INVESTIGAÇÃO OPERACIONAL EM AÇÃO

CASOS DE APLICAÇÃO

RUI CARVALHO OLIVEIRA
JOSÉ SOEIRO FERREIRA
(EDITORES)

IMPRENSA DA UNIVERSIDADE DE COIMBRA
COIMBRA UNIVERSITY PRESS

CASO 15

DESENHO DE PROMOÇÕES DIFERENCIADAS EM EMPRESAS DE RETALHO RECORRENDO À SEGMENTAÇÃO DE CLIENTES

Vera L. Miguéis
Faculdade de Engenharia da Universidade do Porto
vera.migueis@fe.up.pt

Ana S. Camanho
Faculdade de Engenharia da Universidade do Porto
acamanho@fe.up.pt

João Falcão e Cunha
Faculdade de Engenharia da Universidade do Porto
jfcunha@fe.up.pt

Resumo

Uma boa relação entre as empresas e os seus clientes é um factor essencial para a competitividade das empresas. De facto, a fidelização dos clientes é um requisito para o sucesso sustentado de uma empresa. Este artigo propõe um método de apoio ao desenho de promoções diferenciadas para empresas de retalho, que visa reforçar a relação das empresas com os seus clientes. Este método baseia-se no conhecimento extraído de dados transacionais recolhidos através de cartões de fidelização. Inicialmente, os clientes são segmentados usando o algoritmo de agrupamento k-médias e posteriormente o perfil dos clientes integrados em cada grupo é caracterizado através de uma árvore de decisão. De seguida, identificam-se os produtos que são comprados em conjunto pelos clientes de cada segmento, e que constituem a base das promoções diferenciadas propostas. O estudo apresentado é feito em colaboração com uma empresa europeia que opera no domínio do retalho de base alimentar.

Palavras-Chave

Retalho, Promoções, *Data Mining*, Agrupamento, Classificação, Associação

1 Introdução

As recentes alterações económicas e sociais que têm ocorrido na Europa transformaram o sector de retalho. No passado, as empresas centravam-se na venda dos produtos e serviços sem procurarem conhecer os clientes que compravam os referidos produtos ou serviços. Com a proliferação da concorrência, tornou-se mais difícil atrair novos clientes e consequentemente as empresas viram-se obrigadas a concentrar esforços para manter os seus clientes. Para além disso, a evolução da sociedade e das condições económicas fez com que o estilo de vida dos clientes se alterasse, e como resultado estes passaram a ser menos influenciáveis pelas atividades de marketing das empresas. Neste contexto, enumeras empresas abandonaram estratégias centradas no produto e passaram a reger-se por estratégias centradas no cliente. A criação de relações de fidelidade com os clientes tem sido considerada um elemento crítico da estratégica de negócio das empresas. De facto, as empresas que se pretendem manter competitivas apostam na melhoria da sua relação com os seus clientes.

Algumas empresas investiram na criação de bases de dados capazes de armazenar os dados relativos a transações dos clientes que possuem cartões de fidelização. Por cada cliente, milhões de registos são recolhidos, permitindo a análise do histórico completo das compras. Contudo, o conhecimento que potencialmente se pode extrair destes dados é raramente integrado no processo de tomada de decisão, nomeadamente nas campanhas de marketing. A grande quantidade de dados armazenados resulta muitas vezes num problema de sobrecarga de informação acompanhado de escassez de conhecimento. Em geral, os analistas não têm sido capazes de estudar os dados e transformá-los em conhecimento útil para fins de aplicação em campanhas de marketing.

As técnicas de *data mining* têm surgido como ferramentas de grande potencial na análise dos dados resultantes da atividade transacional dos clientes. Estas técnicas podem ser usadas para detectar padrões e regras subjacentes ao comportamento dos clientes. Contudo, o uso das ferramentas de *data mining* no contexto do marketing é ainda incipiente. A segmentação de clientes ou a identificação de grupos de clientes com padrões de comportamento

semelhantes é feita, na maior parte dos casos, de uma forma ad-hoc. Para além disto, a maior parte das empresas ainda usa estratégias de comunicação em massa para chegar aos seus clientes.

Neste contexto, este estudo tem como objectivo segmentar os clientes de uma empresa europeia de retalho e propor políticas promocionais ajustadas aos clientes de cada segmento, com vista a aumentar o nível de serviço percebido pelos clientes e a sua consequente fidelização. Isto é feito com base no conhecimento extraído a partir de uma grande base de dados transacional, com recurso a ferramentas de *data mining*. A referida base de dados resulta do uso do cartão de fidelização por parte dos clientes da empresa.

Este capítulo tem a estrutura que se segue. A Secção 2 inclui uma breve revisão da literatura, com vista à contextualização do estudo. A Secção 3 inclui a apresentação das técnicas usadas: técnicas de agrupamento, classificação e associação. A Secção 4 apresenta a empresa usada como caso de estudo, bem como os resultados da análise efetuada. Finalmente, apresentam-se as conclusões e algumas ideias para trabalho futuro.

2 Revisão da literatura

A fidelização de clientes por parte das empresas tem ganho proeminência na literatura de marketing e gestão. De acordo com Keaveney (1995), a fidelização dos clientes actua frequentemente como factor de retenção, de onde resulta um papel crucial na competitividade das empresas.

Bettencourt and Gwinner (1996) considera que o tratamento personalizado dos clientes é um dos pilares da sua satisfação. Desta forma, é possível fomentar a fidelização dos clientes às empresas através de promoções dirigidas. Contrariamente às políticas promocionais tradicionais que tratam os clientes indistintamente, o desenho de promoções dirigidas requer um conhecimento profundo do comportamento dos clientes. Isto possibilita a recomendação de produtos ou serviços que respondem às necessidades específicas dos clientes. De acordo com Ngai et al. (2009), a caracterização dos clientes é usualmente feita através da segmentação do mercado e a recomendação de produtos é feita com recurso à análise dos cabazes de compra.

