de *Root Mean Squared Error* (RMSE) e *Root Relative Squared Error* (RRSE) confirmam que o J48 é o algoritmo com menor erro na classificação das instâncias, com apenas 46 instâncias classificadas incorrectamente, com a melhor média de precisão de classificação e AUC.

	Instâncias Classificadas Correctamente	Instâncias Classificadas Incorrectamente	RMSE	RRSE	Média da Precisão	Média da AUC
IBK	10341 (90,09%)	1139 (9,92%)	0,1574	49,5447%	90%	0,943
MLP	10285 (89,64%)	1195 (10,41%)	0,1452	45,6839%	89,9%	0,974
J48	11434 (99,69%)	46 (0,40%)	0,0299	9,3977%	99,6 %	0,999
NaïveBayes	7107 (61,81%)	4373 (38,09%)	0,282	88,7457%	77,3%	0,94

Tabela 13 – Medidas de comparação dos modelos em classificação

A avaliação de um modelo passa também por entender a precisão da classificação correcta das instâncias por caso de fraude. Na Tabela 14 é apresentada a precisão dos algoritmos para cada caso de fraude. É possível constatar que o algoritmo J48 apesar de ser o algoritmo com maior número de instâncias classificadas correctamente tem menor precisão que o algoritmo NaïveBayes para os casos de fraude *Automatic Dialer* e *Dispersion*. O que se deve ao facto do NaïveBayes ser baseado num paradigma estatístico em que são calculadas as probabilidades de um evento ocorrer condicionadas pela ocorrência de outro e estes tipos de fraude não serem fáceis de mapear em regras condicionais dos dados devido às suas características. Por outro lado, o caso de fraude *Stuffing* é o que tem menor precisão em todos os algoritmos por se caracterizar num caso de fraude em que os eventos são efectuados variando o IMSI no mesmo IMEI ou vice-versa, podendo todos os outros valores dos restantes atributos, em várias instâncias, serem similares aos de qualquer outro caso de fraude.

Podemos concluir que a nível de *ranking*, tempos de modelação, medidas de comparação de resultados e precisão o algoritmo IBK revela-se como sendo o algoritmo com melhores resultados que os restantes, e o algoritmo de NaïveBayes demonstrou ser o que obteve

piores resultados. Nas secções seguintes são apresentados os resultados de classificação com métodos que nos permitem ir ao detalhe dos resultados obtidos por algoritmos por tipo de fraude.

	IBK	**MLP**	**J48**	**NaïveBayes**
Automatic Dialer	99,9%	99,9%	99,9%	100%
Dispersion	98,5%	99,4%	99,6%	99,8%
Ratio International/National	77,8%	73,8%	99,9%	25,5%
Ratio SMS/Voice	83,6%	78,6%	99,4%	73,5%
Cell Concentration	86,7%	80,5%	98,9%	60,1%
Stuffing	62,5%	67,5%	98,8%	51,3%
Bypass	96,3%	98,7%	99,8%	92,9%
No Fraud	95,5%	97%	99,8%	85,7%

Tabela 14 – Precisão dos modelos por caso de fraude em classificação

5.1.1. Matrizes de Confusão

Na interface *Explorer* do Weka é possível analisar os resultados de classificação obtidos por cada um dos algoritmos em detalhe. De seguida são apresentadas as matrizes de confusão, e as métricas TPR, FPR, precisão e AUC para cada um dos algoritmos, em que as letras representam os seguintes casos de fraude:

- A – *Automatic Dialer*;
- B – *Dispersion*;
- C – Ratio *International/National*;
- D – Ratio SMS/*Voice*;

- E – *Cell concentration*;
- F – *Stuffing*;
- G – *Bypass*;
- H – *No Fraud*.

5.1.1.1. IBK

Na Tabela 15 é apresentada a matriz confusão do algoritmo IBK. É possível constatar que o algoritmo IBK classificou apenas para o caso de fraude *Automatic Dialer* (A) todas as instâncias correctamente, enquanto para o caso de fraude *Stuffing* (F) obteve os piores resultados de classificação com apenas 586 instâncias classificadas correctamente de 1025.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	1190	0	0	0	0	0	0	0
B	0	971	3	0	10	4	9	3
C	0	5	793	0	28	119	0	55
D	0	0	0	883	14	85	0	18
E	0	1	21	13	975	28	4	8
F	0	7	141	115	60	586	15	101
G	0	0	0	0	0	1	999	0
H	1	2	61	45	37	115	10	3944

Tabela 15 – Matriz de Confusão do algoritmo IBK

No sumário apresentado na Tabela 16 podemos ver o sucesso de instâncias classificadas correctamente pelo algoritmo IBK para os casos de fraude *Automatic Dialer* (A) com 100% de instâncias verdadeiras e 0% de falsos positivos seguido do *Bypass* (G) com 99,9% de instâncias classificadas correctamente e 0,4% de falsos positivos. Tal como constatamos na

matriz de confusão, podemos confirmar através das percentagens obtidas de instâncias verdadeiras, falsos positivos, precisão e valor da área ROC que o caso de fraude pior classificado é o de *Stuffing* (F).

Tipo de Fraude	TPR	FPR	Precisão	Área ROC
A	100%	0%	99,9%	1
B	97,1%	0,1%	98,5%	0,985
C	79,3%	2,2%	77,8%	0,889
D	88,3%	1,7%	83,6%	0,935
E	92,9%	1,4%	86,7%	0,958
F	57,2%	3,4%	62,5%	0,776
G	99,9%	0,4%	96,3%	0,998
H	93,6%	2,5%	95,5%	0,957

Tabela 16 – Sumário de resultados obtidos pelo algoritmo IBK por caso de fraude

5.1.1.2. MLP

Na matriz apresentada na Tabela 17 podemos ver a matriz de confusão do algoritmo MLP, em que foram classificadas correctamente todas as instâncias dos casos de fraude *Automatic Dialer* (A) e *Bypass* (G).

O caso de fraude *Stuffing* (F) foi o que teve menor número de instâncias classificadas correctamente com apenas 512 de 1025 instâncias.

No sumário apresentado na Tabela 18 podemos constatar que o caso *Ratio International/National* (C), apesar de ter uma percentagem de instâncias classificadas correctamente acima da média, obteve a maior percentagem de resultados falsos positivos.

