REGRAS DE ASSOCIAÇÃO – *MARKET BASKET ANALYSIS*
ITENS FREQUENTES E ITENS RAROS

por

Filomena Clara Gouveia Anselmo

Tese de Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Orientada por:

Professor Doutor João Manuel Portela da Gama
Professor Doutor Carlos Manuel Abreu Gomes Ferreira

2017

Agradecimentos

"Qualquer conquista começa com a decisão de tentar" (Gail Devers).

Partir da tentativa e chegar a este trabalho final envolveu um longo caminho de dedicação, trabalho, sacrifício, foco e suporte das pessoas que estão mais próximas de mim. No final desta minha etapa académica não podia deixar de prestar alguns agradecimentos:

Ao Professor Doutor João Gama e ao Professor Doutor Carlos Ferreira pela disponibilidade, incentivo e fundamental apoio quando as ferramentas disponíveis pareciam não querer ajudar a que o trabalho avançasse.

Aos responsáveis da empresa alvo de estudo por gentilmente terem autorizado a utilização da base de dados, possibilitando o estudo em questão e a respetiva publicação do mesmo em âmbito académico.

A toda a minha família, em especial aos meus avós pela educação e estrutura da pessoa que sou hoje, aos meus pais pelo incentivo em procurar sempre mais e mais saber, à minha irmã que para além de me incentivar é também um exemplo de dedicação e empenho, e ao Ricardo por me apoiar e acompanhar sempre em cada nova conquista e nunca deixar que me faltem as forças.

Aos amigos "antigos" que me desculpam a ausência e aos que ganhei neste caminho. Deixo o meu agradecimento ao Daniel Magalhães por me ter desafiado a fazer o mestrado, à Vânia Moutinho, ao André Martinez e à Sílvia Carvalho pelos momentos de entreajuda, partilha e apoio ao longo destes dois anos.

Um especial agradecimento à Natália Silva, companhia dos muitos pares de fins-de-semana despendidos neste trabalho, pelo entusiamo, e pelo "está quase"!

A todos que fazem parte do meu percurso o meu muito obrigada porque, acredito mesmo, que tudo que sou e que alcanço é também resultado de um pouco de vós.

Resumo

As técnicas de *Data Mining* e as ferramentas que as sustentam têm como intuito a obtenção de conhecimento útil em base de dados com grandes quantidades de informação com o intuito de auxiliar a antever as tendências que, por exemplo, as vendas numa empresa podem ter no futuro de curto prazo.

A descoberta de Regras de Associação tem um papel fundamental na descoberta de relacionamentos relevantes entre itens numa transação.

Este trabalho tem como principal objetivo o estudo das Regras de Associação com recurso a técnicas de *Data Mining*, especificamente de *Market Basket Analysis*. As transações que foram exploradas pertencem a uma empresa que comercializa produtos ligados à área da saúde de forma a encontrar padrões de consumo. A empresa alvo do estudo comercializa fundamentalmente dois tipos de produtos, os equipamentos médicos e os consumíveis clínicos, que têm diferentes frequências nas transações pois os primeiros são vendidos em muito menor quantidade que os segundos. Por este motivo, este trabalho é dividido em dois grandes pontos: a análise de regras de associação com itens frequentes, cujo foco principal são os consumíveis clínicos, e o estudo de padrões de associação com *itemsets* raros, que se centra fundamentalmente no consumo de equipamentos médicos.

Este estudo proporcionou uma visão suportada do relacionamento entre produtos, o que permite ter aplicações práticas ao nível do posicionamento de produtos em catálogo, bem como, instruir a equipa comercial das regras de associação mais pertinentes por forma a potenciar as vendas e aumentar o vínculo ao cliente.

Palavras-Chave: *Data Mining, Regras de Associação, Market Basket Analysis, Conjuntos de Itens Frequentes, Conjuntos de Itens Raros, Estudo de Caso*

Abstract

The techniques of Data Mining and the tools that support them are intended to obtain useful knowledge in a database with large amounts of information in order to help anticipate the trends that, for example, the sales in a company may have in short-term future.

The discovery of Association Rules has a fundamental role in finding relevant relationships between items in a transaction.

This work has as main objective the study of association rules using Data Mining techniques, specifically Market Basket Analysis. Were explored the transactions of a company that sells products related to health in order to find consumption patterns. The company targeted deals with two types of products: medical equipment's and clinical consumables, which have different frequencies in data base because the first ones are sold in smaller quantities than the second. For this reason, this work is divided in two main points: the analysis of Association Rules with frequent items, which main focus are clinical consumables, and the study of association patterns with rare *itemsets*, the consumption of medical equipment's.

This study provided a supported view of the relationship between products which allows to have practical applications in the positioning of products in catalog as well as to instruct the commercial team of the more pertinent association rules in order to boost sales and increase customer link.

Keywords: *Data Mining, Association Rules, Market Basket Analysis, Frequent Itemsets, Rare Itemsets, Case Study*

Ao Ricardo, meu tudo

Índice

Índice de Figuras

Índice de Tabelas

Índice de Fórmulas

1. Introdução

Ao longo dos últimos tempos temos vindo a assistir a um avanço gigantesco no que diz respeito à obtenção, armazenamento e tratamento de dados. Efetivamente, hoje em dia, vivemos com a presença de grandes volumes de informação. O avanço tecnológico bem como o uso constante e cada vez maior de computadores e tecnologia tem vindo a potenciar a existência de informação em larga escala. Esta realidade, aliada ao baixo custo de armazenamento, explica o aumento de armazenamento de informação a cada dia que passa, e este crescimento continua a progredir a passos largos não se conseguindo prever o fim do seu crescimento.

Desta constante evolução na obtenção de dados surgiu a necessidade de armazenar, tratar e compreender os grandes volumes de dados que se acumulam, com todas as vantagens que tal possa trazer.

As empresas, por exemplo, começaram a olhar de forma diferente e mais interessada para os seus dados e começaram a questionar-se de que forma o tratamento dos dados e posterior análise dos mesmos poderia trazer vantagens sobre o seu negócio, sobre os seus clientes e sobre a otimização das suas decisões. Foi esta evolução e esta necessidade que fez emergir a área de *Business Intelligence* (BI) que teve grande impulso nos últimos anos.

De facto, não é suficiente a obtenção e armazenamento da informação nas empresas. É por isso fundamental a extração de conhecimento que possa estar camuflado nessa informação. De forma a responder a esta necessidade surgiram técnicas inovadoras e ferramentas que permitem a extração e conhecimento de grandes bases de dados. O *Knowledge Discovery in Databases (*KDD) ou Extração de Conhecimento refere-se ao processo de extração de informação de conhecimento relevante através de uma base de dados o qual implica a realização de vários passos: seleção, pré-processamento, transformação, *Data Mining*, interpretação e avaliação.

O *Data Mining* é então uma fase da Extração de Conhecimento em Base de Dados que utiliza ferramentas computacionais, em grandes volumes de dados, os quais são armazenados pelas organizações, com a finalidade de identificar informação potencialmente útil em dados complexos que permita obter conclusões sofisticadas

(Bastos, 2001). Fayyad, Piatetsky & Smyth (1996) referem-se ao *Data Mining* como "o processo não-trivial de identificar padrões novos, válidos, potencialmente úteis e, principalmente, compreensíveis por meio da observação dos dados contidos numa base de dados". Isto significa que a finalidade do *Data Mining* não é apenas identificar padrões e relações na informação mas sim conseguir atingir um conhecimento utilizável de forma útil na tomada de decisão. Esta fase do KDD tem sido um dos maiores alvos de entrega por parte dos investigadores. As técnicas de D*ata Mining* têm vindo a ser utilizadas com sucesso num grande número de casos reais, como é o caso da sua aplicação em diagnósticos médicos ou em análise de dados de vendas em lojas ou empresas (Gama et al., 2012).

Deste modo, este trabalho terá o seu foco nas Regras de Associação através da *Market Basket Analysis*. As Regras de Associação permitem avaliar se a presença de um conjunto de itens nos registos de uma base de dados implica a presença de um outro conjunto distinto de itens nos mesmos registos (Agrawal & Srikan, 1994). A *Market Basket Analysis* é considerada uma das áreas mais antigas de *Data Mining* (Raeder & Chawla, 2011) a qual pretende descrever o comportamento do consumo de clientes e dessa análise retirar conclusões sobre os padrões de compra entre artigos e os conjuntos de artigos frequentes.

Por um lado trata-se de um tema atual e dominante, por outro parece uma mais-valia que a escolha recaia sobre uma temática que possa ter aplicação prática na área do retalho. Tenciona-se assim que a experiência profissional conjuntamente com o conhecimento adquirido no mestrado possam ser um benefício quer para o aprofundamento do estudo académico na área selecionada, quer para o estudo dos dados em causa. Efetivamente, a utilidade para a empresa alvo do estudo destaca-se pelo conhecimento que possa ser adquirido através da análise dos artigos com maior presença nas transações efetuadas e a derivação de regras de compra entre artigos, ou seja, o conhecimento dos padrões de consumo dos seus clientes.

Assim, pretende-se através deste trabalho académico, estudar Regras de Associação entre artigos/itens, de modo a identificar os artigos que estão mais relacionados entre si no momento da compra. Com o presente estudo pretende-se ainda retirar conclusões acerca dos padrões de consumo dos clientes que possam trazer vantagens competitivas e influenciar positivamente as vendas líquidas.

A finalidade deste trabalho baseia-se então, na descoberta das regras de associação que permitem observar se a presença de um conjunto de itens nos registos de uma base de dados origina a presença de um outro conjunto distinto de itens nos mesmos registos.

A descoberta das regras será desenvolvida com base no perfil de transações de uma base de dados, gentilmente cedida por uma empresa que contém todas as transações da empresa alvo de estudo no ano de 2016. A empresa alvo de estudo é uma empresa de direito privado que atua no sector da saúde, na distribuição por grosso de equipamentos médicos, seus acessórios e consumíveis clínicos. Com mais de duas décadas de existência, tem a sua sede no Grande Porto, uma delegação em Lisboa, e dispõe de uma rede comercial que cobre todo o território nacional e ilhas.

Cada transação da base de dados corresponde a uma determinada compra, efetuada num dado momento temporal, que contém todos os artigos adquiridos nessa transação. Cada artigo é identificado por um código de produto único e através da definição intrínseca de codificação adotada pela empresa podemos perceber a que família o produto pertence e qual a sua natureza.

A empresa tem na sua atividade comercial variadíssimos produtos, os quais estão em constante atualização. O grande enfoque que existe no produto, leva à implementação de um produto novo sempre que tal se mostra oportuno e a descontinuação de outros artigos sempre que o mercado exija tal ajuste. Esta possível variedade de produtos oferecidos e a rapidez de adaptação perante variações da procura podem ser determinantes na sustentação do sucesso do negócio. Por isso, poderá ser importante ter esta premissa em conta nas análises efetuadas aos resultados obtidos.

A base de dados que vai ser analisada contém itens com frequências muito distintas. Existem aqueles que aparecem com muita frequência e cujas vendas são traduzidas em milhares de unidades, os consumíveis clínicos, e outros que aparecem em muito menor quantidade nas transações, os equipamentos médicos. Por esse motivo, este trabalho é composto por dois grandes subtemas: a derivação de regras de associação tipicamente estudadas, com algoritmos que têm o seu foco nos *itemsets* frequentes, que no caso em estudo correspondem aos consumíveis clínicos, mas também o conhecimento de padrões de associação com *itemsets* raros ou infrequentes, que na base de dados em causa se centra fundamentalmente no consumo de equipamentos médicos.

As transações presentes na base de dados são compostas por todos os artigos vendidos pela entidade. Conforme mencionado anteriormente, os produtos podem ser divididos em duas grandes categorias: os produtos clínicos ou consumíveis clínicos, e os equipamentos médicos. É importante referir para maior precisão do estudo, que existem outras categorias de produtos vendidos pela empresa como medicamento e peças para reparações técnicas. No entanto, de forma a facilitar o estudo decidiu-se englobar estes produtos na categoria dos consumíveis clínicos.

Apesar do baixo número de "unidades" vendidas comparativamente com os consumíveis clínicos, em 2016, a venda de equipamento médico representou cerca de 25% do volume de negócios, razão pela qual se mostra de extremo interesse utilizar técnicas que incluam os equipamentos médicos, como itens raros, no estudo das Regras de Associação.

Sendo a empresa em estudo uma empresa que atua como retalhista e grossista, é relevante o estudo da relação entre os produtos que comercializa. É então oportuno colocar questões como:

- Existe um conhecimento sobre os padrões de consumo dos clientes?

- Que tipos de produtos são usualmente adquiridos em conjunto e quais têm uma maior relação entre si?

- Existe uma recomendação de produtos aos clientes com base no conhecimento que o histórico de transações possa evidenciar? Havendo uma importância tão significativa do acompanhamento da equipa comercial nos clientes, estão os comerciais da empresa informados sobre o relacionamento entre produtos, para além do conhecimento adquirido à priori fruto da própria atividade comercial?

Este conjunto de questões pode ser incrementado com a questão que determinou também a escolha deste tema: de que forma poderiam as técnicas de *Data Mining* acrescentar algum valor na perceção e melhor compreensão destas questões?

Assim, o objetivo do trabalho passa por tentar compreender a associação/relação entre produtos adquiridos, mas também se quando um artigo é adquirido existe uma maior ou menor probabilidade de outro ou outros artigos serem adquiridos conjuntamente. A conclusão sobre este tipo de padrões poderá levar a uma visão mais clara sobre o comportamento de compra por parte dos clientes indicando que tipo de produtos ou

famílias de produtos usualmente são adquiridos em conjunto e quais os produtos que têm maior relação entre si.

No que diz respeito às regras de associação com base nos itens frequentes, estas permitem determinar, em geral, quais os produtos com maior relação, que representam os padrões de compra dos clientes de forma a ter uma visão dinâmica do comportamento dos clientes. Este tipo de conhecimento permite identificar produtos potencialmente alvo de campanhas de marketing como cross-selling e campanhas promocionais.

Já as regras de associação com itens raros, por apresentarem muito menor suporte, permitem ter uma indicação, ainda que menos forte, de relações entre itens.

O presente estudo explora o tema das regras de associação e inicia-se com uma revisão bibliográfica. Esta revisão constitui o primeiro capítulo que contempla uma breve referência ao processo de descoberta de conhecimento, principais conceitos implícitos no tema das regras de associação, suas medidas de avaliação e principais algoritmos para o estudo das mesmas.

O capítulo 3 refere-se ao estudo do caso e discussão dos resultados obtidos. Seguem-se-lhe as conclusões no capítulo 4 que precedem as referências bibliográficas utilizadas e que dão sustentabilidade a este trabalho.

2. Revisão de Literatura

Neste capítulo é realizada a revisão bibliográfica do trabalho, apresentando-se uma breve contextualização da extração de conhecimento em bases de dados e da importância deste tema do mundo atual.

São detalhados os conceitos gerais das regras de associação que são usados ao longo de todo o estudo bem como as principais medidas de avaliação das mesmas. São ainda explicados os algoritmos utilizados para o estudo das regras de associação bem como os *softwares* envolvidos.

2.1 *Data Mining* - Contextualização

O processo *Knowledge Discovery in Databases* (KDD) ou Processo de Descoberta em Bases de Dados foi definido por Brachman & Anand (1996) como um processo de natureza interativa composto por várias etapas.

Com conceito semelhante, Fayyad, Piatetsky & Smyth (1996) classificam o processo de KDD como interativo e iterativo, envolvendo várias etapas com muitas decisões tomadas pelo utilizador. As fases do ciclo de KDD proposto por Fayyad, Piatetsky & Smyth (1996) encontram-se ilustradas na Figura 1.

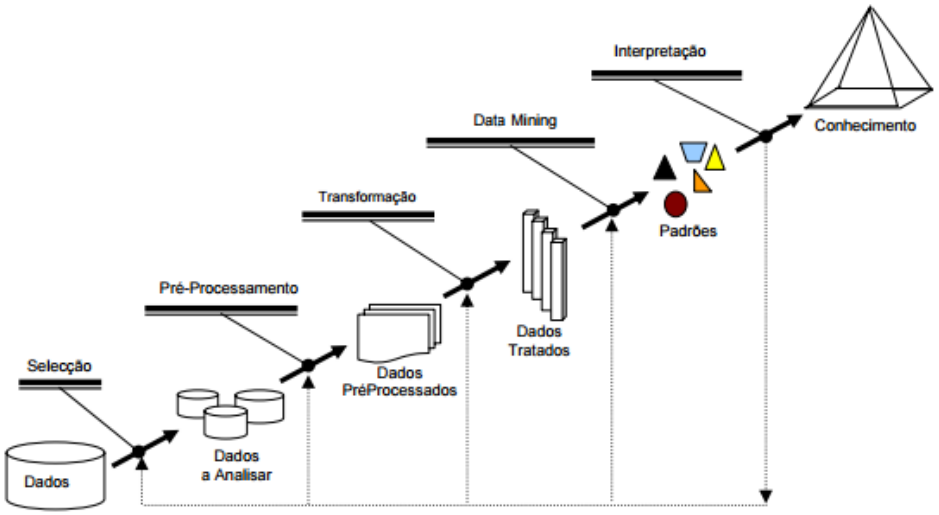

Figura 1 - Fases do Processo de KDD (adaptado de Fayyad, Piatetsky & Smyth (1996)

Em termos gerais, antes de se iniciar qualquer processo, é necessário obter um conhecimento prévio dos dados, bem como identificar o objetivo do processo KDD. Posteriormente, é essencial formar-se um conjunto de dados ou amostra de dados que será alvo de análise.