As primeiras abordagens ao problema de segmentação basearam-se no uso de critérios geográficos. As empresas agrupavam os seus clientes de acordo com a sua área de residência ou de trabalho. Posteriormente, surgiram abordagens baseadas em critérios demográficos, através das quais os clientes eram agrupados de acordo com a idade, sexo, rendimento, ocupação, entre outros. Nos anos 60 a segmentação do mercado sofreu um crescendo de popularidade. Twedt (1964) sugeriu o uso de modelos de segmentação baseados no volume de vendas. Assim, as atividades de marketing deviam centrar-se nos clientes com um número considerável de transações. Esta teoria, denominada «heavy half theory», sugeria que metade dos clientes poderia representar 80% das vendas de uma empresa. Frank et al. (1967) criticou este modelo de segmentação, afirmando que este assume que os clientes mais representativos em termos de vendas tinham características que os diferenciavam dos restantes, o que não foi comprovado pelo modelo de regressão que desenvolveu. Posteriormente, Haley (1968) introduziu um modelo de segmentação que diferencia os clientes de acordo com os benefícios que estes esperavam do produto ou serviço. Assim, o mercado poderia ser segmentado de acordo com as valias esperadas pelos clientes: qualidade, desempenho, imagem, serviço, entre outras. Estes modelos promoveram mais investigação, o que permitiu o desenvolvimento de modelos complexos de segmentação baseados no estilo de vida dos clientes. O conceito de estilo de vida, associado ao marketing por Lazer (1964), é baseado no facto de cada indivíduo possuir padrões de vida próprios, que podem influenciar a sua motivação para a compra de determinados produtos e marcas. Durante os anos 70, a validade destes modelos de segmentação multivariada foi questionada (Green and Wind, 1973), o que levou à modelação teórica do comportamento dos clientes (por exemplo, Blattberg et al. (1978)). Uma década mais tarde, Mitchell (1983) desenvolveu um modelo de segmentação psicográfica que divide o mercado em grupos de acordo com a classe social, estilo de vida e características da personalidade. Contudo, as dificuldades de implementação deste modelo de segmentação mais complexo viriam a revelar-se críticas (Piercy and Morgan, 1993) e (Dibb and Simkin, 1997).

Mais recentemente, a literatura de marketing tem dado ênfase ao abandono de padrões previsíveis de comportamento e consumo por parte dos clientes. A diversidade de necessidades e desejos, influenciada pelo estilo de vida, rendimento e idade, têm tornado menos efetivas as abordagens de segmentação propostas no passado. Assim, os modelos actuais de segmentação de mercado são baseados no comportamento dos clientes inferido a partir dos registos das transações e de inquéritos. Os dados recolhidos são explorados com técnicas de *data mining*, nomeadamente técnicas de agrupamento. Exemplos da aplicação de técnicas de *data mining* na segmentação de clientes, com recurso a inquéritos, incluem Kiang et al. (2006). No contexto dos serviços de comunicação a longa distância, os clientes foram segmentados usando variáveis psicográficas, recolhidas a partir de um inquérito constituído por 68 questões relacionadas com o seu comportamento. Min and Han (2005) agruparam os clientes com gostos semelhantes em relação a filmes, usando a informação fornecida por cada cliente. A classificação relativa dos filmes feita pelos clientes permitiu deduzir o valor percebido por cada cliente em relação a cada filme. Helsen and Green (1991) também identificaram os segmentos de mercado para um novo sistema informático com base em técnicas de agrupamento, usando dados obtidos a partir de um inquérito. A segmentação baseou-se na importância dada às características do produto.

No que se refere às abordagens de segmentação baseadas nos dados transacionais armazenados em bases de dados, o modelo de segmentação «recency, frequency and monetary value» (RFM), introduzido nos anos 20 (Roel, 1988), é um modelo de segmentação muito utilizado. Este modelo considera a data da última compra (recency), a frequência com que o cliente faz compras (frequency) e o valor gasto (monetary value), extraídos a partir de uma base de dados transacional. Liu and Shih (2005) é um exemplo de um estudo recente de segmentação que usa o modelo RFM e que tem como objectivo especificar os segmentos no mercado de retalho de *hardware*.

Agrupados os clientes com características semelhantes, é muitas vezes necessário inferir o seu comportamento ou estilo de vida a partir dos dados transacionais. Para isso poderá ser interessante encontrar padrões comuns nas suas compras, como por exemplo conjuntos de produtos que são normalmente

adquiridos em conjunto. As técnicas de associação de *data mining* são frequentemente usadas para a análise de cabazes de compra, com o objectivo de extrair regras de associação de produtos. O uso destas técnicas para suportar estratégias promocionais diferenciadas pode ser considerado relativamente novo na literatura de marketing. Por exemplo, Van den Poel et al. (2004) usa este tipo de técnicas para definir vendas cruzadas e Brijs et al. (2004) usa-as para definir a disposição dos produtos nas lojas. A literatura que aborda as técnicas de associação centra-se principalmente no desenvolvimento de algoritmos, sendo a sua aplicação em casos reais e a integração com as políticas de marketing menos frequente.

Neste contexto, este estudo propõe a aplicação de técnicas de *data mining*, incluindo técnicas de agrupamento e de associação, para desenhar promoções diferenciadas para uma empresa de retalho europeia.

3 Metodologia

A metodologia usada neste estudo tem como objectivo apoiar o desenho de promoções por forma a aumentar os níveis de fidelização dos clientes. Numa primeira fase, os clientes da empresa usada como caso de estudo são segmentados usando uma técnica de agrupamento. A isto segue-se a caracterização do perfil dos clientes integrados em cada segmento, com recurso a uma árvore de decisão. Posteriormente, identificam-se os produtos que são frequentemente comprados em conjunto pelos clientes de cada segmento, usando uma técnica de associação. Através deste procedimento, é possível definir descontos promocionais para os clientes selecionados, tendo em conta o histórico de compras e as associações de produtos identificadas para os clientes do mesmo segmento. Seguidamente descrevem-se as técnicas subjacentes à aplicação da metodologia descrita.