	A	B	C	D	E	F	G	H
A	1190	0	0	0	0	0	0	0
B	0	940	10	0	32	10	8	0
C	0	0	951	0	4	10	0	35
D	0	0	0	965	0	28	0	7
E	0	5	47	31	863	72	4	28
F	0	0	182	167	114	512	1	49
G	0	0	0	0	0	0	1000	0
H	1	1	99	65	59	126	0	3864

Tabela 17 – Matriz de Confusão do algoritmo MLP

Tipo de Fraude	TPR	FPR	Precisão	Área ROC
A	100%	0%	99,9%	1
B	94%	0,1%	99,4%	0,988
C	95,1%	3,2%	73,8%	0,977
D	96,5%	2,5%	78,6%	0,992
E	82,2%	2%	80,5%	0,933
F	50%	2,4%	67,5%	0,915
G	100%	0,1%	98,7%	1
H	91,7%	1,6%	97%	0,977

Tabela 18 – Sumário de resultados obtidos pelo algoritmo MLP por caso de fraude

5.1.1.3. J48

Na matriz de confusão do algoritmo J48 apresentada na Tabela 19 podemos ver que são classificadas correctamente todas as instâncias de 4 casos de fraude: *Automatic Dialer* (A), *Ratio International/National* (C), *Ratio SMS/Voice* (D) e *Bypass* (G). A classificação com maior número de falsos negativos é o de *Cell Concentration* com apenas 1025 instâncias classificadas correctamente de 1050.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	1190	0	0	0	0	0	0	0
B	0	998	0	0	2	0	0	0
C	0	0	1000	0	0	0	0	0
D	0	0	0	1000	0	0	0	0
E	0	4	1	6	1025	7	2	5
F	0	0	0	0	9	1012	0	4
G	0	0	0	0	0	0	1000	0
H	1	0	0	0	0	5	0	4209

Tabela 19 – Matriz de Confusão do algoritmo J48

O sumário de métricas apresentado na Tabela 20 apresenta os melhores resultados de taxas de verdadeiros positivos e falsos positivos com as melhores precisões de classificação e AUC para todas as instâncias de todos os casos de fraude.

Tipo de Fraude	**TPR**	**FPR**	**Precisão**	**Área ROC**
A	100%	0%	99,9%	1
B	99,8%	0%	99,6%	0,999
C	100%	0%	99,9%	1

D	100%	0,1%	99,4%	1
E	97,6%	0,1%	98,9%	0,994
F	98,7%	0,1%	98,8%	0,998
G	100%	0,1%	99,8%	1
H	99,9%	0,1%	99,8%	0,999

Tabela 20 – Sumário de resultados obtidos pelo algoritmo J48 por caso de fraude

5.1.1.4. NaïveBayes

A matriz de classificação do algoritmo NaïveBayes apresentada na Tabela 21 mostra que o algoritmo apenas classificou correctamente todas as instâncias dos casos de fraude *Automatic Dialer* (A), *Ratio International/National* (C), *Ratio SMS/Voice* (D) e *Bypass* (G). Para todas as instâncias dos restantes casos de fraude obteve um grande número de falsos negativos.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	1190	0	0	0	0	0	0	0
B	0	645	12	3	247	9	68	16
C	0	0	1000	0	0	0	0	0
D	0	0	0	1000	0	0	0	0
E	0	0	87	45	420	330	8	140
F	0	0	202	239	4	536	0	44
G	0	0	0	0	0	0	1000	0
H	0	1	2627	73	28	170	0	1316

Tabela 21 – Matriz de Confusão do algoritmo NaïveBayes

No sumário de métricas da Tabela 22 podemos verificar que apesar de obtermos uma taxa de verdadeiros positivos para os algoritmos *Ratio International/National* (C), *Ratio SMS/Voice* (D) e *Bypass* (G), a taxa de falsos positivos é elevada. Assim como podemos salientar também a baixa taxa de verdadeiros positivos para as restantes instâncias dos outros casos de fraude, sendo o pior caso a classificação de *No Fraud* (H), ou seja onde não existe fraude.

Tipo de Fraude	TPR	FPR	Precisão	Área ROC
A	100%	0%	100%	1
B	64,5%	0%	99,8%	0,992
C	100%	27,9%	25,5%	0,956
D	100%	3,4%	73,5%	0,989
E	40%	2,7%	60,1%	0,927
F	52,3%	4,9%	51,3%	0,909
G	100%	0,7%	92,9%	0,999
H	31,2%	3%	85,7%	0,891

Tabela 22 – Sumário de resultados obtidos pelo algoritmo NaïveBayes por caso de fraude

5.1.2. Curvas ROC

As curvas ROC permitem comparar os algoritmos por classificação de caso de fraude de modo a verificar se existem algoritmos com melhores resultados de classificação para determinados casos de fraude. Um algoritmo com melhores resultados corresponde a uma linha horizontal no topo do gráfico, o que é difícil de obter. Na prática, as curvas consideradas boas estão entre a linha diagonal e a linha perfeita, onde quanto maior a distância da linha diagonal, melhor o modelo. Os gráficos das curvas ROC apresentam no eixo das ordenadas os valores de sensibilidade, a Taxa de Positivos Verdadeiros (TPR), e nas abcissas o complemento da especificidade (1-especificidade), ou seja, a Taxa de Falsos

Positivos (FPR). Quanto menor a especificidade maior a probabilidade de obter classificações de, no contexto de fraude nas telecomunicações, alguém que não está a cometer fraude ser classificado como fraudulento. Quanto menor a sensibilidade menor a probabilidade de classificar um fraudulento correctamente.

Nas seguintes secções são apresentados os gráficos com as curvas ROC dos algoritmos por caso de fraude.

5.1.2.1. *Automatic Dialer*

Na Figura 18 podemos verificar que todos os algoritmos obtiveram a curva ROC ideal de classificação para o caso de fraude *Automatic Dialer*. Nos dados existentes este é o único caso de fraude que realiza chamadas sequenciais para PRSs, o que facilita a sua classificação, apesar de existirem instâncias para os dados de *No Fraud* que realizem também chamadas para PRSs, não existiu nenhuma classificação Falsa Positiva.