Após a fase de seleção surge a fase do pré-processamento ou limpeza de dados. Esta fase assume extrema relevância pois, muitas vezes, os dados extraídos das bases de dados possuem informação imprecisa e inconsistente. Tal deve-se ao facto de, por vezes, as bases de dados apresentarem campos de informação sem qualquer valor ou serem incorretamente preenchidas pelos utilizadores com valores incoerentes o que pode resultar numa grande quantidade de *outliers*. Nesta fase são, então, realizadas algumas operações que incluem o tratamento de problemas como ruídos e dados incompletos (Han, Kamber, & Pei, 2011).

O passo seguinte consiste na tarefa de transformação dos dados. Esta fase é de essencial importância, pois é neste momento que os dados tomam a forma final para serem processados pelas técnicas/algoritmos de *Data Mining*. Na transformação de dados integra-se algum conhecimento adquirido na área em estudo e aplicam-se algumas técnicas sobre as variáveis, de forma a reduzir a dimensionalidade da amostra. A transformação dos dados pode ser obtida através da redução de atributos que não são essenciais ou tão importantes, para determinado objetivo da descoberta.

Na sequência das fases, anteriormente, descritas de homogeneização dos dados, pode ser iniciado o processo de *Data Mining.* Este processo, consiste na análise de grandes volumes de dados com o intuito de descobrir informação útil que normalmente não seria visível ou que dificilmente seria encontrada. Segundo Fayyad, Piatetsky & Smyth (1996), investigadores pioneiros em KDD, "*Data Mining* é um passo no processo de KDD que consiste na aplicação de análise de dados e algoritmos de descobrimento que produzem uma enumeração de padrões (ou modelos) particular sobre os dados".

2.2 Regras de Associação

A temática da descoberta de Regras de Associação foi introduzida por Agrawal & Srikan (1994) como uma maneira de encontrar padrões associativos a partir de dados de um cesto de supermercado. Os dados dos cestos de supermercado consistem em transações em que cada transação é um conjunto de itens comprados por um cliente. A motivação deste tema consiste em aprender sobre padrões de compra num cesto de supermercado e usar essa informação de forma a conseguir vantagens reais para a empresa quer para melhor promoção dos produtos junto dos clientes, aplicação em *design* de catálogo, no *layout* de loja ou em campanhas promocionais.

Desde o seu aparecimento até ao momento, a exploração de Regras de Associação tem sido cada vez mais estudada e aplicada em diversos domínios. Alguns exemplos da sua aplicação são a deteção de fraude de cartões de crédito, os sistemas de deteção de intrusões de rede, os diagnósticos médicos ou as análises de dados genéticos. A título exemplificativo das diversas aplicações das regras de associação, Mccormick, Rudin, & Madigan (2000) coordenaram um estudo onde avaliaram o comportamento de pacientes que visitavam com regularidade os seus médicos de modo a identificar padrões e prever sintomas futuros mediante o seu historial médico.

Outra aplicação cada vez mais utilizada das regras de associação é a sua utilização como base para a aplicação de sistemas de recomendação que têm como objetivo antecipar as preferências de um utilizador com base nas preferências de um grupo de utilizadores (Goldberg, Nichols, Oki, & Terry, 1992). De facto, muitas organizações utilizam as Regras de Associação para conhecer os padrões de consumo e com eles recomendar produtos, músicas ou filmes. A Amazon e o Ebay são dois exemplos de empresas que utilizam a informação dos padrões de consumo dos seus clientes, através do histórico de pesquisa e compra, para criar recomendações de produtos, e assim, potenciar as suas vendas. A Netflix, por exemplo, é uma aplicação que incorpora um algoritmo de recomendação que tem em conta fatores como o género de filmes e séries de TV disponíveis, o histórico de cada utilizador, as classificações atribuídas pelo utilizador bem como a análise de todos os utilizadores da Netflix que possuem gostos similares ao utilizador em questão. O sistema de classificação desta aplicação é uma

maneira fácil de perceber o que cada utilizador mais gosta para proceder a um conjunto de sugestões relacionadas.

Em cada um destes domínios, é necessário analisar dados para identificar padrões que associam atributos diferente e a exploração de regras de associação atende a essa necessidade.

Segundo Brusso (2000) "as Regras de Associação são padrões descritivos que representam a probabilidade de que um conjunto de itens apareça em uma transação visto que outro conjunto está presente".

As Regras de Associação foram introduzidas pelos investigadores Agrawal, Imielinski, & Swami (1993) quando apresentaram um estudo que procurava relações entre os itens nas compras dos clientes numa visita ao supermercado. Com base no conceito de regras fortes introduziram regras de associação para descobrir as regularidades entre os produtos em dados de transações, em grande escala, registados pelos sistemas de pontos de venda nos supermercados. Por exemplo, a regra {Vegetais, Cebolas} ⇒ {Carne}, encontrada nos dados de vendas de um supermercado indicaria que se um cliente comprar vegetais e cebolas juntos, provavelmente também compraria carne.

Esta técnica de *Data Mining* permite descobrir se a presença de um conjunto de itens nos registos de uma base de dados implica a presença de um outro conjunto distinto de itens nos mesmos registos (Agrawal & Srikan, 1994). Os primeiros estudos desta área caíram sobre a análise de dados relativos a cestos de compras num supermercado, para identificar produtos que costumam ser adquiridos em conjunto. Por este motivo, este tipo de análise ficou conhecido como *Market Basket Analysis*. Desde esse momento até aos dias de hoje, muitas outras aplicações se têm dado às Regras de Associação, não se limitando a sua aplicação ao contexto do retalho, como referido anteriormente.

A *Market Basket Analysis* analisa as relações entre itens em cestos de mercado, sendo esta relação facilmente analisada no setor do retalho, onde os clientes compram vários produtos no mesmo "cesto de compras" (Han, Kamber & Pei, 2012). Os itens em cada cesto são registados como uma transação e pela sua análise torna-se evidente que itens específicos são comprados em conjunto, porque ocorrem, conjuntamente, e repetidamente nos registos das transações. Estes padrões que ocorrem em simultâneo são alvo de interesse porque se os clientes compram um destes artigos, provavelmente irão comprar o outro item relacionado também. Isto significa que, por exemplo, uma

promoção direcionada para aumentar as vendas em qualquer item dentro deste grupo de itens poderia também levar ao aumento das vendas nos outros itens.

Uma regra de associação é composta de dois conjuntos de itens: um antecedente ou lado esquerdo (LHS) e um consequente ou lado direito (RHS) e são representadas na forma de Antecedente → Consequente, que se pode interpretar da seguinte forma: "se antecedente então consequente, em que o antecedente e o consequente são *itemsets"*(Gama, Carvalho, Faceli, Lorena, & Oliveira, 2012).

A quantidade de itens pertencentes a um conjunto de itens é chamada de comprimento do conjunto. Um conjunto de itens de comprimento k é referenciado como um k-*itemset*.

2.2.1 Conceitos

Nas regras de associação, as medidas mais usadas são o suporte e a confiança e, portanto, é importante definir e compreender estes dois conceitos que assumem essencial importância tanto na etapa de pós-processamento e na avaliação do conhecimento, como na seleção de *itemsets* durante o processo de geração de regras. Outros conceitos serão também explicados por forma a facilitar o entendimento dos mesmos quando surgirem ao longo do trabalho.

2.2.1.1 Suporte

Nas regras de associação recorre-se a uma medida de incidência para definir quais as associações consideradas como significantes. A mais popular e usada dessas medidas é o suporte (contagem) dos *itemsets*.

O suporte é simplesmente o número de transações que incluem todos os itens na parte antecedente e consequente da regra. Assim, para uma determinada regra de associação $\{X\} \Rightarrow \{Y\}$, o suporte da regra mede o número total de registos de transação que contêm os conjuntos de itens X e Y.

Neste caso, o suporte da regra $\{X\} \Rightarrow \{Y\}$, onde X e Y são conjuntos de itens, seria dados pela seguinte expressão:

$$Suporte\ (\mathrm{X} \rightarrow \mathrm{Y}) = \frac{Frequência\ de\ \mathrm{X\ e\ Y}}{Total\ de\ T}$$

Fórmula 2.1 - Suporte da Regra de Associação (X → Y)

O numerador diz respeito ao número de transações em que X e Y ocorrem simultaneamente e o denominador refere-se ao número total de transações da base de dados.

Segundo Gama et al. (2012), e também de acordo com Agrawal et al. (1993) o suporte pode ser absoluto ou relativo.

O suporte absoluto de um *itemset* representa o número de transações onde se encontra esse *itemset* no banco de dados em questão. O suporte relativo representa a proporção de transações que contêm esse mesmo *itemset* e é calculado através da divisão do suporte absoluto pelo número total de transações.

Transações	Artigos
T1	{Agulha, Penso, Cabo ECG, Bisturi}
T2	{Penso, Cabo ECG, Bisturi}
T3	{Agulha, Eletrocardiógrafo, Cabo ECG, Contentor de 5lt}
T4	{Agulha, Penso, Cabo ECG}
T5	{Penso, Bisturi}
T6	{Contentor de 5lt, Bisturi}

Tabela 1 - Exemplo de base de dados com 6 transações

Na tabela 1 observa-se um exemplo de um conjunto de 6 transações que contêm 6 artigos. Utilizando o exemplo da tabela, o suporte absoluto das agulhas é 3, sendo o seu suporte relativo de (3/6) = 0,5 o que significa que 50% das transações do exemplo contêm o artigo agulhas.

O suporte permite-nos aferir quão comum determinado *itemsets* é na base de dados (Ulas, 1999).

2.2.1.2 Confiança

A outra medida utilizada na descoberta de regras de associação é a confiança do conjunto de itens frequentes. A confiança de uma regra é dada pela seguinte fórmula:

$$Confiança\ (\mathrm{X} \rightarrow \mathrm{Y}) = \frac{supp\ (\mathrm{X} \cup \mathrm{Y})}{supp\ (\mathrm{X})}$$

Fórmula 2.2 - Confiança da Regra de Associação (X → Y)

O numerador refere-se ao número de transações em que X e Y ocorrem simultaneamente. O denominador refere-se à quantidade de transações em que o item X ocorre. Em termos gerais a confiança mede a probabilidade condicional de ocorrer Y dado que ocorreu X.

2.2.1.3 Itens Frequentes

O conceito de item frequente é encontrado ao longo de todo o trabalho. Pode ser brevemente definido como aquele item que aparece em grande quantidade na base de dados.

Em termos gerais, um item é considerado frequente se satisfaz a seguinte condição:

$$\text{Suporte } \{\text{item}\} \geq \text{Suporte Mínimo}$$

2.2.1.4 Itens Frequentes Fechados e Itens Frequentes Máximos

Em certas análises é importante ter os cálculos dos itens máximos e dos itens fechados de forma a ter informação sintetizada dos *itemsets* para se reduzir os e evitar-se alguma redundância. Os conjuntos de itens frequentes fechados e máximos são subconjuntos de *itemsets* frequentes, mas os conjuntos de itens frequentes máximos são uma representação mais compacta porque são subconjuntos de *itemsets* frequentes fechados.

Um *itemset* é fechado se não tiver um superconjunto com a mesma frequência. Por exemplo, se temos um *itemset* de tamanho 2 *ad* que é frequente e temos um *itemset* de tamanho 3 *abd* que tem o mesmo suporte, então ad é frequente mas não é fechado.

Um *itemset* é máximo se nenhum dos seus superconjuntos imediatos são frequentes.

A figura 2 apresenta a relação entre *itemsets* frequentes, frequentes fechados e frequentes máximos.

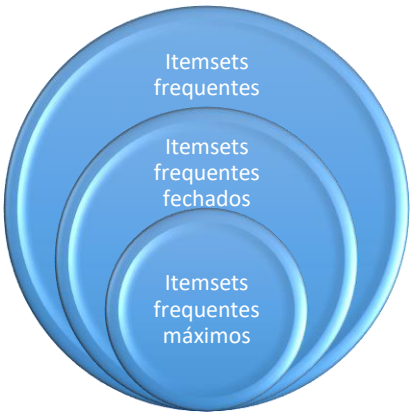

Figura 2 - Esquema relacional entre itemsets frequentes, fechados e máximos

2.2.1.5 Itens Raros

É pertinente perceber o conceito de raridade ou objetos raros ou infrequentes. Intuitivamente, podemos definir *itemsets* raros como aqueles que aparecem juntos em muito poucas transações, ou numa percentagem muito pequena das transações da base de dados.

Alguns autores definem que um item é considerado raro quando satisfaz a seguinte condição:

$$\text{Suporte } \{\text{item}\} \leq \text{Suporte Máximo}$$

2.2.2 Medidas de Avaliação/ Interesse das Regras de Associação

No que diz respeito à descoberta de Regras de Associação, as medidas de avaliação detêm um papel fundamental, pois é a partir dos resultados obtidos com estas medidas que conseguimos analisar qualitativamente as regras que são geradas com os diversos algoritmos.

É, por isso, importante avaliar as regras de associação descobertas quanto ao interesse e conhecimento das mesmas para o utilizador.

O grau de interesse de uma regra pode ser medido através das medidas de avaliação e as mesmas dividem-se em dois grandes grupos: as medidas de avaliação que utilizam critérios objetivos e as que usam critérios subjetivos (Gonçalves, 2005).

2.2.2.1 – Medidas de Avaliação Objetivas

As primeiras medidas objetivas utilizadas foram o suporte e confiança. Porém, têm surgido várias críticas ao longo dos tempos ao modelo Suporte/Confiança dado o enorme número de regras de associação que este modelo pode gerar, tornando árdua a análise das mesmas (Gonçalves, 2005).

Existem várias medidas estatísticas que podem ser utilizadas para avaliar as regras de associação descobertas. A correlação, por exemplo, analisa a força no relacionamento do conjunto. O *gini-index* é outra medida que mede o grau de correlação entre o antecedente e o consequente da regra. O teste do qui-quadrado, por sua vez, é uma medida que testa a independência e a correlação entre os itens da regra de associação. A Alavancagem ou *Leverage* ou PS (Piatetsky Shapiro, 1991) é uma medida que testa o interesse na regra descoberta através da análise da independência do conjunto de itens.

De seguida será abordada uma das medidas mais utilizadas que, por esse mesmo motivo, será a medida de avaliação das regras de associação considerada ao longo deste trabalho, o *Lift*.

Lift ou Interesse

O *lift* ou coeficiente de interesse é uma das medidas mais utilizadas para avaliar dependências entre o antecedente e o consequente da regra. A sua variação é entre zero e infinito e tem uma leitura muito simples: quanto maior o valor do *lift*, mais interessante é a regra, pois maior é a dependência entre os itens que a constituem (Brin, Motwani, Ullman, & Tsur, 1997).

Dada uma regra X → Y, esta medida indica que quanto mais frequente se torna Y quando X ocorre. O *lift* de uma regra é dada pela seguinte fórmula:

$$Lift(\mathrm{X} \rightarrow \mathrm{Y}) = \text{confiança}(\mathrm{X} \rightarrow \mathrm{Y}) / \text{suporte}(\mathrm{Y}) = \frac{\text{suporte}(\mathrm{X} \cup \mathrm{Y})}{\text{suporte}(\mathrm{X}) \times \text{suporte}(\mathrm{Y})}$$

Fórmula 2.3 - Lift da Regra de Associação (X → Y)

Se $Lift\ (X \rightarrow Y) = 1$, então X e Y são independentes.

Se $Lift\ (X \rightarrow Y) > 1$, então X e Y são positivamente dependentes.

Se $Lift\ (X \rightarrow Y) < 1$, então X e Y são negativamente dependentes.

Assim, se *lift* assumir o valor 1, existe independência entre X e Y. Valores para *lift* que se afastam de 1 indicam que a evidência de X fornece informação sobre Y e as regras começam a ter relevância de informação acerca da relação entre eles, pois o *lift* mede a distância para a independência entre X e Y.

O *lift* possuí como característica o fato de ser uma medida simétrica, ou seja:

$$Lift\ (X \rightarrow Y) = Lift\ (Y \rightarrow X)$$

Verifica-se a característica de simetria pois este índice tem como objetivo mensurar dependência entre os itens, e não medir a implicação (a orientação da seta da regra "⇒").

Outra característica que torna o *lift* uma medida muito interessante é que este consegue destacar com maior facilidade a dependência positiva entre conjuntos de itens que possuem suporte baixo.

O interesse das regras de associação pode ser presumido com a aplicação de medidas estatísticas objetivas. No entanto, avaliar as regras de associação unicamente pelas medidas estatísticas não garante o interesse de uma regra.

2.2.2.2 – Medidas de Avaliação Subjetivas

Por vezes pode não ser razoável utilizar uma única medida para analisar o interesse das regras de associação geradas pelos algoritmos. Um dos focos da descoberta de conhecimento tem sido a criação de medidas de interesse na descoberta de padrões (Silberschatz & Tuzhilin, 1995).

Enquanto as medidas de avaliação objetivas dependem apenas dos dados e medem estatisticamente a forma das regras, as medidas de avaliação subjetivas dependem do conhecimento e do interesse do utilizador.

Utilidade e Previsibilidade

De acordo com os critérios subjetivos, as regras de associação podem ser interessantes tendo em conta dois fatores: a regra ser ou não esperada e ser ou não útil. Como é fácil de perceber estes dois fatores são muito dependentes do utilizador que está a analisar a informação obtida nas regras (B. Liu, Hsu, Chen, & Ma, 2000), pelo que avaliar as regras de associação de acordo com estes dois fatores não é fácil pois utilizadores diferentes poderão ter interesses diferentes e até poderá acontecer que a opinião de um utilizador se altere ao longo do tempo.

As regras serão consideradas inesperadas se não forem conhecidas previamente pelo utilizador ou se forem opostas ao conhecimento que o mesmo possui dos dados.