3.1 Análise de agrupamento

As técnicas de agrupamento são técnicas de *data mining* muito populares que agrupam os dados em conjuntos previamente desconhecidos, com base na sua semelhança. Existe uma grande variedade de algoritmos de agrupamento (ver Jain et al. (1999) para uma revisão). Estes algoritmos podem ser

classificados como sendo do tipo particional ou do tipo hierárquico. Os algoritmos particionais permitem dividir os dados em grupos que não se sobrepõem, ou seja, cada objeto pertence unicamente a um grupo. As técnicas particionais requerem a definição à priori do número de grupos. Apesar desta limitação, estas técnicas têm a vantagem de permitir a optimização da função de erro. Os algoritmos hierárquicos podem ser do tipo aglomerativo ou do tipo divisivo. Os algoritmos aglomerativos iniciam-se com cada objeto num grupo e, de seguida, juntam os grupos até que todos os dados estejam no mesmo grupo. Em cada iteração os dois grupos mais semelhantes são agrupados. O algoritmo divisivo inicia-se com um grupo contendo todos os objetos e iterativamente divide este grupo em grupos mais pequenos. Quer o algoritmo aglomerativo quer o algoritmo divisivo produzem uma sequência de grupos, em que o do topo inclui todos os objetos, e em que os do fundo incluem um único objeto. O diagrama resultante da aglomeração ou divisão dos grupos, denominado de dendrograma, permite definir facilmente o número apropriado de grupos para atingir um determinado objectivo. Contudo, o algoritmo hierárquico não permite a realocação de objetos que tenham sido «incorretamente» agrupados ou separados em fases anteriores do procedimento.

O algoritmo k-médias, introduzido por Forgy (1965) e posteriormente desenvolvido por MacQueen (1967), é um dos algoritmos de agrupamento mais populares. O k-médias, que pertence à classe dos algoritmos particionais, tem duas vantagens principais: é fácil de implementar e não exige grande esforço computacional, o que o torna apropriado para grandes bases de dados. Neste estudo aplica-se o algoritmo k-médias para segmentar os clientes, tendo em conta os dados transacionais. Este algoritmo visa distribuir um conjunto de n objetos de dados por k grupos, por forma a obter-se uma semelhança grande intra-grupos e uma semelhança reduzida entre os diferentes grupos. O algoritmo k-médias tem a desvantagem de requerer a definição à priori do número de grupos (tal como todos os algoritmos particionais), e depende das sementes iniciais (objetos de dados definidos como centroides iniciais dos grupos). De facto, este algoritmo implica a definição das sementes para a primeira iteração. A selecção de sementes diferentes pode conduzir à obtenção de grupos diferentes, especialmente quando os dados em análise contêm

valores extremos. Este algoritmo não possui qualquer meio para determinar as sementes iniciais mais apropriadas. Contudo, o procedimento standard é correr o algoritmo com diferentes sementes e escolher aquelas que permitem obter o valor mais baixo para a função de erro, normalmente o erro quadrado. Esta medida de erro avalia a distância entre os objetos de dados e os centroides dos respetivos grupos. Neste estudo usou-se o software RapidMiner, que usa um conjunto aleatório de objetos de dados como centroides iniciais para o agrupamento. Depois de correr o algoritmo k-médias com várias sementes aleatórias, selecionaram-se aquelas que permitiram obter o menor erro quadrado.

Depois da selecção das sementes iniciais, cada objeto é alocado ao grupo mais próximo, de acordo com a distância (tipicamente a distância euclidiana) entre o objeto e o centroide dos grupos. No final de cada iteração, os centroides são atualizados com base na média dos objetos incluídos no grupo, para que se possa iniciar uma nova iteração. Este processo termina quando a função de erro converge para um valor próximo do mínimo.

No que diz respeito à definição do número de grupos, são referidas na literatura várias heurísticas, já que teoricamente não é possível determinar o número óptimo de grupos (ver Tibshirani et al. (2001) para uma revisão). Neste estudo usaram-se dois critérios: o índice de Davies-Bouldin e o erro quadrático médio. O índice de Davies-Bouldin, desenvolvido por Davies and Bouldin (1979), é baseado na razão entre a soma da dispersão interna dos grupos e a distância entre grupos. O número apropriado de grupos é aquele que corresponde a um valor mais baixo deste índice. O erro quadrático médio, usado neste contexto para construir a denominada curva do cotovelo (Aldenderfer and Blashfield, 1984), representa a dispersão dentro dos grupos, definida tipicamente como a soma dos quadrados das distâncias entre todos os objetos e o centroide do grupo correspondente, dividida pelo número de grupos. À medida que o número de grupos aumenta, a medida de erro diminui monotonicamente e, a partir de determinado valor de k, a diminuição deixa de ser significativa. Este «cotovelo» é normalmente usado para definir o número adequado de grupos.

3.2 Regras de classificação

A extração das regras de classificação é um processo que pode ser usado para identificar as características que distinguem os objetos de diferentes grupos. Uma reconhecida técnica de classificação é a árvore de decisão, que permite facilmente obter as regras que caracterizam os grupos. Isto é feito através da identificação dos atributos, ou seja, as variáveis de agrupamento, consideradas relevantes para a descrição dos grupos. Uma árvore de decisão tem a estrutura de uma árvore com nós e ramos. Os nós podem ser de dois tipos: folha e nó de decisão. Uma folha representa uma classificação, ou seja, o nome do grupo, enquanto um nó de decisão está associado a uma escolha entre dois ramos, correspondendo a diferentes valores de um determinado atributo. Os ramos com origem num determinado nó de decisão representam as alternativas possíveis para o intervalo de valores dos atributos.