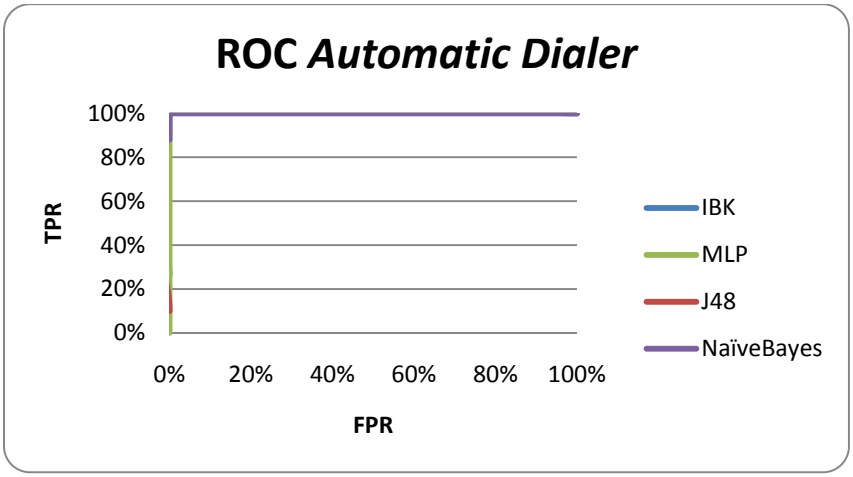

Figura 18 – Curvas ROC para o caso de fraude *Automatic Dialer*

5.1.2.2. *Dispersion*

Na Figura 19 podemos verificar o bom desempenho de todos os algoritmos uma vez que as áreas das respectivas curvas ROC é próxima de 1, apesar da curva do algoritmo J48 mostrar um melhor desempenho da classificação e do NaïveBayes piores resultados, em termos comparativos.

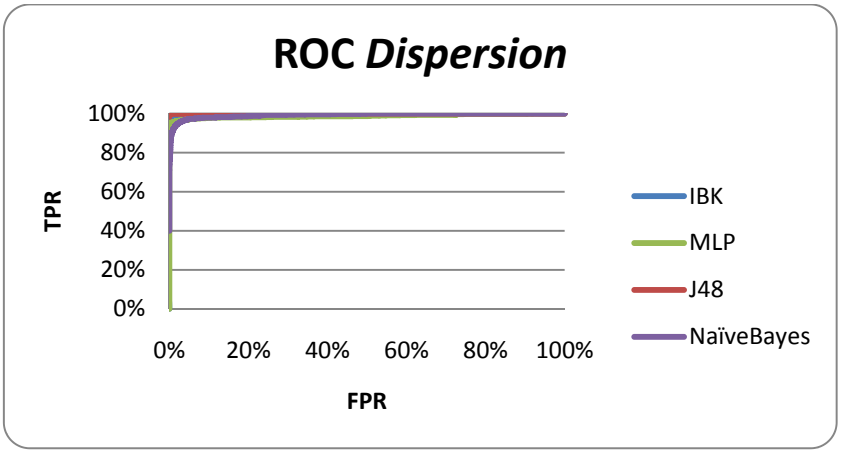

Figura 19 – Curvas ROC para o caso de fraude *Dispersion*

5.1.2.3. *Ratio International/National*

Na Figura 20 é possível visualizar as curvas ROC de classificação dos algoritmos para o caso de fraude *Ratio International/National*. Podemos constatar que o algoritmo J48 obteve a melhor curva ROC de classificação, sendo o algoritmo mais específico e mais sensível, e o algoritmo IBK o que apresenta a curva menos sensível, ou seja, com uma menor taxa de verdadeiros positivos.

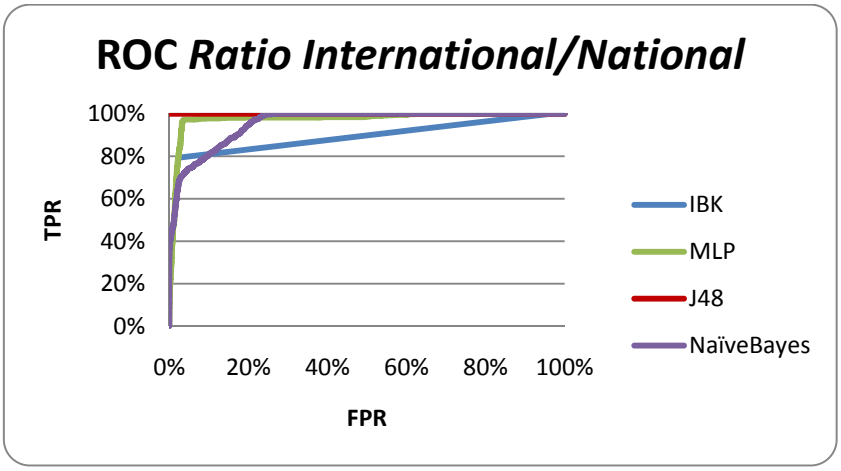

Figura 20 – Curvas ROC para o caso de fraude *Ratio International/National*

5.1.2.4. *Ratio* SMS/*Voice*

Podemos visualizar na Figura 21 uma semelhança entre as curvas ROC do caso de fraude *Ratio International/National* e *Ratio* SMS/*Voice*, em que no segundo o algoritmo NaïveBayes se apresenta mais sensível do que no caso de fraude *Ratio International/National*. Mantendo-se o algoritmo J48 como o mais sensível e mais específico e o IBK o menos sensível.

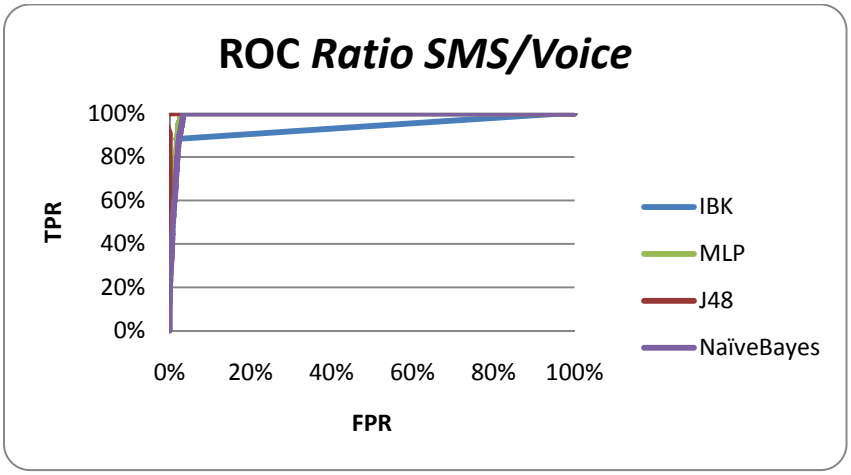

Figura 21 – Curvas ROC para o caso de fraude *Ratio SMS/Voice*

5.1.2.5. *Cell Concentration*

O caso de fraude *Cell Concentration* apresenta todas as curvas ROC acima da linha diagonal, como podemos ver na Figura 22, mas comparativamente com as curvas ROC dos casos de fraude previamente apresentados demonstra ter classificações menos sensíveis e menos específicas por parte de todos os algoritmos. O algoritmo J48 continua a ser o mais sensível e específico de todos.