As regras serão consideradas úteis se trouxerem conhecimento que possa ser utilizado na tomada de decisão. Assim, no que diz respeito às medidas de avaliação segundo critérios subjetivos podemos ter alguns cenários possíveis para as regras:

- Uma regra pode ser esperada e ser útil ou não útil.
- Uma regra pode ser inesperada e ser útil ou não útil.

2.3 Extração de Regras com Itens Frequentes

2.3.1 Algoritmo para Extração de Regras com Itens Frequentes

São variadíssimos os algoritmos mencionados na literatura para a deteção de Regras de Associação que foram surgindo ao longo dos últimos anos:

- Algoritmo Basic – trata-se de um algoritmo simples que apresenta uma forma fácil de incorporar a informação da hierarquia do artigo na derivação de regras de associação através da adição de todos os antepassados do artigo na transação em que ele se encontra (Srikant & Agrawal, 1995). Apesar de se tratar de um algoritmo simples é bastante lento;

- Algoritmos Cumulate e Est Merge - surgiram também em 1995, por Srikant e Agrawal por oposição ao algoritmo Basic que consideravam mais lento;

- Algoritmo DHP (Direct Hashing and Pruning) - apresentado pelos investigadores Chen, Park e Yu (Park, Chen, & Yu, 1995);

- Algoritmo DIC - apresentado pelos investigadores, Brin, Motwani, Uliman e Tsur (Brin et al., 1997);

- Algoritmo Genex - apresentado por Weber (Weber, 1998);

- Algoritmo FP-growth - desenvolvido pelos investigadores Han, Pei e Yin (Han, Pei, & Yin, 2000);

- Algoritmo PHP (Perfect Hashing and Pruning) - apresentado por Özel and Güvenir (Ozel & Guvenir, 2001);

- Algoritmo iFP-growth - apresentado por Siqueira, Prado, Júnior, Carvalho (Siqueira, Prado, Júnior, & Carvalho, 2002). Trata-se de um algoritmo incremental para determinar regras de associação, baseado no FP-growth;

- Algoritmo AFOPT - apresentado por Lu, Guimei, Xu, Wang e Xiao (G. Liu, Lu, Yu, Wang, & Xiao, 2003). Trata-se de uma otimização para descobrir itens frequentes e também baseia a sua estrutura no algoritmo FP-growth.

- Algoritmo AIS – (Agrawal et al., 1993) tem como principal limitação o facto de a descoberta das regras de associação estar limitada a apenas um item no consequente da regra;

2.3.1.1 Algoritmo *Apriori*

O algoritmo *Apriori* foi antecedido pelo algoritmo AIS e foi proposto por Agrawal e Srikant em 1994. Foi o primeiro algoritmo a reduzir de forma eficiente o espaço de pesquisa nos dados, o que melhorou substancialmente o desempenho na descoberta de regras de associação, sendo um dos algoritmos mais utilizados na extração de regras associação em grandes bases de dados. Veio melhorar o AIS no sentido em que veio permitir extrair regras de associação com mais que um item no consequente da regra.

O *Apriori* baseia-se no princípio de que se um *itemset* é frequente então, todos os seus subconjuntos são também frequentes, e utiliza uma estratégia de procura em largura (Gama et al., 2012). Este princípio é devido à seguinte propriedade do suporte:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

Isto significa que o suporte de um *itemset* nunca é maior que o suporte de seus subconjuntos. Tal é conhecido como a propriedade anti-monotónica do suporte (Agrawal et al., 1993).

Este algoritmo possui uma organização que garante uma grande flexibilidade na geração de regras de associação e a sua parametrização é feita com base no suporte mínimo e na confiança mínima. Estes valores mínimos para o suporte e confiança são pré-definidos pelo utilizador.

O algoritmo é executado da seguinte forma:

- Na primeira passagem pelos dados é contabilizado o suporte para cada item individualmente. São, então, determinados e selecionados aqueles que são frequentes, ou seja, os que apresentam suporte igual ou superior ao suporte mínimo estabelecido pelo utilizador. São, assim, constituídos os conjuntos-de-1-item frequentes. Os padrões não frequentes são eliminados. O algoritmo *Apriori* efetua a geração dos itens candidatos com base nos itens considerados frequentes na passagem anterior e cada passagem subsequente inicia-se com esse conjunto de dados pré-determinados.

- Assim, na segunda iteração, são gerados os candidatos a conjuntos-de-2-itens pela junção dos conjuntos-de-1-item propostos como frequentes na primeira iteração. São calculados os seus suportes e selecionados aqueles que são frequentes, ou seja, os que

apresentam suporte igual ou superior ao suporte mínimo estabelecido. Surgem assim os conjuntos-de-2-itens frequentes. Este grupo pré-determinado de itens considerados frequentes vai iniciar a iteração seguinte.

O algoritmo *Apriori* prossegue iterativamente, até que o conjunto-de-k-itens encontrado seja um conjunto vazio.

Numa segunda fase, o algoritmo *Apriori* gera as regras de associação derivadas dos *itemsets* frequentes considerando o valor mínimo definido para a confiança.

Tendo por base o exemplo apresentado na tabela 1, e fixando-se um suporte mínimo de 50% e uma confiança mínima de 70% teríamos:

$$Suporte\ (\text{penso}) = \frac{4}{6} = 66{,}67\%$$

Como suporte {penso} ≥ suporte mínimo estabelecido, então, item {penso} é considerado frequente.

O conjunto de itens frequentes de tamanho 1 é obtido através dos itens cujo suporte ≥ suporte mínimo fixado:

Conjuntos de 1 Item	Suporte
{Agulha}	50%
{Penso}	67%
{Cabo ECG}	67%
{Bisturi}	67%
{Eletrocardiógrafo}	17%
{Contentor de 5lt}	33%

Tabela 2 – Conjuntos de Itens Frequentes de Tamanho 1

São considerados como itens frequentes de tamanho 1: F1: {Agulha}, {Penso}, {Cabo ECG}, {Bisturi}.

Na segunda passagem pelos dados, obtém-se os seguintes conjuntos de tamanho 2:

Conjuntos de 2 Itens	Suporte
{Agulha, Penso}	33%
{Agulha, Cabo ECG}	50%
{Agulha, Bisturi}	17%
{Penso, Cabo ECG}	50%
{Penso, Bisturi}	50%
{Cabo ECG, Bisturi}	33%

Tabela 3 - Conjuntos de Itens Frequentes de Tamanho 2

São considerados como itens frequentes de tamanho 2: F2: {Agulha, Cabo ECG}, {Penso, Cabo ECG}, {Penso, Bisturi}.

F1: {Agulha}, {Penso}, {Cabo ECG}, {Bisturi}.

F2: {Agulha, Cabo ECG}, {Penso, Cabo ECG}, {Penso, Bisturi}.

Na terceira passagem pelos dados, obtém-se os seguintes conjuntos de 3 itens:

Conjuntos de 3 Itens	Suporte
{Agulha, Cabo ECG, Penso}	33%
{Agulha, Cabo ECG, Bisturi}	17%
{Cabo ECG, Penso, Bisturi}	33%

Tabela 4 - Conjuntos de Itens Frequentes de Tamanho 3

Dado que nenhum dos conjuntos de 3 itens tem suporte ≥ suporte mínimo fixado, então, para o exemplo fornecido, não existem conjuntos com três itens considerados como frequentes. Assim, no exemplo apresentado, a descoberta de conjuntos de itens frequentes terminaria aqui, pois foi o momento em que encontrou um conjunto vazio.

Assim, após a primeira fase em que são obtidos os conjuntos de itens frequentes que cumprem com o suporte mínimo fixado, proceder-se à derivação das regras de associação.

Tendo por base uma confiança mínima definida de 70%, seriam geradas as regras de confiança apresentadas na tabela 5 que cumprem com os requisitos definidos de suporte e confiança.

# Regra	Antecedente		Consequente	Suporte	Confiança
[1]	{Agulha}	=>	{Cabo ECG}	50%	100%
[2]	{Cabo ECG}	=>	{Agulha}	50%	75%
[3]	{Bisturi}	=>	{Penso}	50%	75%
[4]	{Penso}	=>	{Bisturi}	50%	75%
[5]	{Penso}	=>	{Cabo ECG}	50%	75%
[6]	{Cabo ECG}	=>	{Penso}	50%	75%

Tabela 5 - Derivação de Regras de Associação

As duas medidas abordadas, suporte e confiança, têm um papel fundamental na definição das Regras de Associação geradas pelo *Apriori*. O problema da exploração das Regras de Associação, de acordo com este algoritmo, pode ser desagrupado em duas partes em que se entende a fulcral importância destes conceitos:

• O suporte para um conjunto de itens é o número das transações que contém este conjunto. Pretende-se encontrar todos os conjuntos de itens que possuam um suporte de transações igual ou superior ao limite mínimo fixado. São assim encontrados os *itemsets* frequentes.

• A partir dos conjuntos de *itemsets* frequentes são geradas as regras de associação. São selecionadas apenas as regras que possuam o grau de confiança mínimo, correspondente à confiança mínima definida.

Assim, dado um conjunto de transações, o problema de descobrir regras de associação está em gerar todas as regras que contenham o suporte e confiança iguais ou maiores do que os valores mínimos determinados pelo usuário, designados como suporte mínimo e confiança mínima, respetivamente.

2.3.1.2 *Software*

Existem vários *softwares* que permitem o estudo das regras de associação com itens frequentes. No presente trabalho este primeiro objetivo – deteção de regras de associação com itens frequentes – será desenvolvido com recurso à utilização do *software R*.

O *R* possui uma extensão, o *arules*, que fornece a infraestrutura necessária para tratar e analisar conjuntos de dados, com a finalidade de encontrar padrões frequentes e facilitar o estudo das regras de associação (Hahsler, Grün, Hornik, & Buchta, 2005).

As transações da base de dados contém um ID de transação e um conjunto de itens. Uma regra num conjunto de regras de associação contém dois conjuntos de itens, um para o LHS (o antecedente da regra) e outro para o RHS (o consequente da regra).

Os conjuntos de itens usados para base de dados de transações e conjuntos de associações podem ser representadas como matrizes de incidência binária com colunas que correspondem aos itens e linhas que correspondem aos conjuntos de itens. A matriz é constituída pelas entradas que representam a presença (1) ou ausência (0) de um item num conjunto de itens específico.

A principal aplicação das regras de associação é a análise da *Market Basket* onde grandes conjuntos de dados de transações são extraídos. Nesta análise, cada transação contém os itens que foram comprados em cada compra de uma loja de retalho (Berry & Linoff, 1997). Os dados das transação normalmente são apresentados no formato:

<*ID* Transação, Item_1 *ID*, Item_2 *ID*, ….>

Alguns comandos úteis no R para exploração da base de dados de transações:

- *read.transactions* (): para importar dados de um arquivo. Esta função lê arquivos estruturados com o formato muito comum ilustrado em cima: uma linha por transação e os itens separados por um caracter predefinido;

- *inspect* ( ): para inspecionar as transações e visualizá-las no ecrã;

- *image* ( ): para criar uma visualização gráfica dos dados em análise;

- *length* ( ):indica o comprimento/número de linhas da base de dados;

A exploração de transação de dados no *arules* resulta em associações.

Alguns comandos no *R* para exploração de regras de associação:

- *summary* ( ):para dar uma breve visão geral do conjunto;

- *inspect* ( ): como referido anteriormente, para inspecionar transações; para exibir associações individuais;

- *length* ( ): para obter o número de elementos no conjunto;

- *items* ( ): para obter, para cada associação, o conjunto de itens envolvidos na associação;

- *sort* ( ): para classificar o conjunto usando os valores de diferentes medidas de qualidade;

- *write* ( ): para escrever associações em forma legível por humanos;

- *save* ( ) e *load* ( ): para salvar e carregar associações em forma compacta.

No *package arules* há várias funções úteis que são implementadas para contagem de suporte, indução de regras, amostragem, entre outras. Algumas dessas funções são explicadas serão descritas com maior pormenor:

Contar suporte para *itemsets*:

Normalmente, o suporte do conjunto de itens é contado durante a exploração da base de dados com uma determinada restrição de suporte mínimo.

Para base de dados com muitos itens e para valores baixos de suporte mínimo, este procedimento pode levar muito tempo a ser processado, pois o número de *itemsets* frequentes pode ser enormíssimo se definido um valor de suporte relativamente baixo.

S*upport* ( ) permite obter informação de suporte para um ou para alguns *itemsets*, sem que seja necessário extrair da base de dados o suporte de todos os *itemsets*.

A função *support* ( ) também é útil para determinar o suporte de *itemsets* infrequentes, com um suporte muito baixo.

Induzir Regras de Associação:

Uma parte do processo de exploração de regras de associação é a geração de regras a partir de conjuntos de itens frequentes usando o comando *apriori* ( ). A implementação do algoritmo *Apriori* usado no *arules* contém um mecanismo de indução de regra e por defeito retorna o conjunto de regras de associação da forma $X \rightarrow Y$ que satisfaçam um mínimo de suporte e um mínimo de confiança, previamente definidas.

Para induzir regras para um dado conjunto de conjuntos de itens, também é necessário armazenar valores de suporte numa estrutura de dados adequada que permite pesquisas rápidas para o cálculo da confiança das regras.

A geração de regras de associação é sempre separada em duas etapas principais: inicialmente extrai-se da base de dados os *itemsets* frequentes através da aplicação do suporte mínimo definido à base de dados e essa informação fica armazena numa estrutura de dados adequada. Depois disso, geram-se conjuntos de regras de associação a partir do

itemsets frequentes gerados anteriormente, aplicando-se a restrição da confiança mínima (Hahsler, Buchta, & Hornik, 2008).

Para selecionar regras interessantes a partir do conjunto de todas as regras possíveis, podem ser usadas restrições em várias medidas de significância e força. As restrições mais usualmente utilizadas são a definição de limites mínimos de suporte e confiança. Os conceitos de suporte e confiança foram anteriormente explicados, na subsecção 2.2.1, onde são introduzidos alguns conceitos.

A abordagem de limitar o suporte e confiança através da aplicação de valores mínimos para estas medidas, tem a vantagem de que se evita a explosão combinatória devido a alguns conjuntos de itens muito raros.

Gerar conjuntos de itens máximos e fechados:

O cálculo dos itens máximos e dos itens fechados permitem sintetizar a apresentação de *itemsets* de forma a evitar alguma redundância. Os conceitos de conjuntos de itens frequentes fechados e máximos foram introduzidos em detalhe na subsecção "Conceitos".

Para encontrar os *itemsets* fechados utiliza-se no R o comando *is.closed ( ).*

Para encontrar os itens máximos utiliza-se o comando: *is.maximal ( ).*

2.4 Extração de Regras com Itens Raros

No mundo, na vida, objetos e acontecimentos raros são muitas vezes de grande interesse e de grande valor. É de extrema relevância interpretá-los e compreendê-los. Tal importância transporta-se também para a análise de dados e consequente para o *Data Mining*, onde muitas vezes os objetos raros são de fundamental interesse.

São variadíssimos os exemplos que traduzem tal importância: em bases de dados médicas descobrir combinações raras de sintomas que forneçam informações úteis para os médicos; em bases de dados financeiras identificar transações financeiras fraudulentas (Chan & Stolfo, 1998) ou em certos equipamentos conseguir perceber sobre falhas ocorridas (Weiss & Hirsh, 1998).

Objetos raros são normalmente muito mais difíceis de identificar e generalizar do que objetos frequentes (Koh & Rountree, 2010), mas quando descobertos podem relevar padrões, potencialmente, interessantes. Segundo o autor, tal importância também se aplica no estudo das regras de associação.

Um dos problemas com os objetos raros é que estes podem ser muito difíceis de detetar, mesmo que a raridade dos objetos seja relativa, ou seja, os objetos não são raros num sentido absoluto, mas são raros em relação a outros objetos que existem na base de dados.

Uma das razões para a grande dificuldade em encontrar objetos raros é que estes não são facilmente localizados quando usados métodos de procura globais. As heurísticas de procura gananciosas têm um problema com casos raros pois os objetos raros podem depender da conjunção de muitas condições e, desta forma, examinar qualquer condição de forma isolada pode não fornecer muita informação. No caso de Market Basket Analysis, se quisermos identificar associação entre dois itens que raramente são comprados existem duas questões: uma é que os itens já são raramente comprados num supermercado (podemos referir-nos por exemplo à compra de uma panela e uma picadora ou de um tapete e uma toalha de banho) e depois, mesmo que os dois sejam comprados em conjunto quando um deles for comprado, pode ser muito difícil encontrar essa associação.

As regras de associação raras têm suporte baixo e confiança alta, em contraste com as regras de associação em geral com itens frequentes que são determinadas por um suporte alto e confiança também alta. A Figura 3 ilustra como a medida de suporte se relaciona com as regras de associação.

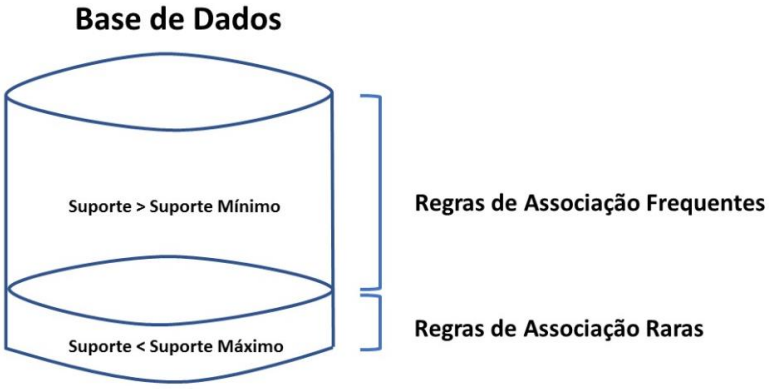

Figura 3 - Regras de Associação de uma Base de Dados

O modelo baseado nos conceitos de suporte e confiança tem tido muito sucesso ao longo dos anos. A maioria dos algoritmos usa este modelo por ser bastante simples e eficaz a extrair informação das bases de dados. No entanto, este modelo também tem arrecadado muitas críticas.