Existem na literatura inúmeros algoritmos de construção de árvores de decisão. O CHAID (Kass, 1980), o ID3 (Quinlan, 1986), o CART (Breiman et al., 1984), o C4.5 (Quinlan, 1992) e o AID (Morgan and Sonquist, 1963) são alguns exemplos. Por forma a remover os ramos da árvore que possuem um reduzido poder de discriminação entre os objetos, a árvore de decisão obtida é frequentemente podada. A principal razão para a execução deste procedimento é a obtenção de árvores genéricas, facilmente interpretáveis e que classificam de uma forma precisa a maior parte dos objetos.

O algoritmo usado neste estudo para a construção da árvore de decisão foi o algoritmo C4.5. Os detalhes sobre a parametrização deste algoritmo podem ser obtidos em Rapid-I (2009). Os valores dos parâmetros usados foram os seguintes: o critério de selecção do atributo foi a razão de ganho, a dimensão mínima para a divisão foi 40.000, a dimensão mínima das folhas foi 20.000, o ganho mínimo foi 0.19, o comprimento máximo foi 20 e o erro pessimista foi 0.5. Esta parametrização foi especificada a partir da comparação da percentagem de clientes classificados corretamente para diferentes configurações. Para avaliar a precisão da árvore de decisão, a base de dados foi dividida em 80% para efeitos de treino e 20% para testes. Esta divisão foi estratificada, de forma que a percentagem de clientes pertencentes aos

diferentes grupos nas amostras de treino e teste fosse aproximadamente idêntica à da base de dados inicial.

A árvore construída pode ser usada para caracterizar cada segmento de mercado. Isto requer que se inicie a caracterização a partir da raiz da árvore e que se faça um movimento ao longo dos ramos, contendo os valores mutualmente exclusivos dos atributos, até se atingir uma folha (contendo o nome do grupo). Assim, para se extrair as regras referentes a um grupo é necessário considerar todos os caminhos desde a raiz até às folhas contendo o nome do grupo.

3.3 Análise do cabaz de compras

A análise do cabaz de compras com recurso a técnicas de associação visa identificar grupos de produtos ou serviços comprados em simultâneo. Uma regra de associação pode ser representada na forma $X \Rightarrow Y$, o que significa que quando X é comprado Y é também comprado. X é denominado de antecedente e Y de consequente, de tal forma que o antecedente desencadeia a compra do consequente. O problema de associação envolve duas fases. A primeira fase diz respeito à descoberta dos produtos comprados mais frequentemente. Os algoritmos mais usados nesta fase são o algoritmo Apriori (Agrawal and Srikant, 1994), o algoritmo Frequent-pattern growth (Han et al., 2000) e o algoritmo Eclat (Zaki et al., 1997). O algoritmo usado neste estudo foi o algoritmo Apriori. A segunda fase da determinação das regras de associação diz respeito ao processo de geração de regras de associação, ou seja, à definição de quais os produtos antecedentes (X) e quais os consequentes (Y).

Os algoritmos de descoberta dos produtos comprados mais frequentemente envolvem várias fases. Na primeira fase, são identificados os produtos que possuem um suporte mínimo definido. A medida de suporte, representada por s(X), corresponde à percentagem de cabazes de compra analisados que contêm o produto X. Em seguida, os produtos com um suporte maior do que o limite especificado são combinados dois a dois, e é calculado o suporte desses conjuntos de produtos, isto é, estima-se o valor de s(X,Y) (ver expressão (1)). Os grupos de dois produtos com um suporte maior do que o mínimo definido pelo analista são considerados no passo seguinte. A cada um deles é

adicionado mais um produto, seleccionado de entre os produtos frequentes, identificados no primeiro passo. Este processo iterativo continua até que não seja possível definir mais conjuntos de produtos com um suporte acima do mínimo especificado.

$$s(X \Rightarrow Y)=P(X \cup Y) \quad (1)$$

Uma vez concluída a identificação dos produtos comprados mais frequentemente, é necessário identificar, dentro dos conjuntos de produtos frequentes, os produtos antecedente e consequente. Isto requer o cálculo da medida de confiança. Considerando-se a regra de associação $X \Rightarrow Y$, a confiança $c(X \Rightarrow Y)$ é a razão entre o número de cabazes que contêm ambos os produtos X e Y e o número de cabazes que contêm apenas X (ver expressão (2)). No caso da medida de confiança ser superior ao limite mínimo definido, a regra é considerada uma regra de associação.

$$c(X \Rightarrow Y)=P(Y \mid X)= s(X \cup Y) \,/s(X) \quad (2)$$

O *lift* é outra medida comum na análise das regras de associação, que avalia o nível de dependência entre os produtos que integram a regra de associação. É obtido dividindo o suporte de X e Y, s(X,Y) (representando a percentagem de cabazes que contêm X e Y em relação ao conjunto total de cabazes) pelo produto do suporte de X e Y, considerados em separado (ver expressão 3).

$$lift(X,Y)=s(X,Y)/(s(X).s(Y)) \quad (3)$$

Se o *lift* for igual a 1, existe independência entre a ocorrência de vendas dos produtos X e Y. Se o *lift* for maior do que 1, os produtos tendem a ser comprados em conjunto, e se for menor do que 1, os produtos tendem a ser comprados separadamente. As regras que apresentam um *lift* inferior a 1 são geralmente ignoradas, já que apenas as regras com *lift* superior a 1 são interessantes para suportar políticas de marketing.

4 Desenho de promoções dirigidas

4.1 Apresentação da empresa usada como caso de estudo

Este estudo desenvolve um método de desenho de promoções no sector de retalho, considerando as associações de produtos observadas nos grupos homogéneos de clientes. Usou-se como caso de estudo uma empresa europeia de retalho. Esta empresa de retalho é constituída por uma cadeia de hipermercados e duas cadeias de supermercados (grandes e pequenas lojas). Os formatos das lojas diferem essencialmente na diversidade e preço dos produtos oferecidos, na área de venda e no tamanho da cidade onde se localizam.