Os algoritmos MLP e NaïveBayes são os menos sensíveis e menos específicos, conseguindo o algoritmo NaïveBayes ser mais sensível a partir da taxa de falsos positivos de 30%.

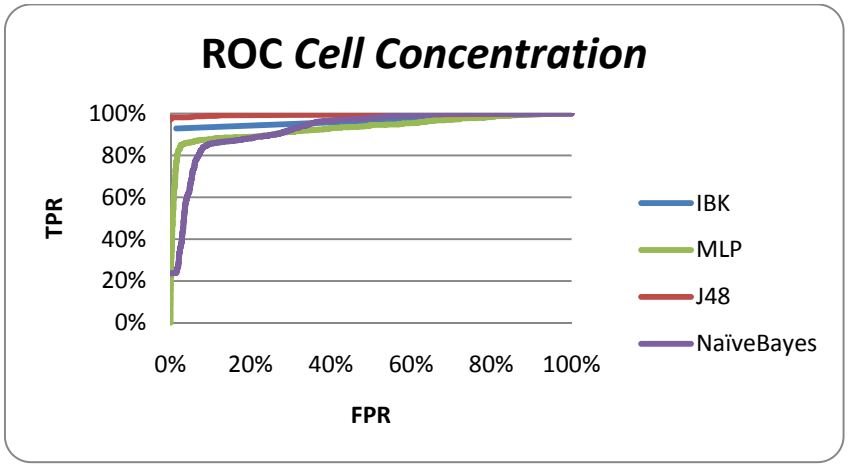

Figura 22 – Curvas ROC para o caso de fraude *Cell Concentration*

5.1.2.6. *Stuffing*

A Figura 23 apresenta as curvas ROC para o caso de fraude *Stuffing* em que o algoritmo J48 é o mais sensível e mais específico e o IBK o menos sensível e menos específico. Os algoritmos MLP e NaïveBayes apresentam uma classificação com taxas de sensibilidade e especificidade similares. Comparativamente com as curvas ROC dos restantes casos de fraude este apresenta, excepto para o algoritmo J48, menor sensibilidade e menor especificidade na classificação por parte dos algoritmos IBK, MLP e NaïveBayes.

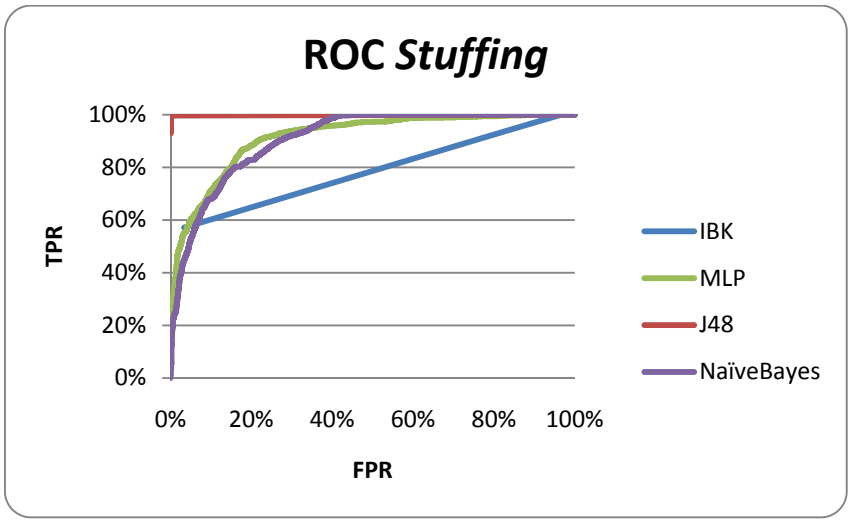

Figura 23 – Curvas ROC para o caso de fraude *Stuffing*

5.1.2.7. *Bypass*

Podemos constatar na Figura 24 que todos os algoritmos obtiveram curvas ROC ideais de classificação para o caso de fraude *Bypass* pelo que apresentam desempenhos idênticos.

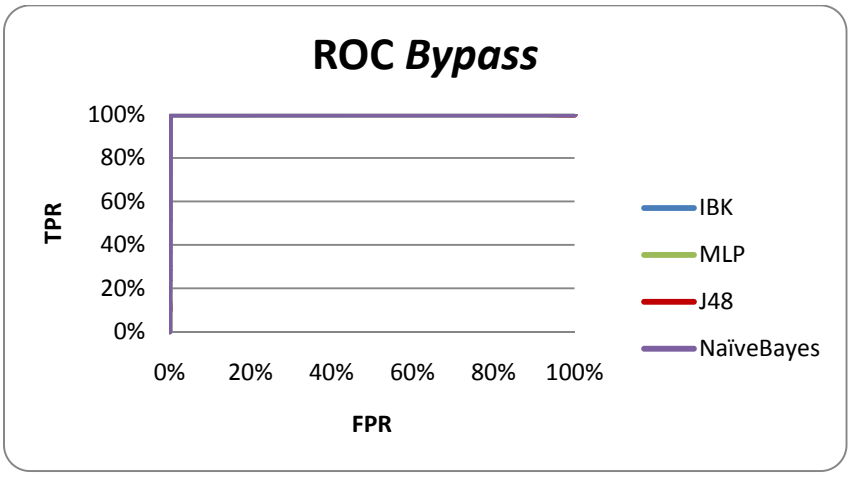

Figura 24 – Curvas ROC para o caso de fraude *Bypass*

5.1.2.8. *No Fraud*

Na Figura 25 é possível visualizar as curvas ROC para as instâncias que representam comportamentos de subscritores normais, com comportamentos similares aos casos de fraude apresentados.

Podemos constatar que o algoritmo NaïveBayes é o menos específico, ou seja, seria o algoritmo que indicaria com maior probabilidade que um subscritor normal é fraudulento. O algoritmo J48 apresenta-se mais uma vez como o mais sensível e mais específico.