A principal limitação apontada a estes algoritmos que têm como base a geração de regras de associação baseado no suporte mínimo está relacionada com a pesquisa de itens frequentes na base de dados. Pode ser definido apenas um suporte mínimo, e caso a base de dados apresente itens da mesma natureza e com frequências idênticas tal não trará limitações à exploração das regras de associação.

Porém, caso estes dois pressupostos não se verifiquem, ao ser definido apenas um suporte mínimo pode verificar-se uma de duas situações:

- Caso o suporte mínimo seja definido com um valor demasiado elevado, o número de regras será normalmente baixo;

- Caso o suporte mínimo seja definido com um valor demasiado baixo (só assim poderiam ser detetadas associações entre a panela e a picadora ou o tapete e a toalha de banho no caso acima identificado) serão geradas demasiadas regras de associação entre produtos.

Assim, para extrair itens raros, a abordagem baseada num único suporte mínimo sofre do chamado dilema do "problema do item raro" (Mannila, 1997).

O problema do item raro é abordado por diversos autores. Para além de Mannila (1997) também Koh & Rountree (2010) o expõe:

- Se o suporte mínimo definido for demasiado elevado, o número de regras gerado será relativamente baixo e não vão ser encontradas regras de associação que envolvam os

itens raros ou infrequentes, pois sobressairiam as regras com itens frequentes. Podem, desta forma, perder-se regras de associação que poderiam ser interessantes.

- Se, por outro lado, for definido um suporte mínimo baixo, o número de regras que são geradas é bastante elevado, envolvendo itens com níveis de suporte substancialmente diferentes e que geralmente têm uma correlação fraca. Esta situação torna praticamente inexequível a análise das regras pelo utilizador e assim, a obtenção de informação relevante sobre a base de dados que possa trazer benefícios à empresa.

Outros autores como B. Liu, Hsu, & Ma (1999) ou Koh & Rountree (2005) sublinham também que o algoritmo *Apriori* dificulta a procura de regras com baixo suporte e confiança alta.

As limitações apresentadas são de fulcral importância no caso da base de dados que vai analisada, pois a mesma tem artigos de naturezas diferentes e de diferentes frequências. Por um lado existem itens que aparecem com muita frequência, os produtos clínicos, e por outro há itens que aparecem com muito menor frequência, os equipamentos médicos.

Têm sido feitos esforços por alguns autores para propor abordagens melhoradas para a deteção de associações raras (Weiss, 2004). Em 1999 B. Liu, Hsu, & Ma (1999) propõe que em vez de se fixar um único suporte mínimo para todos os itens, seja calculado um valor mínimo de suporte para cada item com base no seu suporte. Os itens frequentes são extraídos se um conjunto de itens satisfizer o menor valor de suporte mínimo dos itens nele contido. A sustentação de um artigo é a relação da frequência de um artigo no tamanho do conjunto de dados da transação.

Esta abordagem veio melhorar muito o desempenho em relação às abordagens baseadas num único suporte mínimo. No entanto, e apesar de apresentar melhorias no desempenho esta abordagem sofre ainda dos problemas expostos anteriormente como a “falta de regras" e de "explosão de regras". Se o suporte mínimo para o item for alto, os *itemsets* raros são perdidos, e se o suporte mínimo for um valor baixo, o número de *itemsets* frequentes explode.

Outras abordagens vão sendo sugeridas ao longo do tempo. Por exemplo, em 2007, Zhou & Yau apresentam uma abordagem que extrai as regras de associação considerando apenas itens infrequentes, ou seja, itens com suporte inferior ao mínimo suporte definido. Já em 2009 Kiran & Reddy sugerem no seu artigo uma abordagem

usando a conceito de "diferença de suporte" (SD) no cálculo do valor suporte mínimo para cada item. Através da noção de SD é assegurada uma diferença constante entre o suporte do item e o suporte mínimo para cada item. Deste modo, ao usar a noção de SD, a abordagem proposta extrai exaustivamente os itens frequentes envolvendo itens raros e limita a explosão de itens frequentes envolvendo itens frequentes. Os resultados experimentais em dados sintéticos e no mundo real mostraram que a abordagem proposta descobriu itens frequentes envolvendo itens raros de forma eficiente em comparação com as abordagens existentes até à data.

Koh & Rountree (2005) propõem uma abordagem diferente que, ao invés de definir apenas um suporte, utiliza dois suportes para explorar regras com suportes baixos, mas com confiança elevada. Propõem assim um algoritmo, a que chamaram *Apriori Inverse* que utiliza um limite mínimo e um limite máximo de suporte (limite inferior e superior) para gerar regras de associação ao invés de definir apenas um suporte mínimo.

Este último algoritmo será o utilizado neste trabalho para a descoberta de regras de associação com itens raros.

2.4.1 Métodos para lidar com raridade

A raridade nos dados traduz-se em problemas como a raridade absoluta ou relativa, a fragmentação dos dados ou o ruído dos dados. Weiss (2004) descreve algumas técnicas para lidar com os problemas associados a dados com raridade:

Amostragem

A amostragem é um dos métodos mais usado para lidar com a existência de casos raros. O objetivo da amostragem é eliminar ou reduzir a raridade alterando a distribuição dos exemplos de treino. Os métodos de amostragem podem ser aplicados através da subamostragem, eliminando-se os exemplos da maioria das classes ou através de sobra-mostragem, aumentando exemplos de classes minoritárias. De ambas as formas, diminui-se o nível geral de desequilíbrio de classe, tornando a classe mais rara menos rara.

Estudo apenas da Classe Rara

Ao estudarmos regras de classificação para todas as classes existentes pode originar que as classes raras sejam ignoradas. Uma alternativa é descobrir regras de classificação que preveem apenas a classe rara. Segundo Raskutti & Kowalczyk (2004) *support vector machines* utilizaram esta abordagem para estudar classes raras com algum sucesso. Existem alguns sistemas que aprendem apenas a classe rara como é o caso do sistema Brute que foi usado para procurar falhas no processo de fabricação de Boing (Riddle, Segal, & Etzioni, 1994)

Aprendizagem Sensível ao Custo

Ao introduzir no algoritmo a sensibilidade ao custo, pode explorar-se o facto de que o custo de identificar corretamente a classe rara (classe positiva) supere o custo de identificar corretamente a classe comum (classe negativa). Para situações com duas classes, associa-se um custo maior à deteção de falsos negativos do que a de falsos positivos.

Este tipo de técnica é muito utilizada nas situações de diagnósticos médicos, pois a deteção de um falso positivo leva a procedimentos médico de teste mais abrangentes e mais dispendiosos, enquanto um falso negativo pode muitas vezes colocar em causa a vida do doente.

Segmentação dos Dados

Uma técnica para lidar com os problemas associados a dados com raridade é reduzir o grau de raridade segmentando os dados. Segmentar os dados eficazmente reparte o problema original de exploração de dados em sub-problemas isolados e simplifica o problema.

Conhecimento Humano

O conhecimento que o utilizador possui sobre os dados pode melhorar bastante a forma como os dados são analisados. No que diz respeito às regras de associação, para a exploração das mesmas, o utilizador pode indicar quais os resultados que são interessantes e garantem melhores resultados e quais os que que não são tão interessantes e que não garantem tão boa exploração. Esta interação é especialmente importante quando

se pretende extrair casos raros, uma vez que o utilizador pode ter conhecimento de domínio que pode auxiliar no processo de procura. Nestes casos é mais provável que o utilizador inicie a exploração com uma pequena amostra já filtrada (Kohavi, 1998).

2.4.2 Algoritmos para Extração de Regras com Itens Raros

Apesar da relevância que os itens raros e a descoberta de conjuntos de itens raros possam apresentar em algumas situações, são menos comuns os algoritmos que os estudam que aqueles que procuram regras com itens frequentes. A literatura é ainda bastante reduzida no que diz respeito à exploração com casos raros apesar do interesse que algumas áreas podem beneficiar deste tipo de modelos.

No entanto tem-se verificado uma forte incidência no sentido de desenvolver e impulsionar estes algoritmos. Grande parte do trabalho que tem vindo a ser desenvolvido foca-se na adaptação da exploração com itens frequentes e anda muito em volta do algoritmo *Apriori* (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996).

São vários os autores, entre os quais H. Liu, Lu, Feng, & Hussain (1999) que insistem que embora os padrões fortes apresentados pelos itens frequentes sejam muito úteis na predição, os casos raros ainda que representem um número bastante pequeno de objetos podem ser também muito úteis.

Existem já alguns algoritmos para o estudo das regras de associação com itens raros ou infrequentes. O *Apriori Rare* e o *Apriori Inverse* são dois desses algoritmos que foram testados neste trabalho. No entanto o *Apriori Rare* utiliza apenas um valor de suporte mínimo enquanto o *Apriori Inverse* define dois valores e suporte, um máximo e um mínimo. Por este motivo, pareceu então mais adequado utilizar o *Apriori Inverse* para a deteção de Regra de Associação Esporádicas (com itens raros), pelo que o mesmo será explicado em pormenor de seguida.

Conforme explicado no ponto 2.3.1.1, o algoritmo *Apriori* baseia-se no princípio de que se um *itemset* é frequente então, todos os seus subconjuntos são também frequentes (Gama et al., 2012). Dualmente, e para os *itemsets* raros, um superconjunto de um conjunto de itens raros é necessariamente raro.

2.4.2.1 Algoritmo *Apriori Inverse e Sporadic Association Rules*

O algoritmo *Apriori* gera conjuntos de itens frequentes, ou seja, conjuntos de itens que produzirão regras com suporte superior a um mínimo suporte definido juntando os conjuntos de itens frequentes da passagem anterior e podando os subconjuntos que possuem um suporte menor do que o mínimo pretendido (Agrawal & Srikan, 1994). Desta forma, para gerar regras de associação que possuam baixo suporte, o suporte mínimo deve assumir um valor muito baixo, o que vai aumentar drasticamente o tempo de execução do algoritmo e vai fazer aparecer um número elevadíssimo de regras de associação. Este fenómeno é conhecido como o problema do item raro.

Ao usar-se o *Apriori* podem perder-se regras raras interessantes que desaparecem dos resultados por não atingirem o limite mínimo de frequência.

Desenvolvido por Koh & *Rountree* (2005), o algoritmo *Apriori Inverse* tem como objetivo explorar regras com suportes baixos, mas com confiança elevada. Os autores definiram estas regras como "esporádicas", pois representam casos raros espalhados de forma esporádica através da base de dados, mas com confiança elevada de ocorrer juntos. Propõem assim um algoritmo que utiliza um limite mínimo e máximo de suporte (limite inferior e superior) para gerar regras de associação ao invés de definir apenas um suporte mínimo.

O *Apriori Inverse*, tal como o *Apriori*, utiliza uma procura em largura e é executado da seguinte forma:

- No início é construído um índice invertido utilizando os itens de comprimento 1 como chave e os ID's das transações como dados: (item, [TID-list]). Nesse momento, o suporte de cada item de comprimento 1 da base de dados está disponível com o comprimento das transações (suporte).

- Na primeira passagem pelos dados é contabilizado o suporte para cada item individualmente. São então gerados os *itemsets* esporádicos e tamanho 1. Para isso, para cada artigo é verificado se o seu suporte está compreendido entre os suportes mínimos e máximos definidos. Se isso não acontecer, esse artigo é esquecido. Caso o seu suporte esteja compreendido entre os suportes definidos, esse artigo é aceite. São assim constituídos os *itemset* esporádico de tamanho 1. Posteriormente são gerados os

candidatos a *itemsets* esporádicos de tamanho 2 pela junção dos *itemsets* de tamanho 1 propostos na iteração anterior. Na mesma lógica, se o seu suporte estiver compreendido entre o suporte mínimo e máximo definidos são constituídos os *itemsets* de tamanho 2.

O algoritmo *Apriori Inverse* prossegue iterativamente, até que o conjunto-de-k-itens encontrado seja um conjunto vazio.

O algoritmo *Sporadic Association Rules* aplica primeiro o *Apriori Inverse* para gerar conjuntos de itens perfeitamente raros, que cumpram com o suporte mínimo e o suporte máximo definidos. Em seguida, usa esses conjuntos de itens para gerar as regras de associação que respeitam a confiança mínima definida.

Os autores do *Apriori Inverse* definiram o conceito de regra esporádica que pode ser dividido em dois conceitos: regra perfeitamente esporádica e regra imperfeitamente esporádica:

- Regra esporádica é uma regra onde o suporte está abaixo do limite máximo de suporte definido (*maxsup*) pelo utilizador, mas com confiança acima do nível mínimo de confiança definido (*minconf*) pelo utilizador.

- Regra perfeitamente esporádica é uma regra que não tem qualquer subconjunto acima do suporte máximo definido. O suporte da regra deve ser inferior ao *maxsup*, a confiança deverá ser igual ou superior à *minconf* e nenhum item do conjunto de $X \cup Y$ poderá ter suporte acima do suporte máximo. Assim, as regras perfeitamente esporádicas consistem em antecedentes e consequentes que ocorrem raramente mas, quando ocorrem, tendem a ocorrer em conjunto com uma confiança de pelo menos a *minconf*.

Seja *s* o máximo suporte definido e *c* a confiança mínima, uma regra $(X \rightarrow Y)$ é perfeitamente esporádica se:

$$Confiança\,(X \rightarrow Y) \geq c \text{ e}$$
$$\forall x : x \in (X \cup Y),\ \text{Suporte}\,(x) < s$$

- Regra imperfeitamente esporádica é uma regra em que, apesar de o suporte da regra ser inferior ao *maxsup* e a confiança ser igual ou superior à *minconf*, existe um item do conjunto de $X \cup Y$ que tem suporte acima do suporte máximo. Seja s o máximo suporte definido e c a confiança mínima, uma regra $(X \rightarrow Y)$ é imperfeitamente esporádica se:

$$\text{Confiança}\,(X \rightarrow Y) \geq c, \text{ e}$$

$$\text{Suporte } (X \rightarrow Y) < s, \text{ e}$$
$$\forall x : x \in (X \cup Y) \text{ , Suporte } (x) \geq s$$

De forma a permitir encontrar regras imperfeitamente esporádicas, o algoritmo permite que o suporte máximo seja aumentado ligeiramente para incluir conjuntos de itens com itens que têm suporte acima do suporte máximo.

O algoritmo tem, então, por objetivo gerar apenas regras perfeitamente esporádicas, sem ter que percorrer muitas regras que possuem suporte alto (maior que o suporte máximo definido) e, por isso, não são esporádicas. É também objetivo do *Apriori Inverse* não ter de gerar um grande número de regras triviais, por exemplo, as regras da forma $X \rightarrow Y$ onde o suporte de Y é muito alto mas o suporte de X é bastante baixo.

3. Estudo de Caso

Neste capítulo procede-se a uma apresentação do problema e à descrição dos principais conceitos utilizados bem como a metodologia aplicada. Será feita a caraterização da base de dados que será utilizada/analisada e que servirá de base ao trabalho desenvolvido. Este capítulo tem também como finalidade a avaliação dos resultados obtidos.

3.1 Identificação e Caracterização da Base de Dados

Conforme referido anteriormente, para este estudo foi utilizada uma base de dados atenciosamente cedida por uma empresa portuguesa que atua na área da saúde que comercializa equipamentos médicos e consumíveis clínicos.

O estudo começou com o levantamento, no sistema de gestão da empresa, de todas as transações efetuadas pela empresa nos últimos anos. Por forma a conciliar os dados fornecidos pela empresa e a oportunidade da informação, decidiu-se analisar o último ano de transações, o ano de 2016. A seleção do período de vendas a ser analisado é já um primeiro passo do pré-processamento. A base de dados relativa ao ano selecionado para análise contém 7240 transações e 552 diferentes artigos.

3.1.1 Itens Frequentes

3.1.1.1 Preparação dos Dados

Esta base de dados de transações, no seu estado inicial, continha todo a informação exportada para um ficheiro Excel, não estando num primeiro momento limpa de registos e informação ruidosa para a análise. Assim, numa primeira fase foi necessário proceder à limpeza e pré-seleção de dados. Esta tarefa foi efetuada em *Excel*. Esta fase de limpeza e pré-seleção de dados estava à partida prevista por se tratar das fases inicias do processo

de KDD e de extrema relevância uma vez que é a partir dos dados limpos e selecionados que será processada toda a análise e serão estudados os resultados obtidos.

Nesta fase de limpeza e pré-seleção dos dados foram realizadas tarefas como:

- Eliminação de linhas de texto sem qualquer informação adicional útil para a análise (por exemplo: número de encomenda ou proposta que lhe deu origem, número da guia de transporte, informação de contacto da pessoa que recebe a mercadoria no cliente ou informação sobre garantia dos equipamentos);

- Eliminação do cliente uma vez que esta informação não era relevante para o estudo das regras de associação;

- Eliminação da designação do artigo uma vez que esta informação não era relevante para o estudo das regras de associação;

- Em artigos compostos, eliminação do artigo composto deixando apenas informação dos componentes;

- Eliminação de referência repetida quando a mesma aparecia mais do que uma vez na mesma transação uma vez que não interessa o número de vezes que o artigo aparece numa transação. Apenas interessa o número de transações em que o artigo é comprado face ao número total de transações existentes, pois o que se pretende estudar são padrões de compra passíveis de serem adotados por um grande número de clientes;

- Eliminação de referências relativas a artigos descontinuados, dado que o estudo de artigos entretanto descontinuados não se mostra oportuna para conclusões atuais.