O estabelecimento de relações de fidelidade com os clientes tornou-se um elemento essencial da estratégia desta empresa. O desenvolvimento dos sistemas de informação da empresa e a implementação de um programa de fidelização têm permitido à empresa recolher os dados relativos ao perfil de cada cliente (ex: nome do cliente, morada, data de nascimento, sexo, número de pessoas no agregado familiar, número de telefone e número de um documento identificativo) e às suas transações (data, hora, loja, produtos e preços). Este programa é suportado essencialmente por um cartão de fidelização, e atualmente cerca de 80% do número de transações são registradas usando o cartão de fidelização.

Atualmente, os clientes da empresa são segmentados de duas formas. Uma delas consiste no agrupamento de clientes com base nos seus hábitos de consumo. Este modelo de segmentação é uma versão simplificada do modelo RFM, e é denominado internamente como modelo de Frequência e Valor Monetário (FM). De acordo com os valores destas duas variáveis, a empresa especifica 8 grupos de clientes. Cada cliente integra um destes grupos, com base no número médio de compras efetuadas num período de 8 semanas e no valor médio gasto por compra. As alterações na percentagem de clientes pertencentes a cada grupo são usadas para sinalizar uma eventual necessidade de promover ações para melhorar o relacionamento com o cliente. Por exemplo, se o número de clientes nos grupos com mais visitas às lojas sofrer uma redução, a empresa é alertada para lançar campanhas de marketing, a fim

de motivar os clientes a ir às lojas com mais frequência. O outro método de segmentação baseia-se nas necessidades e preferências dos clientes. Neste caso, os clientes são agrupados em 7 segmentos de acordo com o balanço entre as categorias de produtos que compram. Para isto, calculam-se as percentagens de produtos que cada cliente compra pertencentes a cada categoria de produtos predefinida pela empresa. Os clientes que apresentam percentagens semelhantes são agrupados, usando um algoritmo de agrupamento.

Quanto à estratégia promocional da empresa, existem essencialmente 3 tipos de políticas:

- Descontos em produtos específicos anunciados nas prateleiras das lojas e folhetos, que podem ser usados por todos os clientes com um cartão de fidelização;
- Descontos nas compras feitas em determinados dias (desconto percentual ou absoluto sobre o valor total das compras). Estes descontos podem ser usufruídos por clientes que apresentem o cupão de desconto enviado pelo correio no ponto de venda (POS);
- Descontos em produtos específicos em dias selecionados. Estes podem ser enviados pelo correio ou emitidos no POS.

Os dois primeiros tipos de promoções não fazem distinção entre os clientes de diferentes segmentos. O terceiro tipo, em vez de utilizar os modelos de segmentação descritos anteriormente, utiliza um modelo baseado no histórico das compras do produto incluído na promoção. Os descontos só são emitidos para os compradores mais frequentes do produto, ou para aqueles clientes que normalmente não compram o produto, para os incentivar a tornarem-se compradores.

A análise descrita neste capítulo baseia-se nos dados transacionais de clientes com o cartão de fidelização e é independente dos métodos usados atualmente pela empresa. A metodologia proposta apoia-se essencialmente em técnicas de *data mining*, ao contrário da metodologia usada atualmente pela empresa. A base de dados utilizada inclui os registos do último trimestre de 2009. Cada transacção inclui o número de identidade do cliente, a data e a hora da transação, o produto transacionado e o preço do produto. Além das informações transacionais, a empresa forneceu informações demográficas sobre

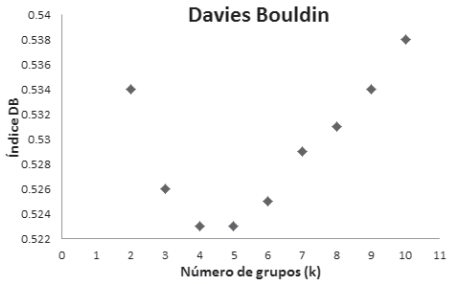
cada cliente: código postal da residência, cidade, data de nascimento, sexo e número de pessoas no agregado familiar. A preparação da base de dados para a análise exploratória envolveu a integração dos dados provenientes de diferentes fontes, a eliminação dos extremos e a seleção dos dados relevantes para a análise. Os clientes cujo valor médio gasto por compra ou cujo número médio de compras por mês estão fora do intervalo definido pela média mais três desvios padrão foram excluídos da análise. Estes extremos representavam 1,96% dos clientes incluídos na base de dados original. Dado que se está interessado no desenho de promoções para famílias, selecionou-se para a análise apenas os clientes cuja média do valor gasto por compra era igual ou inferior a €500. Estes clientes representavam 97,28% dos clientes incluídos na base de dados original. Habitualmente as compras superiores a este valor são feitas por pequenos retalhistas que revendem os produtos em lojas concorrentes, pelo que não devem ser alvo de programas promocionais. Após o processo de preparação dos dados a analisar, a base de dados passou a conter 2.142.439 clientes.

4.2 Segmentação

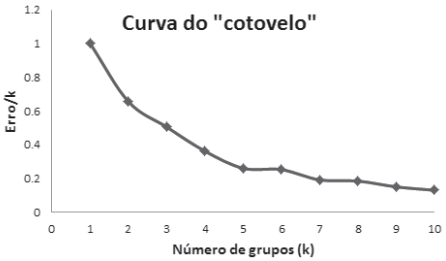
Neste estudo, a segmentação de clientes baseou-se na frequência e valor monetário das compras dos clientes. Estes indicadores representam os hábitos de compra dos clientes. À semelhança de Marcus (1998), definiram-se como variáveis exploratórias de agrupamento as seguintes variáveis: o número médio de compras feitas por mês e o valor médio gasto por compra. Embora a literatura sugira o uso das variáveis RFM, neste estudo não se incluiu a variável referente à data da última compra, já que se considerou que o período de análise não era suficientemente longo para permitir a diferenciação de clientes nesta dimensão.