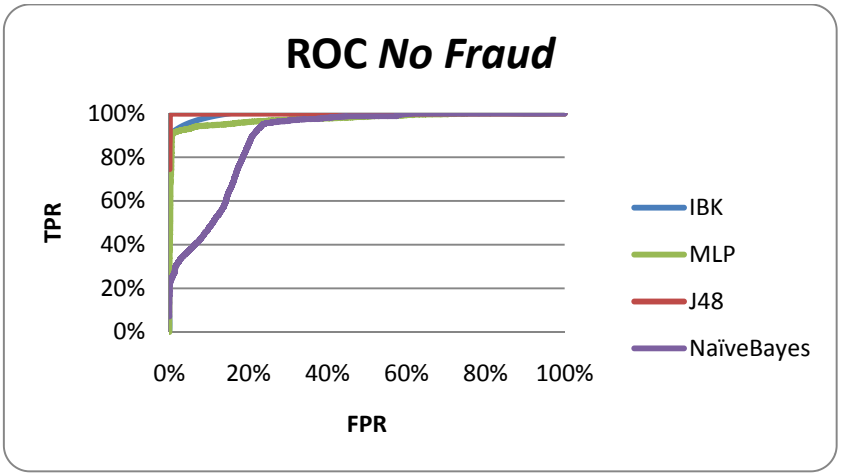

Figura 25 – Curvas ROC para o caso de fraude *No Fraud*

5.2. Regressão

Ao correr os algoritmos IBK, MLP e J48 com os dados gerados no modo de regressão foram obtidos os valores apresentados na Tabela 23. O algoritmo NaïveBayes foi excluído por não suportar o tipo de dados numérico, pois em modo de regressão todos os outros algoritmos apenas suportavam *datasets* com valores numéricos. O algoritmo com piores valores de *Root Mean Squared Error* (RMSE) e *Root Relative Squared Error* (RRSE) é o MLP, mas apesar de em termos comparativos ser o pior obteve uma percentagem de RRSE abaixo de 100% o que significa que o método de previsão avaliado é melhor do que o método simples da média.

	RMSE	RRSE
IBK	0,00	0%
MLP	0,01	40%
REPTree	0,00	0%

Tabela 23 – Medidas de comparação dos modelos em Regressão

5.3. Sumário

Ao analisar os resultados obtidos foi possível verificar que em classificação o algoritmo de árvores de decisão, J48, foi o algoritmo com melhores resultados a nível de percentagem correcta de classificações, performance, tempo de treino e precisão.

Nos gráficos de curvas ROC podemos constatar que todos os algoritmos obtiveram classificações para todos os casos de fraude acima da linha diagonal que representa uma classificação aleatória. Para todos os casos de fraude o algoritmo J48 demonstrou ser o mais próximo das

Em modo de regressão podemos concluir que o algoritmo REPTree, de árvores de decisão, e o IBK, de raciocínio baseado em casos, obtiveram valores óptimos de RRSE e RMSE.

O caso de fraude com menor número de instâncias correctamente classificadas foi o *Stuffing*, sendo os CDRs muito similares aos dos outros casos de fraude, variando apenas o IMSI do A *number* para o mesmo IMEI justificando-se assim a fraca capacidade de classificação deste caso por todos os algoritmos.

Capítulo 6

Conclusão

Neste capítulo são apresentadas e discutidas as conclusões resultantes do trabalho desenvolvido nesta dissertação. Tem início com a síntese da dissertação, seguindo-se uma discussão sobre a mesma com referência às limitações existentes e considerações sobre o trabalho. Finalmente, são apresentadas algumas possibilidades de trabalho futuro.

6.1. Síntese

A fraude é um problema que afecta as grandes operadoras de telecomunicações e as suas receitas anuais. Num mercado competitivo existe pressão para as operadoras aumentarem a sua eficiência, imagem, qualidade de serviço e reduzirem os seus custos. São coleccionados e armazenados grandes volumes de dados como resultado do aumento do uso dos serviços de comunicações móveis. O enorme volume de actividades de chamadas numa rede significa que a detecção e prevenção de fraude e sua análise são um grande desafio. A informação e conhecimento derivado destes dados podem fornecer às operadoras uma vantagem competitiva.

O processo de descoberta de conhecimento em Bases de Dados implementa a extracção semi-automática de conhecimento de alto nível a partir dos dados armazenados. Caracteriza-se por várias etapas das quais se destaca o *Data Mining*. É nesta etapa que são utilizados algoritmos de extracção de conhecimento.

Nesta dissertação foram abordados modelos de aprendizagem supervisionada, adequados para o objectivo de classificação e previsão, de modo a encontrar a similaridade de eventos de subscritores na rede para aplicação do método de prevenção de fraude *fingerprinting*. Os

modelos abordados enquadram-se em quadro paradigmas distintos: Raciocínio Baseado em Casos, Redes Neuronais, Árvores de Decisão e Modelos Bayesianos. A cada modelo estão associados algoritmos para identificar padrões e relacionamentos.

Os algoritmos dos modelos de *Data Mining* foram treinados e testados com os dados gerados, com experiências de classificação e regressão, de modo a ser possível comparar o seu desempenho na detecção dos diversos tipos de fraude. Para concretizar estas fases foi necessário desenvolver uma ferramenta que simulasse eventos de subscritores, num formato similar ao de um CDR, e produzisse dados de acordo com as fases de pré-processamento de *Data Mining* de selecção de dados, transformação de variáveis e substituição de valores omissos. Foram gerados dados de subscritores para 7 casos de fraude mais comuns e dados de subscritores não fraudulentos aproximados de cada um dos casos de fraude escolhidos.

Dado que uma análise exaustiva de todas as ferramentas de *Data Mining* se encontra fora do âmbito desta dissertação, optou-se somente pela utilização da ferramenta Weka pois implementa algoritmos para as técnicas de *Data Mining* escolhidas e fornece uma forma rápida e fácil de explorar e analisar dados. Contudo, existe um grande número de ferramentas com diversas implementações de algoritmos para cada técnica. A ferramenta Weka implementa as técnicas RBC, RNAs, Árvores de Decisão/Regressão e Redes de Bayes, através dos algoritmos IBK, MLPs, J48/REPTree e NaïveBayes, para cada técnica, respectivamente, podendo estas ser aplicadas tanto à classificação como à regressão.