No final desta fase obteve-se uma base de dados apenas com o número das faturas e as referências dos produtos que as constituem. Cada transação tem um número identificativo próprio (número da fatura) e corresponde à compra de um ou mais artigos efetuada por um cliente num determinado momento no tempo.

Cada linha representa uma transação que é identificada pelo número da fatura e tem um produto associado através da referência interna que o identifica. Faturas com mais do que um produto transacionado serão apresentadas na base de dados com tantas linhas quantos os produtos que a constituem, conforme exemplificado na tabela 6.

Fatura	Referência
…	…
6320	6930003
6321	6320091
6321	6350113
6321	6320036
6321	6320006
…	…

Tabela 6 - Amostra de Base de Dados

Na amostra extraída da base de dados, podemos perceber que a fatura com o número 6320 é constituída apenas por um artigo com a referência 6930003. No entanto, a fatura com o número 6321 é constituída por quatro artigos com as referências 6320091, 6350113, 6320036 e 6320006.

Cada artigo é identificado na transação através da sua referência interna, ou seja, cada artigo tem um código que o representa e que obedece a uma estrutura definida na empresa:

- Os primeiros dois dígitos representam o fornecedor/família do artigo;

- O terceiro e quarto dígitos, em conjunto, indicam se a natureza do artigo (**10** – equipamento médico, **20** – consumível médico para equipamento, **30**, **40** e **50** – consumíveis clínicos em geral);

- Os últimos três dígitos da referência identificam o produto e são sequências consoante o fornecedor e natureza do produto.

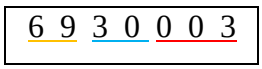

Figura 4 - Exemplo de Referência de Produto

No exemplo da figura 4 observamos a composição do artigo com código 6930003. Analisando os dígitos que o constituem sabemos que se trata de um artigo do fornecedor número 69, cuja natureza do artigo é 30 - consumível clinico em geral, e que se trata do artigo 003.

Para que o *software* utilizado, o *R*, leia corretamente os dados é necessário que os mesmos estejam adaptados para que o programa os consiga ler de forma correta. Para a

leitura dos dados, no caso das regras de associação, havia duas formas possíveis para o *software R* ler os dados:

- Sob a forma {T; x1; x2; ⋯; xn} em que T representa a transação e x os códigos dos artigos comercializados. As estruturas deverão estar separadas por ponto e vírgula sendo que não há limite de estruturas para a mesma transação.

- Sob a forma {T; x1}{T; x2} ⋯ {T; xn} em que T representa a transação e x os códigos dos artigos comercializados e em que cada conjunto se encontra numa linha.

No estudo efetuado optou-se pelo uso da segunda estrutura de dados para leitura da base de dados no *software R*, conforme exemplificado na tabela 6.

Este capítulo será sustentado por um *software* que permite realizar o estudo proposto, o *software R*. Este *software* é gratuito e permite a análise e tratamento dos dados bem como a formação dos resultados propostos através do *package arules*, conforme explicado anteriormente no ponto 2.3.1.2 Software.

3.1.1.2 Descoberta de Regras de Associação com Itens Frequentes

Numa fase inicial de análise exploratória dos dados foram obtidos alguns conhecimentos e padrões que permitiram conhecer melhor os dados.

Relativamente à totalidade das transações analisou-se de forma geral os dados.

Conforme referido anteriormente, para a descoberta de regras de associação com itens frequentes foi utilizado o *software R*. Numa fase inicial é necessário instalar o *package arules*, pois quando iniciamos o programa este *package* não faz parte das funções iniciais. Para isso basta utilizar o comando da figura 5.

```
install.packages("arules")
```

Figura 5 - Comando para instalar o package arules no R

Para ter uma primeira ideia sobre o número de *itemsets* frequentes fez-se variar o suporte e viu-se a variação do número de conjuntos de itens frequentes conforme o suporte definido. A figura 6 mostra o comando utilizado no *R* para a obtenção dos

resultados de *itemsets* frequentes e a tabela nº 7 resume os resultados obtidos do número de conjuntos de itens frequentes conforme a variação do valor mínimo de suporte:

```
itemsets <- eclat(tr, parameter = list(supp = 0.01))
```

Figura 6 - Comando no R para identificação de itemsets frequentes

Suporte	*Itemsets* Frequentes
0,001	201
0,005	54
0,01	23
0,05	3
0,06	3
0,07	2

Tabela 7 - Número de Itemsets Frequentes vs Suporte

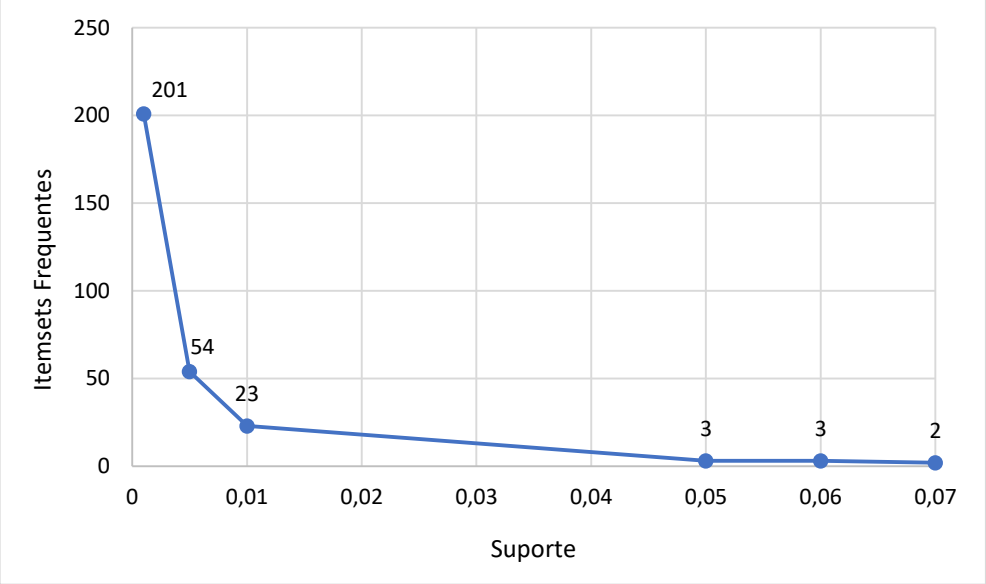

Figura 7 - Gráfico de Tamanho dos Itemsets vs Suporte

Quanto maior o valor de suporte menos *itemsets* são considerados frequentes e quanto menor o valor de suporte maior será o numero de *itemsets* descobertos. Conforme podemos verificar na tabela 7, para um suporte mínimo de 1% seriam considerados 23

itemsets frequentes enquanto para um valor de suporte mínimo de 0,1% o número de *itemsets* frequentes aumentaria para 201.

Definindo-se um valor para o suporte mínimo, todos os conjuntos de itens que não verifiquem o suporte mínimo definido não são considerados.

A tabela 9 mostra a listagem dos 15 *itemsets* mais frequentes na base de dados. Os três artigos com maior suporte, sendo assim os mais frequentes, são o 1430001-gel lubrif.esteril anestésico 11ml, o 6930001-esponja de gelatina standard e o 1430002-gel lubrif.esteril anestésico 6ml com suportes de cerca de 0.07925987, 0.07608395, 0.06835128 respetivamente. Isto significa que o item 1430001 está presente em 7,9% das transações, o item 6930001 em 7,6% e o item 1430002 em 6,8% das transações totais analisadas.

Itemsets	Descrição	Suporte
{1430001}	GEL LUBRIF.ESTERIL ANESTESICO 11ML	0,0792599
{6930001}	ESPONJA DE GELATINA STANDARD	0,0760840
{1430002}	GEL LUBRIF.ESTERIL ANESTESICO 6ML	0,0683513
{1030052}	CONTENTOR 7.5 LTS	0,0439105
{2020250}	PUNHO MONOPOLAR DISPOSABLE	0,0412869
{5240004}	MEDICAMENTO DE CONTRASTE	0,0306545
{1030040}	CONTENTOR 4 LTS	0,0283071
{1030030}	CONTENTOR 2 LTS	0,0280309
{1730057}	ESPONJA DE GELATINA STANDARD	0,0247169
{2030400}	STRIPPER DE VARIZES CONVENCIONAL DISPOSABLE	0,0230599
{2020202}	PLACA ADULTO SIMPLES	0,0222314
{1430001,1430002}	GEL LUBRIF.ESTERIL ANESTESICO 11ML, 6ML	0,0201602
{2020211}	PLACA UNIVERSAL DUPLA	0,0194698
{2030107}	LOÇÃO HIDRATANTE 1000 ML	0,0168462
{2730002}	DISPOSITIVO INTRA-UTERINO	0,0142226

Tabela 8 - Listagem dos 15 itemsets mais frequentes na Base de Dados

Na verdade, até os artigos com maiores suportes não apresentam suporte muito elevado. O suporte é relativamente baixo o que valida a grande diversidade e importância repartida dos produtos vendidos pela empresa.

É com base neste conhecimento prévio do suporte de cada item e de um melhor conhecimento da base de dados que foi possível definir melhor e ir ajustando o suporte para a descoberta de regras de associação.

A figura 8 representa o comando que permite derivar regras de associação através do algoritmo *Apriori* conforme valores de suporte e confiança definidos. No caso da figura 8 foram definidos suporte e confiança mínimos de 1%.

regras <- *Apriori* (tr, parameter = list(sup = 0.01, conf = 0.01))

Figura 8 - Comando no R para identificação de itemsets frequentes e derivação de regras de associação

A medida de avaliação *lift* foi a medida escolhida para avaliar e selecionar os padrões de associação mais fortes.

Conforme explicado no ponto 2.2.2.1 (fórmula 2.3) o *lift* é uma medida simétrica o que significa que o *lift* da regra (X → Y) apresenta o mesmo valor do *lift* da regra (Y → X).

Deste modo, e apesar de os valores de confiança das regras poderem ser diferentes optou-se por não considerar a segunda regra ao longo dos resultados que forem sendo apresentados.

Regras que apresentem *lift* um foram também desconsideradas, uma vez que o valor de *lift* um significa que os itens são independentes e o que se pretende é descobrir dependências, ou seja, associações interessantes entre os itens.

Considerando apenas as regras de associação máximas e um suporte e confiança de 0,001 foram encontradas as regras de associação que se apresentam na sua totalidade no Anexo 2 deste trabalho. A tabela 9 apresenta as mais relevantes, listadas por ordem decrescente de interesse (*lift*).

Para uma melhor observação gráfica, recorreu-se ao *add-in* Think-Cell do Microsoft PowerPoint. O gráfico da figura 9 apresenta no eixo das abscissas os antecedentes das regras e no eixo das ordenadas os consequentes. Graficamente, quanto maior o circulo, maior o *lift* da regra. Os círculos vermelhos representam as regras de associação cujo *lift* é inferior a 1 e que, significam portanto, que os itens são negativamente dependentes.

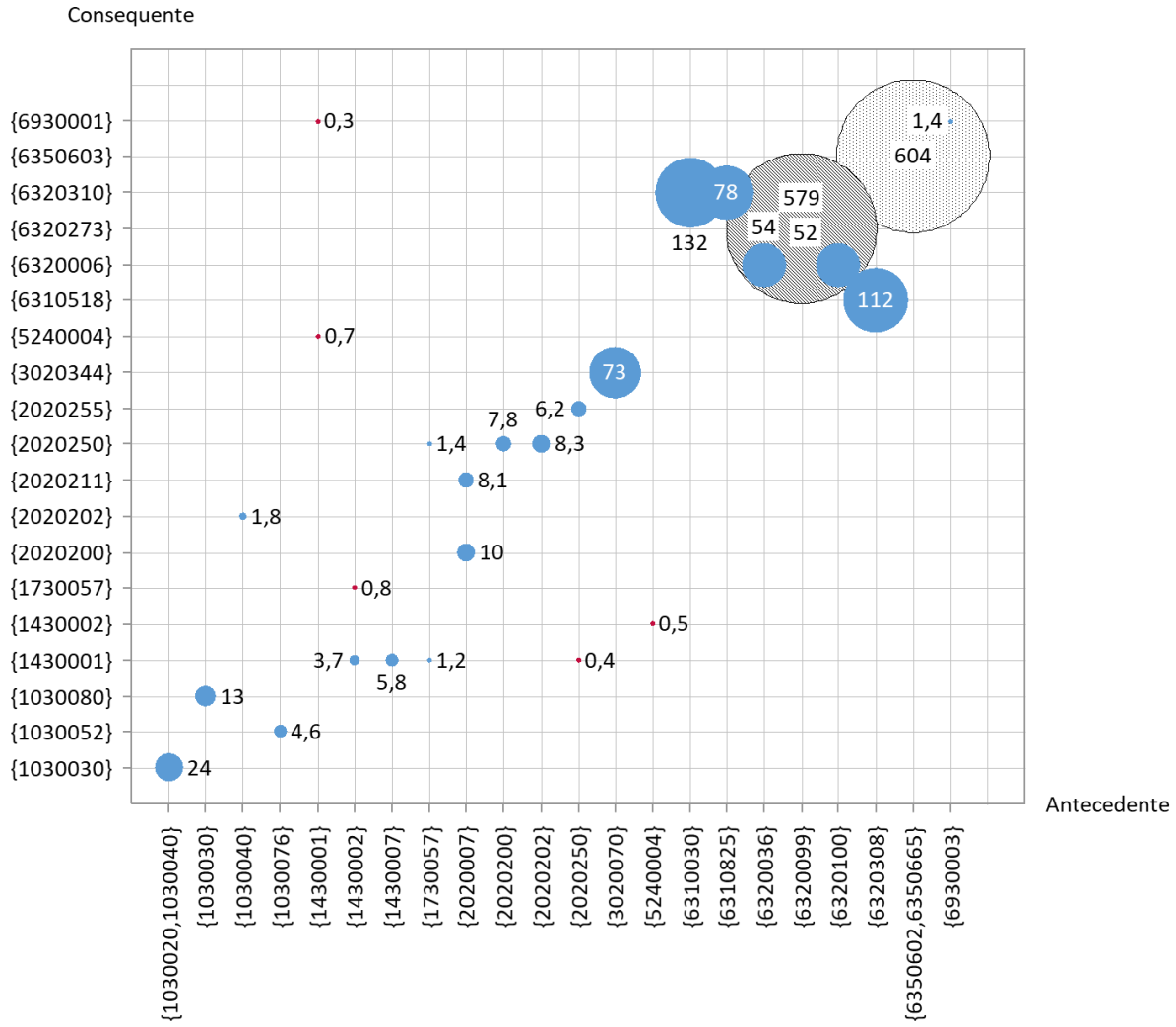

Figura 9 - Representação Gráfica das Regras de Associação com Apriori

#	Antecedente		Consequente	Suporte	Confiança	*Lift*
1	{6350602,6350665}	=>	{6350603}	0.001104667	1.000000000	603.5000000
2	{6320099}	=>	{6320273}	0.001104667	0.800000000	579.3600000
8	{6310030}	=>	{6320310}	0.001242751	0.818181818	131.6727273
9	{6320308}	=>	{6310518}	0.007042254	0.927272727	111.9218182
10	{6310825}	=>	{6320310}	0.002347418	0.377777778	78.1676190
11	{3020070}	=>	{3020344}	0.002485501	0.580645161	72.5005562
13	{6320036}	=>	{6320006}	0.001795084	0.406250000	54.4826389
14	{6320100}	=>	{6320006}	0.001242751	0.391304348	52.4782609
15	{1030020,1030040}	=>	{1030030}	0.001104667	0.666666667	23.7832512
17	{1030030}	=>	{1030080}	0.001242751	0.044334975	13.3780788
20	{2020007}	=>	{2020200}	0.001518917	0.134146341	9.5243902
21	{2020202}	=>	{2020250}	0.007594587	0.341614907	8.2741644
22	{2020007}	=>	{2020211}	0.001795084	0.158536585	8.1427089
23	{2020200}	=>	{2020250}	0.004556752	0.323529412	7.8361204
24	{2020250}	=>	{2020255}	0.001795084	0.043478261	6.1739130
25	{1430007}	=>	{1430001}	0.004832919	0.460526316	5.8103338
26	{1030076}	=>	{1030052}	0.001380834	0.200000000	4.5547170
27	{1430002}	=>	{1430001}	0.020160177	0.294949495	3.7212966
29	{1030040}	=>	{2020202}	0.001104667	0.039024390	1.7553704
30	{6930003}	=>	{6930001}	0.001104667	0.108108108	1.4209055
31	{1730057}	=>	{2020250}	0.001380834	0.055865922	1.3531137
32	{1730057}	=>	{1430001}	0.002347418	0.094972067	1.1982364
33	{1430002}	=>	{1730057}	0.001380834	0.020202020	0.8173354
34	{1430001}	=>	{5240004}	0.001657001	0.020905923	0.6819851
35	{5240004}	=>	{1430002}	0.001104667	0.036036036	0.5272181
36	{2020250}	=>	{1430001}	0.001242751	0.030100334	0.3797676
37	{1430001}	=>	{6930001}	0.001657001	0.020905923	0.2747744

Tabela 9 - Resumo das Regras de Associação descobertas com Apriori

Da análise das regras que se encontram na tabela 9, podemos retirar algumas conclusões relativamente às regras de associação descobertas:

- Regra geral, as associações encontradas são entre artigos da mesma família/fornecedor;

- Cinco das regras descobertas têm *lift* inferior a 1 (regras # 33 a #37) o que significa que apenas as últimas 5 regras da tabela apresentam itens negativamente dependentes. Itens negativamente dependentes não são interessantes como regras de

associação pois tal significa que, nestas regras, o consequente da regra é menos provável estar presente com o antecedente da regra do que a frequência base do consequente na base de dados. Por este motivo apresentam-se como menos destaque na tabela.