A segmentação dos clientes por meio do algoritmo k-médias requer a definição à priori do número de grupos (k). A fim de se definir o número de segmentos, foi analisada a curva do cotovelo e calculado o índice de Davies-Bouldin para diferentes valores de k, como representado na Figura 1. De acordo com o índice de Davies-Bouldin, o número mais adequado de segmentos seria quatro ou cinco, uma vez que estes correspondem ao menor

valor do índice. A partir da curva do cotovelo, podemos concluir que cinco grupos parece ser a opção mais adequada. Por este motivo, os clientes analisados foram agrupados em cinco grupos.



(a) Índice de Davies-Bouldin.



(b) Curva do erro quadrático médio em torno dos centroides.

Figura 1 - Medidas de erro para diferentes valores de k.

A percentagem de clientes incluídos em cada grupo especificado pelo algoritmo k-médias foi a seguinte: 37% no *Cluster* 4, 27 % no *Cluster* 2, 20 % no *Cluster* 3, 8 % no *Cluster* 0 e 8% no *Cluster* 1.

A fim de caracterizar o perfil de clientes pertencentes a cada segmento, extraíram-se as regras subjacentes a esta classificação utilizando uma árvore de decisão. Os perfis dos segmentos resultantes da árvore de decisão são ilustrados na Figura 2, e podem ser descritos como se segue. O *Cluster* 0 inclui os clientes que fazem compras mais do que 6,2 vezes por mês. O *Cluster* 3 corresponde aos clientes que vão às compras entre 3,2 e 6,2 vezes por mês. O *Cluster* 1 inclui os clientes fazem compras menos do que 3,2 vezes por mês e gastam mais do que €135,9 por visita. O *Cluster* 2 inclui os clientes que fazem compras entre 1,5 e 3,2 vezes por mês e gastam menos do que €135,9 por visita. O *Cluster* 4 corresponde aos clientes que fazem compras menos do que 1,5 vezes por mês e gastam menos de €135,9 por visita. Importa salientar que a percentagem de clientes classificados corretamente com recurso à árvore de decisão é de 97.1%, o que significa que as referidas regras podem ser um bom suporte à segmentação de novos clientes.

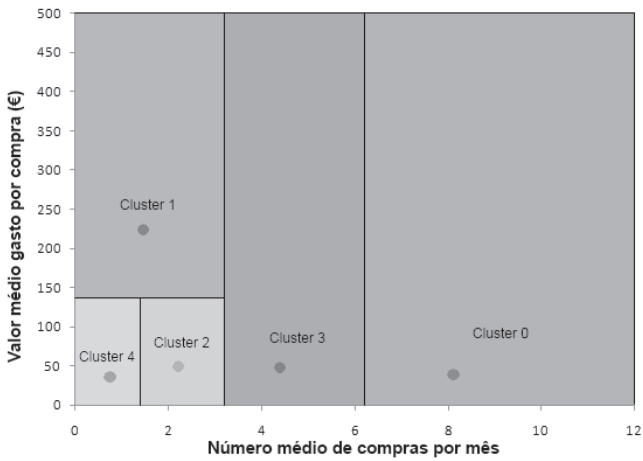


Figura 2 - Caracterização dos grupos.

4.3 Desenho de promoções diferenciadas

A análise dos cabazes de compra, destinada à identificação de associações de produtos dentro dos grupos, pode apoiar o desenho de promoções diferenciadas. Segundo Morales (2005) e Palmatier et al. (2009), estas promoções podem promover a fidelização dos clientes pois fazem com que os clientes sintam benefícios em se manter fiéis à empresa e permitem criar um sentimento de gratidão para com a empresa. As promoções diferenciadas visam premiar o relacionamento do cliente com a empresa e sugerir a futura aquisição de produtos que podem suscitar o interesse do cliente.

Para o propósito desta análise, um cabaz é o conjunto de todos os produtos que foram comprados por um cliente nos três meses analisados. Note-se que este estudo não se foca na análise dos produtos que foram comprados na mesma transação. A análise dos cabazes de compra é feita ao nível da subcategoria do produto, e não ao nível do produto, já que se pretende descobrir o tipo de produtos que pode ser potencialmente interessante para os clientes de um determinado segmento.

Vamos supor que, para fins ilustrativos, a empresa está interessada em promover ações de fidelização dirigidas aos clientes que apresentam uma baixa frequência de visitas à loja e baixo valor das compras. Estes clientes

provavelmente fazem a maior parte das suas compras em lojas da concorrência, de modo que a empresa pode estar interessada em motivá-los a visitar as lojas com maior frequência e ampliar a diversidade de produtos adquiridos. O segmento alvo que se enquadra neste perfil é o *Cluster* 4. Assim, procedeu-se à aplicação do algoritmo de associação considerando os cabazes de compra deste grupo (785.679 cabazes).

As subcategorias foram consideradas associadas quando se verificaram as seguintes condições: um *lift* maior que 1, uma confiança maior ou igual a 50% e um suporte maior ou igual a 2%. Isto significa que tem de haver pelo menos 15.713 cabazes que incluam as subcategorias consideradas associadas. Considerando estas condições, identificaram-se 29 regras de associação de subcategorias.

A maior parte das regras de associação obtidas incluem produtos pertencentes à mesma categoria, tais como peras e maças. Um grande número de regras incluem os produtos comprados mais regularmente, como por exemplo o arroz e o leite, como poderia ser antecipado. Para além destas, foram também identificadas algumas relações entre produtos pertencentes a diferentes categorias. A título de exemplo, a Tabela 1 apresenta algumas das regras de associação de subcategorias identificadas, ordenadas de acordo com a confiança.