A análise de resultados é baseada no número de classificações correctas para cada caso de fraude e de não fraude, avaliando cada modelo através de matrizes de confusão, tempos de execução, percentagens de erro e curvas ROC. Os melhores resultados foram obtidos pelo algoritmo de Árvores de Decisão J48 para todos os casos de fraude.

6.2. Discussão

Nesta dissertação, pretendeu-se avaliar qual a qualidade, em termos de previsão, obtida pelos algoritmos IBK, MLP, J48/REPTree e NaïveBayes das técnicas de *Data Mining* RBC, RNAs, Árvores de Decisão e Redes de Bayes, respectivamente. Estas técnicas têm

diversos parâmetros, bem como variações de algoritmos, que afectam o seu desempenho final. Foram efectuadas duas experiências, uma de classificação e uma de regressão, com validação cruzada de 10-*folds* e 20 repetições.

Os resultados obtidos revelam uma necessidade de tempo de treino muito elevada por parte do algoritmo das RNAs (MLP), o que se torna relevante quando se pretende implementar num domínio de aplicação com quantidades de dados de elevada dimensão. As árvores de decisão obtiveram os melhores resultados de classificação (J48) e regressão (RETree) com percentagens de erro e número de instâncias classificadas correctamente acima das restantes técnicas.

A avaliação dos modelos através das curvas ROC permitiu constatar que as árvores de decisão (J48) obtiveram as curvas mais próximas da curva ROC ideal para todos os casos de fraude. O caso de fraude com maior dificuldade de classificação para todos os algoritmos foi o *Stuffing* o que se deve ao facto de este caso de fraude poder ter instâncias que, isoladamente, sejam similares às de outros casos de fraude.

6.2.1. Limitações

Importa referir que existem limitações no que diz respeito à análise comparativa dos algoritmos, nomeadamente:

- Os conjuntos de dados de teste foram gerados por uma ferramenta, para apenas 7 casos de fraude, não existindo garantias que os resultados obtidos correspondam ao que se espere encontrar no mundo real;
- Não foram analisadas métricas de avaliação dos modelos como: tempo de processamento com grandes volumes de dados, facilidade de compreensão, novidade e utilidade do conhecimento adquirido.

No entanto, convém salientar que os resultados obtidos pelos algoritmos são indicativos e perspectivam a aplicabilidade do *Data Mining* no contexto de prevenção de fraude nas telecomunicações.

6.2.2. Considerações Sobre o Trabalho Realizado

Criou-se um cenário de estudo introduzindo os conceitos de fraude nas telecomunicações salientando a diferença entre detecção e prevenção de fraude. Apresentam-se os principais tipos de fraude existentes assim como os principais métodos de detecção e prevenção de fraude usados no mercado e algumas das ferramentas existentes na actualidade.

Efectuou-se um estudo do processo de descoberta de conhecimento em bases de dados e as técnicas mais adequadas de *Data Mining* para o objectivo de prevenção de fraude nas telecomunicações. Estudou-se os processos e os algoritmos existentes para as seguintes técnicas: Raciocínio Baseado em Casos, Redes Neuronais, Árvores de Decisão e Redes de Bayes.

Foi desenvolvida uma ferramenta de geração de dados para simular eventos de subscritores fraudulentos com foco nos seguintes casos de fraude: *Automatic Dialer*, *Dispersion*, *Ratio International/National*, *Ratio* SMS/*Voice*, *Cell Concentration*, *Stuffing* e *Bypass*. Foram também gerados dados de subscritores com eventos que não resultaram em nenhum caso de fraude.

Foi usada uma ferramenta de utilização livre, o Weka, cuja utilização é explicada com detalhe, de modo a guiar um utilizador inexperiente em aplicações semelhantes. O Weka permitiu realizar experiências de classificação e regressão sobre algoritmos das diferentes técnicas de *Data Mining*, IBK, MLP, J48/REPTree e NaïveBayes. Os modelos foram treinados e testados com os dados gerados.

A análise dos resultados obtidos de classificação e regressão foi efectuada com as técnicas de avaliação de modelos matrizes de confusão, curvas *Receiver-Operating Characteristic* (ROC), precisão, tempos de treino e medidas de regressão.

6.3. Trabalho Futuro

De modo a complementar e dar continuidade ao estudo da aplicabilidade de técnicas de *Data Mining* na prevenção de fraude nas telecomunicações, com foco no método *fingerprinting*, são proporcionadas diversas perspectivas de trabalho futuro, nomeadamente:

- Na presente dissertação foram seleccionadas 4 técnicas de *Data Mining* com o objectivo de previsão. A escolha das técnicas pode ser alargada para abranger também o paradigma evolucionário. O número de algoritmos por técnica pode também ser alargado de modo a comparar resultados obtidos por diversos algoritmos dentro da mesma técnica;
- Os dados utilizados foram dados gerados de modo a simular 7 casos de fraude. Com a utilização de conjuntos de dados, para treino e teste dos modelos, de grandes volumes e do mundo real será possível obter melhores resultados de classificação dos modelos;
- A ferramenta seleccionada para realizar as experiências foi o Weka. Existem no mercado diversas ferramentas que implementam algoritmos das técnicas de *Data Mining*. A fase das experiências para avaliação dos algoritmos pode ser alargada a mais ferramentas;
- As experiências realizadas, em classificação e regressão, basearam-se na validação cruzada dos dados com 10-*folds* e 20 repetições. Podem ser efectuadas mais experiências de modo a comparar a performance dos modelos com base nas percentagens de dados usados para treino e teste;
- A avaliação dos resultados, obtidos das experiências de classificação e regressão dos modelos pode considerar mais métricas como tempo de processamento com grandes volumes de dados, facilidade de compreensão, novidade e utilidade do conhecimento adquirido. a compreensão dos modelos criados é deveras importante em aplicações de *Data Mining*. Mais do que apenas aplicar o modelo, permite ao utilizador validar e utilizar o conhecimento extraído.

Bibliografia

1. **Fayaad, Usama M., et al.** *Advances in Knowledge Discovery and Data Mining.* USA : AAAI Press/MIT Press, 1996.

2. **Brown, Stephen.** *Telecommunication Fraud Management.* s.l. : Waveroad Securit, 2005.

3. *CRISP-DM 1.0 - Step-by-Step Data Mining Guide.* **Chapman, Pete, et al.** 2000. CRISP-DM Consortium.

4. **Nicholson, Mike.** Solution to Eliminate the Risk Inherent of Telecom Subscribers. [Online] 2008. http://www.articlesbase.com/communication-articles/solution-to-eliminate-the-risk-inherent-of-telecom-subscribers-687383.html.