- A regra #1 {6350602,6350665} => {6350603} é a que apresenta maior *lift* (603,5) e tem confiança 1. Isto significa que o antecedente da regra nunca aparece sem o consequente. Assim, é de esperar, com 100% de confiança, que todas as transações que contêm os itens 6350602 - caixa frontal e 6350665 - rear panel assembly também contenham o item 6350603 - label p/ caixa frontal.

O algoritmo *Apriori* baseia-se no princípio de que se um *itemset* é frequente então, todos os seus subconjuntos são também frequentes. Assim, seria de esperar à partida também a descoberta das regras #3, #5 e #7, de tamanho dois, apresentadas na listagem completa do Anexo 1:
{6350602} => {6350603}; {6350665} => {6350603}; {6350665} => {6350602}

- O mesmo sucede com a regra #15 {1030020,1030040} => {1030030} que aparece com uma confiança aproximada de 67%. Segundo esta regra é de esperar que em 67% das transações em que são comprados os itens 1030020 - Contentor 1 lts e 1030040 - Contentor 4 lts seja comprado o artigo 1030030 - Contentor 2 lts.

Considerando a referida regra máxima de tamanho 3, seria de esperar à partida a descoberta das regras de tamanho 2, regras #18, #19 e #28, com os seus subconjuntos também frequentes:
{1030020} => {1030030}; {1030020} => {1030040}; {1030030} => {1030040}

Efetivamente esta regra de associação faz todo o sentido pois estes três artigos encontram-se no catálogo nacional da saúde, onde a empresa alvo do estudo está representada. Adicionalmente tratam-se de produtos que, quando introduzidos num cliente, facilmente a sua compra se torna fiel e se estende a toda a gama do produto, o que se verifica principalmente nos produtos de uso mais intensivo como é o caso das medidas referidas nesta regra.

- A regra #2 é a segunda regra com maior *lift* {6320099} => {6320273} e tem confiança 0,8 o que indica que é de esperar que em 80% das transações o item 6320099 -

sensor spo2 neonatal implique a compra do item 6320273 - tubo nibp neonatal. Tanto o antecedente como o consequente desta regra são consumíveis do mesmo tipo de equipamento que se destinam a pacientes neonatais.

- A regra #32 {1730057} => {1430001} apresenta um *lift* de 1.1982364 o que significa que há 19,8% mais hipóteses de ocorrência de 1430001-gel lubrif.esteril anestésico 11ml, dado que 1730057-esponja de gelatina standard também está na transação. Ou, por outras palavras, a probabilidade de encontrar 1430001 em todas as transações que possuem o produto 1730057 é 19,8% a mais do que a probabilidade de encontrar o produto 1430001 em todas as transações. Os itens desta regra são dois produtos de farmácia que a empresa vende em grande quantidade.

As regras analisadas foram avaliadas segundo critérios objetivos.

No entanto, conforme visto no ponto 2.2.2.2, as regras de associação podem ser medidas por indicadores subjetivos: o interesse e a previsibilidade. Assim, pensando-se na previsibilidade das regras a descobrir, era esperado à partida que aparecessem regras com os componentes dos *kits* (artigos que a empresa vende em pacote):

- Regra #11 {3020070} => {3020344}, Componentes do *kit* 3020114 - *kit* p/argon (punho+placa)
- Regra #21 {2020202} => {2020250} Componentes do *kit* 2020261 - *kit* punho+placa simples
- Regra #23 {2020200} => {2020250} Componentes do *kit* 2020263 - *kit* punho+placa dupla
- Regra #24 {2020250} => {2020255} Componentes do *kit* 2020259 - *kit* punho+lixa

O item 2020250 - punho monopolar descartável, aparece em várias regras de associação encontradas. Trata-se também de um item com forte participação nas vendas quando vendido individualmente. É o quinto artigo com maior suporte (aproximadamente 4,13%) o que significa que é o quinto artigo que mais aparece no total das transações.

Para além de se encontrar nas regras de associação que relacionam os componentes dos *kits* que ele integra, este artigo aparece também na regra # 31 como consequente da regra {1730057} => {2020250}. Nesta regra aparece com o antecedente 1730057 - esponja de gelatina standard, no entanto com menor confiança e *lift* que as regras que integra.

Dos *kits* disponíveis para venda, todos aparecem nas regras de associação da tabela 9, para um suporte e confiança mínimos definidos de 0,1% o que seria já esperado dado que, ao serem vendidos em *kit*, já estão, por si só, interligados. Não obstante serem esperadas, as regras que os incluem vêm-se apresentar como bastante úteis para reforçarem e validarem o conhecimento existente à priori sobre os dados.

O artigo 2020202 - placa simples para além de se encontrar em regras de associação que relacionam os componentes dos *kits*, também se encontra na regra #29 {1030040} => {2020202} com o antecedente 1030040 - Contentor 4 lts, no entanto também com confiança e *lift* muito menores que a regra em que participa e que diz respeito ao *kit* do qual é componente.

- A regra # 27 {1430002} => {1430001} poderia ser uma regra também esperada à partida. Os itens 1430002-gel lubrif.esteril anestésico 6ml e 1430001-gel lubrif.esteril anestésico 11ml referem-se ao mesmo produto. Tratam-se ambos de lubrificantes, apenas diferindo na capacidade sendo um de 6ml e outro de 11ml. Funcionam como lubrificante para algaliação, sendo o primeiro mais direcionado a mulheres e o segundo a homens, respetivamente.

- A regra #20 {2020007} => {2020200} tinha também previamente uma forte previsibilidade de ocorrência pois o cabo é indispensável para o funcionamento das placas e também porque uma das medidas comerciais adotadas na venda destes dois artigos é de que na compra de determinada quantidade mínima da referencia 2020200-placa ad. dupla disposable, é vendida a referência 2020007-cabo p/ placas a um preço mais baixo que o valor base de venda.

- As regras #8 {6310030} => {6320310}, #9 {6320308} => {6310518} e #10 {6310825} => {6320310}, não seriam esperadas à partida na descoberta de associações

com itens frequentes, uma vez que se trata de equipamentos e seus acessórios e, supostamente, os mesmos se encontram em menor número que os consumíveis médicos.

Apesar disso, são três regras que estão no top de 10 de regras com maior *lift*, para os valores definidos de suporte e confiança.

O artigo 6320310-suporte rodado p/imec aparece em duas destas regras como consequente o que significa que é esperado que um cliente que compre monitor paciente portátil com ou sem touch (6310030 e 6310825) compre também o suporte rodado, com uma confiança de aproximadamente 82% e 38%, respetivamente.

- As regras #13 {6320036} => {6320006} e #14 {6320100} => {6320006} relacionam acessórios de equipamentos. O consequente em ambas é o 6320006-sensor dedo spo2 reut. adulto, sendo de esperar que quem compra 6320036-extensão spo2 e 6320100-extensão spo2 p/beneview compre também o sensor de dedo, em cerca de 40% das transações. Tanto os antecedentes como os consequentes destas regras são acessórios de spo2 que se destinam a ler o mesmo parâmetro (spo2) pelo que a sua relação seria provável.

Para além dos pontos analisados é de mencionar que as regras de associação com maiores *lifts* pertencem a artigos da mesma família/fornecedor. De relevar também que os resultados obtidos recaem sobre meia dúzia de fornecedores principais. Maioritariamente, os produtos cujas relações são mais fortes entre si pertencem aos fornecedores números 10, 14, 17, 20, 30, 63 (primeiros dois dígitos das referências dos artigos).

3.1.2 Itens Raros

3.1.2.1 Preparação dos Dados

Tem havido também forte investigação no que diz respeito a algoritmos que extraem regras de associação com itens raros. No entanto, ainda não é tão frequente encontrar estes algoritmos implementados à partida nos principais *softwares*.

Para a descoberta de regras de associação com itens raros foi utilizado o *software* SPMF que é uma biblioteca aberta de *Data Mining* em linguagem *Java.*

O SPMF pode ser usado através de uma interface de usuário simples e oferece implementações de vários algoritmos de *Data Mining.* Para este ponto foi utilizado o algoritmo *Apriori Inverse* explicado em detalhe no ponto 2.4.2.1.

O *software* SPMF utiliza ficheiros de texto como entrada dos dados. Assim, para a sua utilização foi necessário converter a base de dados de transações no formato que fosse legível pelo *software*.

Muito do trabalho de limpeza e preparação de dados já tinha sido efetuado no ponto 3.1.1.1 de preparação dos dados para a descoberta de regras com itens frequentes. Nesta fase foi, então, necessário adaptar a base de dados ao formato exigido pelo *software* SPMF. Com essa finalidade transformaram-se os dados de forma que cada linha correspondesse a uma transação, que em cada transação os itens aparecessem separados por um único espaço e que os códigos dos itens dentro de uma mesma transação (linha) estivessem ordenados por ordem crescente. É requisito fundamental que não existam repetições de artigos em cada transação. Esta tarefa foi efetuada em Excel.

3.1.2.2 Descoberta de Regras de Associação com Itens Raros

A venda de equipamentos médicos detém uma fatia muito significativa nas vendas. De facto, não obstante não serem artigos frequentes, é de extrema relevância considerá-los e estudá-los pelo seu significante contributo nas vendas.

Assim, numa fase inicial de análise exploratória dos dados foi utilizado o *Apriori-Inverse* para obter algum conhecimento e padrões que permitiram conhecer melhor os dados e na tentativa de obter alguma informação mais profunda sobre os itens mais raros – os equipamentos médicos.

O Anexo 3 do presente trabalho mostra um *print* do *interface* do SPMF para a aplicação do *Apriori-Inverse* onde são definidos os ficheiros de *input* - base de dados das transações, ficheiro de *output* - resultados, e os valores de suportes mínimos e máximos

pretendidos. O *output* de resultados do *Apriori-Inverse* é um ficheiro de texto onde cada linha representa um conjunto de itens perfeitamente raros.

O Anexo 4 mostra um print da tabela resumo de resultados do interface.

O Anexo 5 apresenta a tabela original de resultados para um suporte mínimo de 0.0002 e um suporte máximo de 0.001. Os itens máximos da tabela original - Anexo 5, encontram-se resumidos na tabela 10.

Itemsets
3020050 3020108 #SUP: 4
3020421 3020423 #SUP: 2
4030066 4030097 #SUP: 2
4030112 4030113 #SUP: 2
4030152 4030153 #SUP: 2
4530061 4530080 #SUP: 2
6310048 6320324 #SUP: 5
6310067 6320281 #SUP: 5
6310150 6310507 #SUP: 2
6320027 6320044 #SUP: 2
6310309 6320146 6320263 #SUP: 3
6350686 6350712 6350857 #SUP: 2
6310720 6320400 6320401 6320638 6320646 #SUP: 2

Tabela 10 - Itemsets Apriori Inverse

Apesar de aparecerem alguns consumíveis como *itemsets* com suportes definidos entre 0.02% e 0.01% (4030066,….,4530080) surgem também os primeiros equipamentos. Pela primeira vez temos dois equipamentos relacionados 6310150 e 6310507, apesar de um suporte baixo de 2, o que significa que estes dois artigos aparecem relacionados em apenas duas transações. No entanto, estes dois artigos tratam-se de Unidade de Telemetria e de Transmitter p/Telemetria, dois artigos que, mesmo sendo definidos como equipamentos, estão relacionados à partida, pois um é complemento do outro.

Na verdade, uma das principais limitações do estudo das regras de associação na base de dados em questão é a própria base de dados. É que os equipamentos aparecem, por regra, em transações isoladas, isto é, os equipamentos são quase sempre faturados separadamente. Tal acontece por várias razões: a maioria dos clientes são entidades públicas que abrem concurso público e cabimento específicos para determinados equipamento em separado. E para cada cabimento deve corresponder apenas uma fatura.

Muitas vezes também acontece que um mesmo cliente tem locais de entrega diferentes e por esse motivo é emitida uma fatura por cada local de entrega.

Por este motivo, compactou-se as transações por dia e por cliente e fez-se nova análise dos resultados encontrados.

Com este objetivo foi necessário preparar os dados de forma a juntar as transações de modo a que estas aparecessem juntas por dia e por cliente. Isto significa que se tiverem sido emitidas duas faturas a um cliente, no mesmo dia, deixaremos de ter duas transações e passaremos a ter apenas uma que engloba todos os artigos adquiridos por esse cliente, nesse dia. Mais uma vez relembra-se que não foram considerados artigos descontinuados por não ser oportuno o seu estudo e não foram considerados artigos repetidos numa transação agregada pois, uma vez mais, não interessa o número de vezes que o artigo aparece numa transação. Interessa apenas o número de transações em que o artigo é comprado face ao número total de transações existentes, pois o que se pretende estudar são padrões de compra passíveis de serem adotados por um grande número de clientes.

A base de dados que inicialmente era composta por 7240 transações passou, depois de as transações terem sido agrupadas por dia e por cliente a ter um total de 5370 transações.

A tabela 11 resume as regras de associação encontradas após a conversão da base de dados para valores definidos de suporte mínimo de 0,1%, suporte máximo de 1% e confiança de 0,1%.

O valor do *lift* não aparece calculado por defeito no SPMF, todavia, pareceu importante proceder ao seu cálculo separadamente e apresentá-lo nas tabelas de resultados para uma maior homogeneização da apresentação de resultados e uma melhor interpretação da informação:

Antecedente		Consequente	Suporte Absoluto	Confiança	*Lift*
{6310067}	=>	{6320281}	6	1.000000000	767.1428571
{6350603}	=>	{6350602}	6	1.000000000	767.1428571
{6320090; 6320273}	=>	{6320099}	6	1.000000000	537.0000000
{6320273}	=>	{6320091}	6	0.600000000	268.5000000
{4030028}	=>	{4030027}	7	0.875000000	117.4687500
{6320253}	=>	{6320021}	6	0.375000000	111.8750000
{6320308}	=>	{6310518}	40	0.930232558	111.0077519
{6310030}	=>	{6320310}	9	0.818181818	107.1618625
{4030014}	=>	{4030004}	7	0.538461538	67.24508044
{6310318}	=>	{6310825}	6	0.333333333	61.72413786
{6310825}	=>	{6320310}	13	0.448275862	58.71320436
{3020070}	=>	{3020344}	17	0.566666666	55.32727266
{6320090}	=>	{6320308}	6	0.400000000	49.95348837
{6320100}	=>	{6320006}	10	0.454545454	49.81447118
{6320036}	=>	{6320006}	14	0.437500000	47.94642857
{6320253}	=>	{6320006}	6	0.375000000	41.09693877
{4030050}	=>	{4030004}	7	0.225806451	28.19954981
{4030004}	=>	{4030003}	7	0.162790697	16.81127005
{4030213}	=>	{4030003}	6	0.162162162	16.74636172

Tabela 11 – Resumo das Regras de Associação descobertas com Apriori-Inverse

Nos resultados obtidos salienta-se a forte participação de associações com acessórios de equipamentos, (artigos cujas referências têm nos terceiro e quarto dígitos 20). De facto são vários os exemplos em que podemos ver os equipamentos associados aos seus acessórios: {6310067} => {6320281}; {6310030} => {6320310} o que seria bastante expectável acontecer dado os acessórios serem componentes dos equipamentos, aparecendo nas transações sempre como complementos dos equipamentos. Para além disso, verifica-se também que os acessórios de equipamentos estão fortemente associados entre si: {3020070} => {3020344}; {6320090} => {6320308}; {6320100} => {6320006}; {6320036} => {6320006}; {6320253} => {6320006} o que também seria esperado dado que se os acessórios aparecem associados aos respetivos equipamentos, então os acessórios dos equipamentos entre si também se relacionam.

Por este motivo decidiu-se, como último passo no estudo das regras de associação com itens raros, experimentar analisar as regras sem que os acessórios constassem na base de dados, ou seja, eliminando-os da base de dados e deixando apenas as referências de

consumíveis e equipamentos para melhor tentar perceber o que aconteceria relativamente à descoberta de relação entre eles.

Em termos práticos, foram retirados da base de dados todos os artigos cujos terceiro e quarto dígitos eram 20, ficando assim a base de dados composta apenas por artigos cujos terceiro e quarto dígitos são 10 (equipamentos médicos), 30, 40 ou 50 (consumíveis clínicos). A base de dados passou, assim, a ser constituída por 5189 transações.

Após a referida alteração à base de dados foram obtidos os resultados das regras de associação que se encontram listadas na sua totalidade na tabela do Anexo 6 para valores definidos de suporte mínimo de 0,05%, suporte máximo de 3% e confiança de 10%.

As regras mais relevantes do Anexo 6 encontram-se resumidos na tabela 12.

Uma vez mais, dado que o *lift* não aparece nos resultados das regras de associação do SPMF, procedeu-se ao seu cálculo separadamente e apresentou-se os valores na tabela de resultados para uma melhor conclusão dos mesmos.