Tabela 1: Regras de associação para o Cluster 4

Antecedente (x)	Consequente (y)	Conf.	*Lift*	s(x,y)	s(x)	s(y)
Condicionador de cabelo	Champô	64%	5,50	3%	5%	12%
Tomates	Vegetais para a salada	60%	4,55	5%	9%	14%
Couve	Vegetais para sopa	58%	3,68	6%	10%	16%
Arroz	Massa	58%	3,08	9%	16%	19%
Fiambre em fatias	Queijo flamengo	57%	3,89	7%	12%	15%
Sal	Arroz	53%	3,40	4%	8%	16%
Peras	Maçãs	51%	3,68	6%	12%	14%
Carne picada	Miudezas de porco	51%	3,52	4%	8%	15%
Óleo	Arroz	51%	3,26	6%	12%	16%
Vegetais embalados	Vegetais para a sopa	51%	3,22	3%	6%	16%

Considere-se que a empresa quer motivar os clientes para a compra de produtos que podem ser interessantes para estes, embora não tenham vindo a fazer parte das suas listas de compras recentes. Para isto, a empresa pode emitir um cupão de desconto no balcão de pagamento que vise o produto

consequente da regra de associação que não tem sido comprado, apesar do produto antecedente ter vindo a ser comprado. Por exemplo, a empresa pode sugerir um desconto na próxima compra de champô aos clientes que têm comprado condicionador de cabelo, mas que não têm comprado champô. Através da análise da base de dados da empresa concluiu-se que esta ação promocional pode ser relevante e motivar cerca de 15 mil clientes a comprar champô, dado que recentemente apenas compraram condicionador. Este tipo de promoções não só permite motivar os clientes a visitar mais vezes as lojas, mas também a aumentar a diversidade de produtos comprados. Para além disto, este tipo de ações tendem a fidelizar o cliente, na medida em que este se pode sentir parte integrante da estratégia da empresa.

A fim de verificar se as ações promoções seriam diferentes se a empresa definisse como alvo outro grupo de clientes, apresenta-se no Apêndice algumas das regras de associação de produtos que resultam da análise dos cabazes de compra dos clientes dos restantes segmentos. A primeira regra apresentada para cada grupo de clientes é aquela que apresenta maior confiança. Impondo critérios semelhantes aos definidos para o *Cluster* 4, *lift* (>1), confiança (≥50%) e suporte (≥2%), o número total de regras identificadas foi de 18.866 para o *Cluster* 0, 4.911 para o *Cluster* 1, 1.761 para o *Cluster* 2 e 10.408 para o *Cluster* 3. É interessante verificar que, para os pequenos segmentos, tais como o *Cluster* 0, utilizando os mesmos critérios, é possível obter mais regras de associação de subcategorias do que para os segmentos maiores como o *Cluster* 4. Dado que os clientes do *Cluster* 4 fazem compras esporádicas, é mais difícil encontrar padrões de compra. Embora algumas das regras identificadas para os diferentes grupos de clientes sejam idênticas, tais como o facto de a compra de condicionador de cabelo despoletar a compra de champô, pode verificar-se pela análise da Tabela 1 que os clientes têm hábitos de compra diferentes. A regra com maior confiança identificada para cada grupo de clientes é diferente para todos os grupos. Portanto, acredita-se que o procedimento sugerido neste estudo, que consiste no desenvolvimento da segmentação antes do processo de descoberta das regras de associação, pode contribuir para a melhoria da relação entre a empresa e os seus clientes.

5 Conclusão

Este estudo propõe um modelo para a segmentação dos clientes de uma empresa europeia de retalho e propõe políticas promocionais dirigidas aos clientes de cada segmento, com o objetivo de reforçar as suas relações de fidelização.

A aplicação de técnicas de *data mining* permitiu encontrar grupos naturais de clientes, com base nos dados transacionais armazenados na base de dados relativa ao cartão de fidelização da empresa. A segmentação baseou-se na frequência (número médio de compras feitas por mês) e no valor monetário (valor médio gasto por compra) das compras dos clientes. Usando uma técnica particional de agrupamento, os clientes foram agrupados em cinco grupos de acordo com os seus hábitos de compras. A análise também envolveu a construção de uma árvore de decisão, a fim de extrair as regras subjacentes à segmentação dos clientes. Assim, foi possível definir um perfil dos clientes de cada segmento, que pode ser usado para classificações futuras de clientes com elevada precisão.

O estudo descrito também identificou regras de associação de produtos dentro de cada segmento, tendo em conta os cabazes de compra dos clientes. Estas regras permitiram o desenho de promoções diferenciadas, que podem ser fundamentais para motivar os clientes a aumentar as suas compras e a manter-se fieis à empresa.

Como trabalho futuro espera-se que seja feita a implementação da metodologia proposta na empresa usada como caso de estudo. Após a implementação, seria importante entrevistar clientes pertencentes a cada grupo, a fim de verificar se estão satisfeitos com as promoções que lhes são dirigidas. Seria também interessante verificar se o comportamento dos clientes revela uma intensificação da sua relação com a empresa, como resultado das acções promocionais diferenciadas baseadas na metodologia proposta neste trabalho. Deverão ser considerados indicadores de fidelização, como por exemplo a taxa de aumento do consumo e a relação entre os descontos oferecidos e usados.

REFERÊNCIAS

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. pages 487-499.

Aldenderfer, M. and Blashfield, R. (1984). *Cluster Analysis*. Number 07-044. Sage Publications, NewburyPark, California.

Bettencourt, L. A. and Gwinner, K. (1996). Customization of the service experience: the role of the frontline employee. *International Journal of Service Industry Management*, 7(2):3–20.

Blattberg, R., Buesing, T., Peacock, P., and Sen, S. (1978). Identifying deal prone segment. *Journal of Marketing Research (JMR)*, 15(3):369–377.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classiffication and Regression Trees*, volume 1. Chapman and Hall/CRC, New York.

Brijs, T., Swinnen, G., Vanhoof, K., and Wets, G. (2004). Building an association rules framework to improve product assortment decisions. *Data Mining and Knowledge Discovery*, 8(1):7–23.

Davies, D. and Bouldin, D. (1979). Cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 1(2):224–227.

Dibb, S. and Simkin, L. (1997). A program for implementing market segmentation. *Journal of Business & Industrial Marketing*, 12(1):51 – 65.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21:768–769.