5. *Classification, detection and prosecution of fraud in mobile networks.* **Gosset, P. e Hyland, M.** Sorrento, Itália : s.n., 1999. Proceedings of ACTS Mobile Summit.

6. **Bolton, R. J. e Hand, D. J.** Statistical Fraud Detection: A Review. *Statistical Science.* 2002.

7. **Hilas, C. e Sahalos, J.** *User profiling for Fraud Detection in Telecommunication Networks.* Aristotle University of Thessaloniki : s.n.

8. *Application of Familiar Technology to Minimize Losses.* **Hawes, Timothy K.** 1999, Information Management Magazine.

9. *A hybrid system for fraud detection.* **Verrelst, Herman, et al.** Bruges (Belgium) : D-Facto public., 1999. ESANN'1999 proceedings - European Symposium on Artificial Neural Networks. pp. 447-454.

10. **Dasgupta, Koustuv, et al.** Social Ties and their Relevance to Churn in Mobile. University of Maryland Baltimore County : IBM India Research Lab.

11. *Fraudulent Ways.* **Rebordit.** 2005, Telecommunications International.

12. **Adriaans, P. e Zantinge, D.** *Data Mining.* s.l. : Addison-Wesley, 1996.

13. **Rud, Olivia Parr.** *Data Mining Cookbook.* USA : John Wiley & Sons, Inc, 2001.

14. **Rezende, O. S.** *Sistemas Inteligentes Fundamentos e Aplicações.* Brazil : Editora Manole, Lda, 2003.

15. **Mitchel, T. M.** *Machine Learning.* Boston, USA : McGraw-Hill, 1998.

16. **Santos, M. Filipe e Azevedo, Carla.** *Data Mining.* s.l. : FCA, 2005.

17. **Novais, Paulo e Neves, José Maia.** Raciocínio Baseado em Casos. Universidade do Minho, Escola de Engenharia, Departamento de Informática : s.n., 1998.

18. **Kolodner, J. e Leake, D.** *A tutorial introduction to CBR. Case-Based Reasoning: Experiences, Lessons, and Future Directions.* s.l. : Menlo Park: AAAI Press/The MIT Press, 1996.

19. **Watson, Ian.** *Applying Case-Based Reasoning: tecniques for enterprise systems.* San Francisco : Morgan Kaufmann, 1997.

20. *Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches.* **Aamodt, A. e Plaza, E.** 1994. Artificial Intelligence Communications.

21. **Weber, Rosina.** Intelligent Jurisprudence Research. *Phd Thesis.* Florianópolis : UFSC, 1998.

22. **Leake, D.** *Case-Based Reasoning: Experiences, Lessons, and Future Directions.* Menlo Park : AAAI Press/The MIT Press, 1996.

23. **Kolodner, J.** *Case-Based Reasoning.* Los Altos : Morgan Kaufmann, 1993.

24. **Vergara e Walter, H.** *Simulação cognitiva da tomada de decisão em situações complexas: modelagem do raciocínio humano por meio de casos.* Florianópolis : UFSC, 1995.

25. **Berson, Alex, Stephen, Smith e Thearling, Kurt.** *Building Data Mining Applications for CRM.* USA : McGraw-Hill, 2000.

26. **Moreau, Yves, et al.** Novel Techniques for Fraud Detection in Mobile Telecommunication Networks. 1996.

27. **Cortez, Paulo e Neves, José.** *Redes Neuronais Artificiais.* Braga, Portugal : Universidade do Minho, 2000.

28. **Patterson, D.** *Artificial Neural Networks - Theory and Applications.* Singapura : Prentice Hall, 1996.

29. **Chester, Michael.** *Neural Network - A tutorial.* USA : PTR Prentice Hall, Inc., 1993.

30. *Faster Learning Variations on Back-Propagation: An Empirical Study.* **Fahlman, S.** USA : Morgan Kaufmann Publishers, 1998. Proceedings of Connectionist Models Summer School.

31. *Supervised Learning in Multilayer Perceptrons – from Backpropagation to Adaptive Learning Techniques.* **Riedmiller, M.** s.l. : Computer Standards an Interfaces, 1994, Computer Standards and Interfaces.

32. **Zupan, J. e Gasteiger, J.** *Neural Networks For Chemists: An Introduction.* New York : VCH, 1993.

33. **Rich, Elaine e Knight, Kevin.** *Artificial Intelligence.* Singapura : McGraw Hill, 1991.

34. **Michalski, Ryszard S., Bratko, Ivan e Miroslav, Kubat.** *Machine Learning and Data Mining Methods and Applications.* England : John Wiley & Sons, Inc., 1998.

35. *Problems in the analysis of survey data, and a proposal.* **Morgan, J. N. e Sonquist, J. A.** s.l. : Journal of the American Statistical Association, 1963, Vol. 58.

36. **Hartigan, J.** *Clustering Algorithms.* New York : Wiley, 1975.

37. **Breiman, L., et al.** *Classification and Regression Trees.* Belmont, CA : Wadsworth, 1984.

38. **Quinlan, J. R.** *Induction of decision trees.* s.l. : Machine Learning, 1986. Vol. 1.

39. **Pearl, J.** *Probabilistic reasoning in Intelligent Systems: Networks of Plausible Inference.* s.l. : Morgan Kaufmann, 1988.

40. **Heckerman, David.** *A Tutorial on Learning with Bayesian Networks.* s.l. : Microsoft Research Advanced Technology Division, 1996. Technical Report MSR-TR-95-06.

41. **Jiawei, Han e Kamber, Micheline.** *Data Mining: Concepts and Techniques.* USA : Morgan Kaufmann Publishers, 2001.

42. **Krause, P. J.** *Learning Probabilistic Networks.* UK : Philips Research Laboratory, 1998. Technical Report.

43. **Cios, Krzysztof J., Pedrycz, Witold e Swiniarski, Roman W.** *Data Mining Methods for Knowledge Discovery.* USA : s.n., 1998.

44. **Jensen, Finn V.** *An Introduction to Bayesian Networks.* s.l. : UCL Press, 1996.

45. **Kohavi, R. e Provost, F.** *Machine Learning.* 1998.

46. *The use of the area under the ROC curve in the evaluation of machine learning algorithms.* **Bradley, Andrew P.** Pattern Recognition, 1997, Vol. 30.