Para uma melhor visualização, apresenta-se o gráfico da figura 10 elaborado com recurso ao add-in Think-Cell do Microsoft PowerPoint. No eixo das abscissas apresentam-se os antecedentes das regras e no eixo das ordenadas os respetivos consequentes. Graficamente, quanto maior o círculo, maior o *lift* da regra:

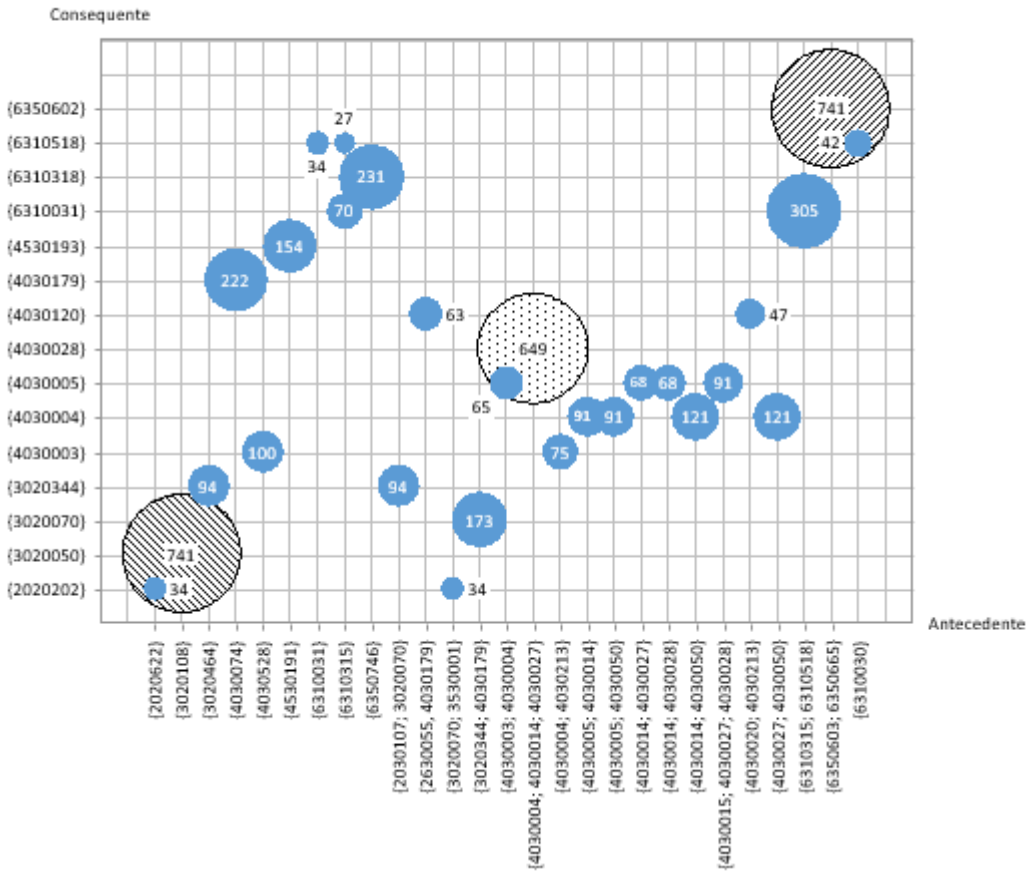

Figura 10 - Representação Gráfica das Regras de Associação com Apriori-Inverse

Efetivamente, após a alteração da base de dados em que se excluiu os acessórios das transações, surgem então associações entre equipamentos, objetivo primordial do estudo das regras de associação com itens raros. As mesmas estão sinalizadas a negrito na tabela 12.

É um facto que o suporte absoluto das regras descobertas é bastante baixo, variando entre 3 e 5. Realmente, o que se procura obter nesta fase do trabalho são regras de associação com itens raros, pelo que não é de esperar valores de suporte altos.

Ainda assim, as regras apresentam confiança e *lift* bastante significativos, havendo mesmo uma regra com confiança de 1: {6310315; 6310518} => {6310031}.

A regra com maior *lift* apresenta um valor para esta medida de cerca de 741, sendo que a regra que relaciona equipamentos com maior *lift*, indica um valor de aproximadamente 305 o que significa que a venda dos equipamentos 6310315-monitor

desfibrilhador beneheart d3 e 6310518-monitor sinais vitais potencia a venda do equipamento 6310031-monitor paciente portátil em cerca de 300%.

Antecedente		Consequente	Suporte Absoluto	Confiança	*Lift*
{3020108}	=>	{3020050}	4	1.0000000	741.285714
{6350603; 6350665}	=>	{6350602}	5	1.0000000	741.285714
{4030004; 4030014; 4030027}	=>	{4030028}	3	1.0000000	648.625000
{6310315; 6310518}	**=>**	**{6310031}**	**3**	**1.0000000**	**305.235294**
{6350746}	=>	{6310318}	4	0.8000000	230.622222
{4030074}	=>	{4030179}	3	0.6000000	222.385714
{3020344; 4030179}	=>	{3020070}	3	1.0000000	172.966666
{4530191}	=>	{4530193}	5	0.6250000	154.434523
{4030014; 4030050}	=>	{4030004}	4	1.0000000	120.674418
{4030027; 4030050}	=>	{4030004}	3	1.0000000	120.674418
{4030528}	=>	{4030003}	3	1.0000000	99.7884615
{3020464}	=>	{3020344}	3	1.0000000	94.3454545
{2030107; 3020070}	=>	{3020344}	3	1.0000000	94.3454545
{4030015; 4030027; 4030028}	=>	{4030005}	4	1.0000000	91.0350877
{4030005; 4030014}	=>	{4030004}	3	0.7500000	90.5058139
{4030005; 4030050}	=>	{4030004}	3	0.7500000	90.5058139
{4030004; 4030213}	=>	{4030003}	3	0.7500000	74.8413461
{6310315}	**=>**	**{6310031}**	**3**	**0.2307692**	**70.4389140**
{4030014; 4030027}	=>	{4030005}	3	0.7500000	68.2763157
{4030014; 4030028}	=>	{4030005}	3	0.7500000	68.2763157
{4030003; 4030004}	=>	{4030005}	5	0.7142857	65.0250626
{2630055, 4030179}	=>	{4030120}	3	1.0000000	62.5180722
{4030020; 4030213}	=>	{4030120}	3	0.7500000	46.8885542
{2020622}	=>	{2020202}	3	1.0000000	33.9150326
{6310030}	**=>**	**{6310518}**	**4**	**0.3636363**	**41.9313131**
{6310031}	**=>**	**{6310518}**	**5**	**0.2941176**	**33.9150326**
{3020070; 3530001}	=>	{2020202}	3	1.0000000	33.9150326
{6310315}	**=>**	**{6310518}**	**3**	**0.2307692**	**26.6102564**

Tabela 12 - Resumo das Regras de Associação descobertas com Apriori-Inverse após alteração da Base de Dados (sem acessórios de equipamentos)

O algoritmo *Apriori* baseia-se no princípio de que se um *itemset* é frequente então, todos os seus subconjuntos são também frequentes. Dualmente, para os *itemsets* raros, um superconjunto de um conjunto de itens raros é necessariamente raro.

- As regras {6310315; 6310518} => {6310031} e {6310315} => {6310031} revelam que os artigos 6310315-monitor desfibrilhador beneheart d3 e 6310518-monitor sinais vitais potenciam a venda do equipamento 6310031-monitor multi-parâmetros tamanho ecrã 1.

- As regras {6310031} => {6310518} e {6310315} => {6310518} relaciona ainda os equipamentos mencionados na regra anterior, alternando apenas o antecedente e consequente da regra.

- A regra {6310030} => {6310518} associa 6310030-monitor multi-parâmetros tamanho ecrã 2 com 6310518-monitor sinais vitais.

Curiosamente a maioria das transações que relacionam os equipamentos referidos ocorrem para clientes privados. Apesar de ser difícil tirar conclusões generalistas com base no estudo das regras de associação com itens raros, as regras referidas podem ser um indício de que, havendo mais vendas de equipamentos para o sector privado, as mesmas podem vir a reforçar as regras de associações encontradas. Na verdade a análise deste resultado é bastante importante uma vez que as vendas para o sector privado significa ainda uma pequena fatia das vendas.

De realçar também que os resultados de regras de associação com itens raros recaem apenas sobre o fornecedor com o número 63 (os dois primeiros dígitos da referência do artigo ditam a família a que pertencem). Trata-se do principal fornecedor de equipamentos médicos.

4. Conclusões

As técnicas de *Data Mining* permitem, cada vez mais, explorar padrões e retirar conhecimento para a gestão dos negócios e tomada de decisão para que haja melhoria contínua nos processos. Em particular, a *Market Basket Analysis* revela-se uma área de grande relevância no estudo das transações e tem como grande finalidade encontrar relações de dependência entre artigos vendidos e descrever o comportamento do consumo dos clientes retirando daí informação sobre padrões de compra de forma a ganhar vantagem competitiva. Nesta dissertação, o principal objetivo foi estudar esses relacionamentos numa empresa de distribuição por grosso de equipamentos médicos e de consumíveis clínicos.

Uma vez que os consumíveis clínicos são artigos que aparecem com muita frequência nas transações e os equipamentos médicos aparecem em muito menor quantidade, a abordagem seguida baseou-se na exploração de regras de associação com itens frequentes, mas também na descoberta de regras com itens raros, apesar de ser ainda uma área menos estudada, por se acreditar ser uma mais-valia para o negócio. Assim, este estudo utilizou como principal *software* o *R*, com recurso ao algoritmo *Apriori* para o estudo de regras de associação com itens frequentes, e o SPMF recorrendo ao algoritmo *Apriori-Inverse*, para o estudo de regras de associação com itens infrequentes.

De um modo geral, chegou-se a uma conclusão sobre os artigos mais comercializados: gel lubrificante estéril anestésico 11ml, esponja de gelatina standard, gel lubrificante estéril anestésico 6ml, contentor 7.5 lts, punho monopolar disposable, medicamento de contraste, contentor 4 lts, contentor 2 lts, esponja de gelatina standard, stripper de varizes convencional disposable, placa adulto simples. Apesar de serem os mais frequentes, estes produtos apresentam um suporte relativamente baixo, cerca de 2%, o que vem reforçar a forte aposta da empresa na diversidade de produtos.

No que diz respeito às regras de associação com itens frequentes, algumas das que apresentam maiores níveis de confiança e maiores lifts são:

- Com 100% de confiança e com o maior *lift*, as transações que contêm os itens 6350602 - caixa frontal e 6350665 - rear panel assembly também contêm o item 6350603 - label p/ caixa frontal.

- A compra de sensor spo2 neonatal implica a compra do item 6320273 - tubo nibp neonatal, por se tratarem de consumíveis do mesmo tipo de equipamento que se destinam a pacientes neonatais.

- A relação entre a venda de contentor 1 lts e contentor 4 lts influencia positivamente a venda do contentor 2 lts em cerca de 67% das transações.

As regras de associação que incluem itens raros destacam-se em artigos que pertencem ao mesmo fornecedor e referem quatro produtos principais: monitor desfibrilhador beneheart d3, monitor sinais vitais e monitores multi-parâmetros de tamanho de ecrã 1 e 2. Estas relações referem-se maioritariamente a vendas para clientes privados, que representam uma percentagem muito pequena das vendas totais, o que revela um potencial elevado destes produtos no sector e que pode resultar num crescimento relevante em vendas através de uma melhor promoção comercial no terreno.

Os resultados encontrados serão de grande utilidade para a empresa em domínios de marketing e logística, nomeadamente para:

- Influenciar a visibilidade dos artigos e grupos de artigos mais vendidos no catálogo de produtos – uma estratégia que poderia ser utilizada seria colocar juntos artigos bastante relacionados e, afastar conjuntos de itens que pertencem a regras com elevado nível de confiança, obrigado assim a que o cliente percorra mais páginas do catálogo e que veja mais produtos.

- Promover produtos pertencentes a conjuntos de produtos em ações de cross-selling e promoções. Por exemplo, na regra {contentor 1 lts} → {contentor 2 lts} em que a confiança é bastante elevada, verificando-se que o antecedente da regra atinge as vendas esperadas e o consequente não, poderá ser proposto algum tipo de mecanismo em que, na compra do antecedente se oferece uma promoção no consequente da regra, potenciando assim a venda deste último.

- Reorganizar a disposição dos materiais em armazém de forma a aumentar a eficiência logística no *picking* dos produtos.

- Sensibilizar a equipa comercial, principal meio de comunicação comercial e uma das principais fontes de conhecimento dos produtos nos clientes. Estando a equipa comercial mais conhecedora das regras de associação mais fortes, poder-se-á ser mais bem-sucedido nas recomendações ao cliente o que significará um passo mais para o fortalecimento das vendas. Não há dúvidas de que um dos grandes benefícios das regras

de associação é antecipar as preferências de um cliente com base nas preferências/histórico de um grupo de clientes.

4.1 Limitações

Uma das limitações a este estudo está relacionada com a base de dados a analisar que apresenta algumas barreiras no que diz respeito à formação das transações e separação dos equipamentos e consumíveis nas mesmas. Na verdade, uma das principais restrições do caso em estudo prendeu-se com próprias transações que englobam os equipamentos médicos, os considerados itens raros, pois estes apareciam, geralmente, em transações isoladas dos restantes artigos. Isso obrigou a que se compactassem as transações por dia e por cliente de forma a contornar o problema.

Outro aspeto que deve ser apontado neste ponto prende-se com o número de transações alvo do estudo. Definiu-se como período de estudo apenas o ano de 2016, com 7240 transações, pelo que o número de transações existentes para análise pode ser considerado baixo e impeditivo de alcançar conclusões mais significativas.

É também importante referir que apenas foi utilizada uma medida de avaliação objetiva das regras de associação: o *lift*, o que poderá apresentar-se também como uma limitação.

4.2 Trabalho Futuro

Um dos trabalhos futuros passaria por tentar perceber sequências de associações com o objetivo de compreender, por exemplo, se o início a venda de um artigo para um cliente poderá implicar o início da venda de um outro artigo.

Um outro aspeto que seria interessante analisar seria considerar o estudo em distintos períodos ano e tentar perceber se existem períodos, estações ou eventos que influenciem os resultados das regras encontradas. Desta forma, considerando diferentes períodos para análise, o estudo conduziria a resultados sobre a sazonalidade das vendas, o que poderia ser interessante para a gestão de *stocks* da empresa. O próprio estudo da

frequência dos artigos através do seu suporte, permite também o conhecimento de artigos com alta e baixa rotação permitindo identificar os artigos que em que se deve ou não deve apostar em abastecimento, rentabilizando o espaço em armazém.

Referências Bibliográficas

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, *22*(2), 207–216.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). *Fast Discovery of Association Rules*. (MIT Press, Ed.). California, USA: American Association for Artificial Intelligence Menlo Park.

Agrawal, R., & Srikan, R. (1994). Fast Algorithms for Mining Association Rules. *Proc. 20th Int. Conf. Very Large Data Bases,VLDB*, *1215*, 487–499.

Bastos, G. M. (2001). Algumas Aplicações Práticas da Tecnologia Data Mining. *Sebrae*. Obtido de http://livrozilla.com/doc/1021304/algumasaplica%25C3%25A7%25C3%25B5espr%253%25A1ticas-da-tecnologia-data-mining acedido em 17 de Dezembro de 2016.

Berry, M. J. A., & Linoff, G. S. (1997). *Data mining techniques: for marketing, sales, and customer support* (Second Edi). Indianapolis, Indiana: Wiley Publishing, Inc.

Brachman, R. J., & Anand, T. (1994). The process of Knowledge Discovery in Databases: A first sketch. *KDD Workshop*, *3*, 1–12.

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. *ACM SIGMOD Record*, *26*(2), 255–264.

Brusso, M. J. (2000). *Access Miner: Uma proposta para a extração de regras de associação aplicada à mineração do uso da Web*.

Chan, K., & Stolfo, J. (1998). Toward Scalable Learning with Non-uniform Class and

Cost Distributions : A Case Study in Credit Card Fraud Detection. *KDD, 1998*, 164–168.

Fayyad, U., Piatetsky, S. G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI magazine*, *17*(3), 24–26.

Gama, J., Carvalho, A., Faceli, K., Lorena, A., & Oliveira, M. (2012). *Extração de Conhecimento de Dados – Data Mining*. (Edições Sílabo, Ed.) (Primeira E). Lisboa.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information Tapestry. *Communications of the ACM*, *35*(12), 61–70.

Gonçalves, E. C. (2005). Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas Objective and Subjective Measures for Association Rules. *INFOCOMP–Journal of Computer Science*, 10.

Hahsler, M., Buchta, C., & Hornik, K. (2008). Selective association rule generation. *Computational Statistics*, *23*(2), 303–315.

Hahsler, M., Grün, B., Hornik, K., & Buchta, C. (2005). Introduction to arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, *14*(15), 1–25.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Techniques* (Third Edit). Waltham, USA: Morgan Kaufmann Publishers.

Han, J., Pei, J., & Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation. *ACM sigmod record*, *29*(2), 1–12.

Kiran, R. U., & Reddy, P. K. (2009). An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules. *Computational Intelligence and Data Mining*, *24*, 340–347.

Koh, Y. S., & Rountree, N. (2005). Finding Sporadic Rules Using Apriori-Inverse. *Advances in Knowledge Discovery and Data Mining*, 153–168.

Koh, Y. S., & Rountree, N. (2010). *Rare Association Rule Mining and Knowledge Discovery : Technologies for Infrequent and Critical Event Detection*. Hershey, USA: IGI Global.

Kohavi, R. (1998). Data Mining with MineSet : What Worked , What Did Not , and What Might. *Proceeding of the KDD-98 workshop on the Commercial Success of Data Mining*, 1–6.

Liu, B., Hsu, W., Chen, S., & Ma, Y. (2000). Analyzing the Subjective Interestingness of Association Rules. *IEEE Intelligent Systems and their Applications*, *15*(5), 47–55.

Liu, B., Hsu, W., & Ma, Y. (1999). Mining Association Rules with Multiple Minimum Supports. *ACM SIGKDD international conference on Knowledge discovery and data mining*, 337–341.

Liu, G., Lu, H., Yu, J., Wang, W., & Xiao, X. (2003). AFOPT: An Efficient Implementation of Pattern Growth Approach. *Fimi*.

Liu, H., Lu, H., Feng, L., & Hussain, F. (1999). Efficient Search of Reliable Exceptions. *Methodologies for knowledge discovery and data mining*, 194–204.

Mannila, H. (1997). Methods and problems in data mining. *Database Theory—ICDT*, 41–55.

McCormick, T. H., Rudin, C., & Madigan, D. (2011). A Hierarchical Model for Association Rule Mining of Sequential Events : an approach to automated medical symptom prediction.

Ozel, S. A., & Guvenir, H. A. (2001). An algorithm for mining association rules using

perfect hashing and database pruning. *10th Turkish Symposium on Artificial Intelligence and Neural Networks*, 257–264.

Park, J. S., Chen, M. S., & Yu, P. S. (1995). An Effective Hash-Based Algorithm for Mining Association Rules. *IBM*, *24*(2), 175–186).

Piatetsky Shapiro, G. (1991). Discovery, Analysis, and Presentation of Strong Rules. *Knowledge Discovery in Databases*, 229–248.

Raeder, T., & Chawla, N. V. (2011). Market basket analysis with networks. *Social Network Analysis and Mining*, *1*(2), 97–113.

Raskutti, B., & Kowalczyk, A. (2004). Extreme Re-balancing for SVMs: a case study. *ACM Sigkdd Explorations Newsletter*, *6*(1), 60–69.

Riddle, P., Segal, R., & Etzioni, O. (1994). Representation Design and Brute-force Induction in a Boeing Manufacturing Domain. *Applied Artificial Intelligence*, *8*(1), 125–147.

Silberschatz, A., & Tuzhilin, A. (1995). On Subjective Measures of Interestingness Discovery in Knowledge Bell Laboratories Measures. *KDD*, *95*, 275–281.

Siqueira, G. M., Prado, T. K., Júnior, W. M., & Carvalho, M. L. (2002). piFP-growth: Um Algoritmo Paralelo para Geração Incremental de Regras de Associação. *Anais WSCAD*, 76–83.

Srikant, R., & Agrawal, R. (1995). Mining Generalized Association Rules. *IBM*, 407–419.

Ulas, M. A. (1999). *Market Basket Analysis for Data Mining*.

Weber, I. (1998). On pruning strategies for discovery of generalized and quantitative

association rules. *Proceedings of Knowledge Discovery and Data Mining Workshop, Singapore*.

Weiss, G. M. (2004). Mining with Rarity: A Unifying Framework. *ACM Sigkdd Explorations Newsletter*, *6*(1), 7–19.

Weiss, G. M., & Hirsh, H. (1998). Learning to Predict Rare Events in Event Sequences. *KDD*, 359–363.

Zhou, L., & Yau, S. (2007). Association Rule and Quantitative Association Rule Mining among Infrequent Items. *Rare Association Rule Mining and Knowledge Discovery*, 15–32.

Anexos

Anexo 1 – Resumo de Resultados do Software R

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen
      0.001    0.1    1 none FALSE            TRUE       5   0.001      1     10
 target   ext
  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 7

set item appearances ...[0 item(s)] done [0.01s].
set transactions ...[554 item(s), 7242 transaction(s)] done [0.06s].
sorting and recoding items ... [168 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.02s].
writing ... [236 rule(s)] done [0.00s].
creating S4 object  ... done [0.12s].
```

Figura 11 - Resumo de Resultados Obtidos no Software R

Anexo 2 – Regras de Associação com Itens Frequentes

#	Antecedente (lhs)		Consequente (rhs)	Suporte	Confiança	*Lift*
1	{6350602,6350665}	=>	{6350603}	0.001104667	1.000000000	603.5000000
2	{6320099}	=>	{6320273}	0.001104667	0.800000000	579.3600000
3	{6350602}	=>	{6350603}	0.001518917	0.916666667	553.2083333
4	{6350603,6350665}	=>	{6350602}	0.001104667	0.888888889	536.4444444
5	{6350665}	=>	{6350603}	0.001242751	0.818181818	493.7727273
6	{6350602,6350603}	=>	{6350665}	0.001104667	0.727272727	478.8099174
7	{6350665}	=>	{6350602}	0.001104667	0.727272727	438.9090909
8	{6310030}	=>	{6320310}	0.001242751	0.818181818	131.6727273
9	{6320308}	=>	{6310518}	0.007042254	0.927272727	111.9218182
10	{6310825}	=>	{6320310}	0.002347418	0.377777778	78.1676190
11	{3020070}	=>	{3020344}	0.002485501	0.580645161	72.5005562
12	{1030030,1030040}	=>	{1030020}	0.001104667	0.400000000	70.6536585
13	{6320036}	=>	{6320006}	0.001795084	0.406250000	54.4826389
14	{6320100}	=>	{6320006}	0.001242751	0.391304348	52.4782609
15	{1030020,1030040}	=>	{1030030}	0.001104667	0.666666667	23.7832512
16	{1030020,1030030}	=>	{1030040}	0.001104667	0.615384615	21.7395872
17	{1030030}	=>	{1030080}	0.001242751	0.044334975	13.3780788
18	{1030020}	=>	{1030030}	0.001795084	0.317073171	11.3115463
19	{1030020}	=>	{1030040}	0.001657001	0.292682927	10.3395598
20	{2020007}	=>	{2020200}	0.001518917	0.134146341	9.5243902
21	{2020202}	=>	{2020250}	0.007594587	0.341614907	8.2741644
22	{2020007}	=>	{2020211}	0.001795084	0.158536585	8.1427089
23	{2020200}	=>	{2020250}	0.004556752	0.323529412	7.8361204
24	{2020250}	=>	{2020255}	0.001795084	0.043478261	6.1739130
25	{1430007}	=>	{1430001}	0.004832919	0.460526316	5.8103338
26	{1030076}	=>	{1030052}	0.001380834	0.200000000	4.5547170
27	{1430002}	=>	{1430001}	0.020160177	0.294949495	3.7212966
28	{1030030}	=>	{1030040}	0.002761668	0.098522167	3.4804758
29	{1030040}	=>	{2020202}	0.001104667	0.039024390	1.7553704
30	{6930003}	=>	{6930001}	0.001104667	0.108108108	1.4209055
31	{1730057}	=>	{2020250}	0.001380834	0.055865922	1.3531137
32	{1730057}	=>	{1430001}	0.002347418	0.094972067	1.1982364
33	{1430002}	=>	{1730057}	0.001380834	0.020202020	0.8173354
34	{1430001}	=>	{5240004}	0.001657001	0.020905923	0.6819851
35	{5240004}	=>	{1430002}	0.001104667	0.036036036	0.5272181
36	{2020250}	=>	{1430001}	0.001242751	0.030100334	0.3797676
37	{1430001}	=>	{6930001}	0.001657001	0.020905923	0.2747744

Tabela 13 - Listagem Completa das Regras de Associação com Itens Frequentes

Anexo 3 – *Interface* do *Software* SPMF - Aplicação de Algoritmo

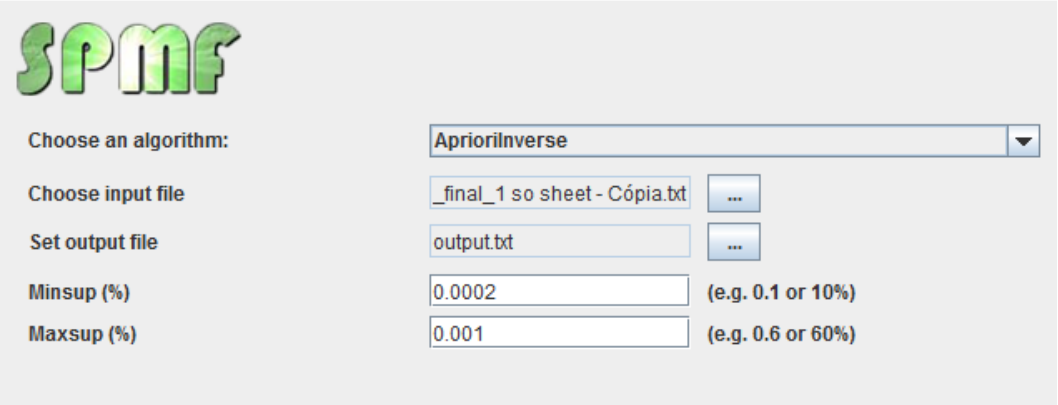

Figura 12 - Interface do Software SPMF - Aplicação do Apriori-Inverse

Anexo 4 – *Interface* do *Software* SPMF - Obtenção de Resultados

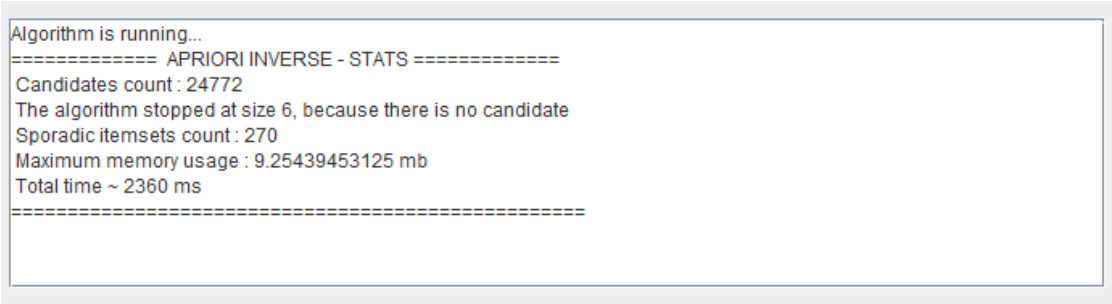

Figura 13 - Interface do Software SPMF - Resumo de Resultados

Anexo 5 – *Interface* do *Software* SPMF - Tabela de Resultados

Resultados
3020050 3020108 #SUP: 4
3020421 3020423 #SUP: 2
4030066 4030097 #SUP: 2
4030112 4030113 #SUP: 2
4030152 4030153 #SUP: 2
4530061 4530080 #SUP: 2
5210011 5210012 #SUP: 2
5210011 5220012 #SUP: 2
5210012 5220012 #SUP: 2
6310048 6320324 #SUP: 5
6310067 6320281 #SUP: 5
6310150 6310507 #SUP: 2
6310309 6320146 #SUP: 3
6310309 6320263 #SUP: 4
6310720 6320400 #SUP: 3
6310720 6320401 #SUP: 2
6310720 6320638 #SUP: 2
6310720 6320646 #SUP: 2
6320027 6320044 #SUP: 2
6320146 6320263 #SUP: 3
6320400 6320401 #SUP: 2
6320400 6320638 #SUP: 2
6320400 6320646 #SUP: 2
6320401 6320638 #SUP: 2
6320401 6320646 #SUP: 2
6320638 6320646 #SUP: 2
6350686 6350712 #SUP: 3
6350686 6350857 #SUP: 2
6350712 6350857 #SUP: 2
5210011 5210012 5220012 #SUP: 2
6310309 6320146 6320263 #SUP: 3
6310720 6320400 6320401 #SUP: 2
6310720 6320400 6320638 #SUP: 2
6310720 6320400 6320646 #SUP: 2
6310720 6320401 6320638 #SUP: 2
6310720 6320401 6320646 #SUP: 2
6310720 6320638 6320646 #SUP: 2
6320400 6320401 6320638 #SUP: 2
6320400 6320401 6320646 #SUP: 2

6320400 6320638 6320646 #SUP: 2
6320401 6320638 6320646 #SUP: 2
6350686 6350712 6350857 #SUP: 2
6310720 6320400 6320401 6320638 #SUP: 2
6310720 6320400 6320401 6320646 #SUP: 2
6310720 6320400 6320638 6320646 #SUP: 2
6310720 6320401 6320638 6320646 #SUP: 2
6320400 6320401 6320638 6320646 #SUP: 2
6310720 6320400 6320401 6320638 6320646 #SUP: 2

Tabela 14 - Listagem Completa Itemsets com Apriori Inverse

Anexo 6 – Regras de Associação com Itens Raros

Antecedente		Consequente	Suporte Absoluto	Confiança	*Lift*
3020108	=>	3020050	4	1.0000000	741.285714
6350603; 6350665	=>	6350602	5	1.0000000	741.285714
4030004; 4030014; 4030027	=>	4030028	3	1.0000000	648.625000
6350601	=>	6350112	3	0.5000000	432.416666
6350112	=>	6350602	3	0.5000000	370.642857
6350601	=>	6350602	3	0.5000000	370.642857
6310315; 6310518	**=>**	**6310031**	**3**	**1.0000000**	**305.235294**
6350746	=>	6310318	4	0.8000000	230.622222
4030074	=>	4030179	3	0.6000000	222.385714
4030521	=>	4030356	4	0.4444444	192.185185
3020344; 4030179	=>	3020070	3	1.0000000	172.966666
4530191	=>	4530193	5	0.6250000	154.434523
2020302	=>	3020027	4	0.3333333	123.547619
4030014; 4030050	=>	4030004	4	1.0000000	120.674418
4030027; 4030050	=>	4030004	3	1.0000000	120.674418
4030015	=>	4030014	3	0.3000000	119.746153
2030550	=>	4030070	3	0.3333333	108.104166
4530238	=>	4530032	4	0.3636363	104.828282
4030528	=>	4030003	3	1.0000000	99.7884615
3020464	=>	3020344	3	1.0000000	94.3454545
2030107; 3020070	=>	3020344	3	1.0000000	94.3454545
4030015; 4030027; 4030028	=>	4030005	4	1.0000000	91.0350877
4030005; 4030014	=>	4030004	3	0.7500000	90.5058139
4030005; 4030050	=>	4030004	3	0.7500000	90.5058139
6310050	=>	6310825	3	0.4285714	76.6847290
4030004; 4030213	=>	4030003	3	0.7500000	74.8413461
1730127	=>	9030473	3	0.1000000	74.1285714
6310315	**=>**	**6310031**	**3**	**0.2307692**	**70.4389140**
3530002	=>	3530001	5	0.5000000	68.2763157
4030014; 4030027	=>	4030005	3	0.7500000	68.2763157
4030014; 4030028	=>	4030005	3	0.7500000	68.2763157
4030003; 4030004	=>	4030005	5	0.7142857	65.0250626
2630055, 4030179	=>	4030120	3	1.0000000	62.5180722
1230003	=>	1230002	3	0.1875000	60.8085937
4030258	=>	4030050	5	0.3571428	59.7811059
6310318	=>	6310825	6	0.3333333	59.6436781
4030020; 4030213	=>	4030120	3	0.7500000	46.8885542
4030022	=>	4030213	3	0.3000000	42.0729729
6310030	**=>**	**6310518**	**4**	**0.3636363**	**41.9313131**

4530032	=>	4530013	3	0.1666666	41.1825396
4030254	=>	4030213	4	0.2857142	40.0694980
4030254	=>	4030050	3	0.2142857	35.8686635
2020622	=>	2020202	3	1.0000000	33.9150326
6310031	**=>**	**6310518**	**5**	**0.2941176**	**33.9150326**
3020070; 3530001	=>	2020202	3	1.0000000	33.9150326
4030003, 4030213	=>	4030120	3	0.5000000	31.2590361
4030015	=>	4030003	3	0.3000000	29.9365384
4030254	=>	4030003	4	0.2857142	28.5109890
2630054	=>	4530008	3	0.2500000	28.2010869
6310315	**=>**	**6310518**	**3**	**0.2307692**	**26.6102564**
4030356	=>	4030003	3	0.2500000	24.9471153
2630054	=>	3020344	3	0.2500000	23.5863636
1730127	=>	2020056	5	0.1666666	22.7587719
4030021	=>	2030107	3	0.4285714	22.0183875
4030254	=>	4030005	3	0.2142857	19.5075187
4030022	=>	4030120	3	0.3000000	18.7554216
1430005	=>	1430007	3	0.2500000	18.2711267
4030002	=>	4030005	4	0.1904761	17.3400167
3830231	=>	2930016	3	0.2142857	14.8257142
1730127	=>	1230004	4	0.1333333	14.4138888
4030002	=>	4030016	4	0.1904761	14.1197278
9030473	=>	2030400	3	0.4285714	14.0750452
4030254	=>	4030120	3	0.2142857	13.3967297
1230003	=>	2020200	4	0.2500000	13.1035353
4030050	=>	4030120	6	0.1935483	12.1002720
3530001	=>	4530008	4	0.1052631	11.8741418
4530193	=>	1730121	3	0.1428571	11.5825892
4030002	=>	3830500	3	0.1428571	10.4406438
1730127	=>	3830500	4	0.1333333	9.74460093
2030505	=>	2030400	5	0.2941176	9.65934475
2020010	=>	2020007	3	0.1428571	9.62708719
2630055	=>	2020202	5	0.2631578	8.92500859
1730123	=>	3020222	3	0.1071428	8.68694196
2020006	=>	2020211	3	0.2142857	8.68694196
2020010	=>	2020211	4	0.1904761	7.72172619
2020007	=>	2020200	11	0.1428571	7.48773448
2020007	=>	2020211	14	0.1818181	7.37073863
2730001	=>	2030400	6	0.2222222	7.29817158
4030179	=>	2020202	3	0.2142857	7.26750700
1730122	=>	2030107	3	0.1363636	7.00585058
4530193	=>	2020202	4	0.1904761	6.46000622
1230003	=>	2020202	3	0.1875000	6.35906862

3020096	=>	2030400	8	0.1904761	6.25557564
4030027	=>	4030120	4	0.1000000	6.25180722
2030401	=>	2020211	5	0.1515151	6.14228219
1730127	=>	2030400	5	0.1666666	5.47362869
2020212	=>	2020211	4	0.1142857	4.63303571
3020026	=>	2030400	4	0.1290322	4.23764801
4530008	=>	2020202	5	0.1086956	3.68641659
2020056	=>	2030400	4	0.1052631	3.45702864
4030027	=>	2020202	4	0.1000000	3.39150326
4030027	=>	2030400	4	0.1000000	3.28417721

Tabela 15 - Listagem Completa das Regras de Associação com Itens Raros