47. **Witten, I. H. e Frank, E.** *Data Mining - Pratical Machine Learning Tools and Techniques.* s.l. : Elsevier, 2005.

48. **Witten, Ian H. e Frank, Eibe.** *Data Mining: Practical machine learning tools and techniques.* San Francisco : Morgan Kaufmann, 2005.

49. *From Data Mining to Knowledge.* **Fayyad, U., Shapiro, G. e Smyth, P.** 1996, AI Magazine.

50. **Han, Jiawei e Kamber, Micheline.** *Data Mining: Concepts and Techniques.* USA : Morgan Kaufmann Publishers, 2001.

51. **Abidogun, O.** Data Mining, Fraud Detection and Mobile Telecommunications: Call Pattern Analysis with Unsupervised Neural Networks. *Dissertation for the degree of Master of Science.* 2005.

52. **Ramos, C., Neves, J. e Machado, J.** *Progress in Artificial Intelligence.* 13th Portuguese Conference on Artificial Intelligence : Springer, 2007.

53. **Riva, G., et al.** *Ambient Intelligence: The Evolution Of Technology, Communication And Cognition Towards The Future Of Human-Computer Interaction.* s.l. : IOS Press, 2005.

54. Ambient Intelligence and Ubiquitous Computing. [autor do livro] H. H. Adelsberger, et al. *Handbook on Information Technologies for Education and Training.* s.l. : Springer, 2008, pp. 79-100.

55. *Using reporting and data mining techniques to improve knowledge of subscribers. Applications to customer profiling and fraud management.* **Amat, Jean-Louis.** Alcatel, France : s.n., 2001. New trends in telecommunications.

56. **Rocha, Miguel, Cortez, Paulo e Neves, José Maia.** *Análise Inteligente de Dados - Algoritmos e Implementação em JAVA.* s.l. : FCA, 2008.

57. *An Introduction to Case-Based Reasoning.* **Blake, Deirdre.** 2008, Theory, History, Application Development.

58. **Hollm'en, J.** User Profiling and Classification for Fraud Detection in Mobile Communications Networks. *Dissertion for the degree of Doctor of Science.* Helsinki University of Technology : s.n., 2000.

Glossário

AA	Aprendizagem Automática
ADSL	*Asymmetric Digital Subscriber Line*: tecnologia de comunicação de dados que permite uma transmissão de dados mais rápida através de linhas de telefone do que um modem convencional pode oferecer.
AG	Algoritmos Genéticos
AID	*Automatic Interaction Detection*
ARFF	*Attribute Relationship File Format*
AUC	Área Abaixo da Curva
BD	Bases de Dados
BSS	*Business Support Systems*
CAAT	*Computer Assisted Auditing Techniques*
CART	*Classification and Regression Trees*
CDMA	*Code Division Multiple Access*
CDR	*Call Detail Record*
CHAID	*Chi-square Automatic Interaction Detection*
CLI	*Command Line Interface*
CRISP-DM	*CRoss-Industry Standard Process for Data Mining*

CSV	*Comma Separated Values*
DATAMART	Sub-componente de uma DW para dar resposta a um conjunto específico de análises
DM	*Data Mining:* processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.
DW	*Data Warehouse*: estrutura de informação só de consulta concebida para facultar a análise de dados corporativos de diversos sistemas ou fontes de empresa
DCBD	Descoberta de Conhecimento das Bases de Dados
DSS	*Decision Support System*
ESN	*Electronic Serial Number*
FIINA	Fórum Internacional de Acesso Irregular à Rede
FMS	*Fraud Management System*
FN	Falsos Negativos
FP	Falsos Positivos
GPRS	*General Packet Radio Service:* tecnologia que aumenta as taxas de transferência de dados nas redes GSM existentes.
GSM	*Global System for Mobile Communications*
GUI	*Graphic User Interface*
HUR	*High Usage Report*
IA	Inteligência Artificial

ID	*Iterative Dichtomizer*
IMSI	*International Mobile Subscriber Identity*
IMEI	*International Mobile Equipment Identity*
IP	*Internet Protocol*
IREG	Grupo de Especialistas do Ramo de *Roaming* Internacional
KDD	*Knowledge Discovery in Databases*: processo de extracção de informações de base de dados, que cria relações de interesse que não são observadas pelo especialista no assunto, bem como auxilia a validação de conhecimento extraído.
LOO	*Leave-one-out*
MAD	*Mean Absolute Deviation*
MCC	Código Móvel do País
MLP	Redes Perceptrão Multicamada
MMS	*Multimedia Message Service*
MNC	Código da Rede Móvel
MNI	*Mobile Number Identification*
MSE	*Mean Squared Error*
MSIN	Número de Identificação na Estação Móvel
NN	*Nearest Neighbour*
OLAP	*On-Line Analytical Processing*
OSS	*Operational Support Systems*
PBX	*Private Branch Exchange*

PIN	*Personal Identification Number*
PKI	*Public Key Infrastructure*
PMS	*Property Management Systems*
PRS	*Premium Rate Services*
RBC	Raciocínio Baseado em Casos
RDIS	Rede Digital Integrada de Serviços
REP	*Reduced Error Pruning*
RFMC	Rede *Feedforward* Multi-camada
RMI	*Remote Method Invocation*
RMSE	*Root Mean Squared Error*
RN	Rede Neuronal
RNA	Rede Neuronal Artificial
RNR	Rede Neuronal Recorrente
ROC	*Receiver-Operating Characteristic*
RPROP	*Resilient Backpropagation*
RRSE	*Root Relative Squared Error*
SAS	Sistema de Segurança de Acreditação
SIM	*Subscriber Identity Module*
SMDR	*Station Message Detail Recording*
SMS	*Short Message Service*
SS7	Sistema de Sinalização por Canal Comum Número 7: sistema

	com um canal específico para troca de sinalização, isto é, um dos canais ao invés de enviar informações digitalizadas da conversação é utilizado somente para enviar informações de sinalização comum a diversas chamadas.
SSE	*Sum Squared Error*
SSL	*Secure Socket Layer*
TAP	*Transferred Account Procedure*
TAS	*Tactical Analysis System*
TFP	Taxa de Falsos Positivos
TN	Negativos Verdadeiros
TP	Positivos Verdadeiros
TPR	Taxa de Positivos Verdadeiros
UMTS	*Universal Mobile Telecommunication System*
VoIP	Voz sobre IP
VPN	*Virtual Private Network*
WAP	*Wireless Application Protocol*