Frank, R. E., Massy, W. F., and Boyd, H. W. (1967). Correlates of grocery product consumption rates. *Journal of Marketing Research (JMR)*, 4(2):184–190.

Green, P. E. and Wind, Y. (1973). *Multiattribute Decisions in Marketing*. Dryden Press.

Haley, R. I. (1968). Benefit segmentation: A decision-oriented research tool. *Journal of Marketing*, 32(3):30–35.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 1–12, Dallas, Texas, United States. ACM.

Helsen, K. and Green, P. E. (1991). A computational study of replicated clustering with an application to market-segmentation. *Decision Sciences*, 22(5):1124–1141.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119.

Keaveney, S. M. (1995). Customer switching behavior in service industries: An exploratory study. *The Journal of Marketing*, 59(2):71–82.

Kiang, M. Y., Hu, M. Y., and Fisher, D. M. (2006). An extended self-organizing map network for market segmentation–a telecommunication example. *Decision Support Systems*, 42(1):36–47.

Lazer, W. (1964). Lifestyle concepts and marketing. In *Toward scientific marketing*. American Marketing Association, Chicago.

Liu, D. and Shih, Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management*, 42(3):387–400.

MacQueen, J. B. (1967). Some methods for classification and analysis of MultiVariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. *Journal of Consumer Marketing*, 15(5):494–504.

Min, S. and Han, I. (2005). Detection of the customer time-variant pattern for improving recommender systems. *Expert Systems with Applications*, 28(2):189–199.

Mitchell, A. (1983). *The nine American lifestyles: Who we are and where we're going*. Warner, New York.

Morales, A. C. (2005). Giving firms an "E" for effort: Consumer responses to high-effort firms. *Journal of Consumer Research*, 31(4):806–812.

Morgan, J. N. and Sonquist, J. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415–434.

Ngai, E., Xiu, L., and Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2, Part 2):2592–2602.

Palmatier, R. W., Jarvis, C. B., Bechkoff, J. R., and Kardes, F. R. (2009). The role of customer gratitude in relationship marketing. *Journal of Marketing*, 73(5):1–18.

Piercy, N. and Morgan, N. (1993). Strategic and operational market segmentation: a managerialanalysis. *Journal of Strategic Marketing*, 1:123–140.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Quinlan, J. R. (1992). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc.

Rapid-I (2009). *RapidMiner 4.6 User Guide*. Rapid-I GmbH, Germany.

Roel, R. (1988). Direct marketing's 50 big ideas. *Direct Marketing*, 50:45–52.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 63:411–423.

Twedt, D. W. (1964). How important to marketing strategy is the "Heavy user"? *The Journal of Marketing*, 28(1):71–72.

Van den Poel, D., Schamphelaere, J. D., and Wets, G. (2004). Direct and indirect effects of retail promotions on sales and profits in the do-it-yourself market. *Expert Systems with Applications*, 27(1):53–62.

Zaki, M. J., Parthasarathy, S., Ogihara, M., and Li, W. (1997). New algorithms for fast discovery of association rules. Technical report, University of Rochester.

APÊNDICE

Tabela A1: Regras de associação.

Cluster 0						
Antecedente (x)	**Consequente (y)**	Conf.	*Lift*	s(x,y)	s(x)	s(y)
Doces sazonais	Chocolates	92%	1,19	26%	28%	73%
Tomates	Vegetais para a salada	86%	1,27	45%	53%	68%
Peras	Maças	84%	1,21	50%	61%	70%
Farinha	Açucar	81%	1,22	41%	51%	67%
Leites infantis	Fraldas	79%	5,65	2%	3%	14%
Fiambre fatiado	Queijo flamengo	79%	1,26	42%	54%	63%
Maçãs	Peras	73%	1,21	50%	70%	61%
Papel higiénico	Guardanapos	73%	1,20	43%	60%	61%
Guardanapos	Papel higiénico	72%	1,20	43%	61%	60%
Sumos/refrigerados	Legumes transformados	63%	1,55	2%	3%	40%
Cluster 1						
Antecedente (x)	**Consequente (y)**	Conf.	*Lift*	s(x,y)	s(x)	s(y)
Legumes secos	Massas	86%	1,48	6%	7%	58%
Corn Flakes	Leite UHT	83%	1,31	5%	6%	64%
Pão ralado	Especiarias secas	72%	1,94	3%	5%	28%
Refeições pré-preparadas	Vegetais em conserva	72%	1,83	2%	3%	39%
Cebolas	Ovos	64%	1,80	13%	20%	35%
Carne de coelho embalada	Carne de frango embalada	61%	2,59	2%	4%	23%
Barras de cereais	Bolachas doces	59%	1,59	4%	7%	37%
Boião de fruta	Iogurtes infantis	58%	3,23	3%	5%	18%
Insecticida	Acessórios de limpeza	54%	1,79	4%	7%	30%
Pastelaria/padaria	Natas/Chantilly	53%	2,06	2%	4%	25%
Cluster 2						
Antecedente (x)	**Consequente (y)**	Conf.	*Lift*	s(x,y)	s(x)	s(y)
Couves	Vegetais para a sopa	74%	1,79	21%	28%	41%
Condicionador de cabelo	Champô	72%	2,40	10%	13%	30%
Tomates	Vegetais para a salada	72%	2,04	18%	24%	35%
Polpas de tomate	Vegetais em conserva	62%	1,72	14%	22%	36%
Frutas tropicais	Bananas	61%	1,50	18%	29%	40%
Rolos de cozinha	Papel higiénico	58%	1,80	12%	21%	32%
Artigos de criança	Bolachas doces	57%	1,80	10%	18%	32%
Cebolas	Batatas	55%	2,25	11%	17%	24%
Farináceos	Açúcar	54%	1,70	14%	25%	36%
Fraldas	Toalhetes bébé	53%	4,76	4%	8%	11%
Cluster 3						
Antecedente (x)	**Consequente (y)**	Conf.	*Lift*	s(x,y)	s(x)	s(y)
Achocolatados	Leite UHT	85%	1,15	16%	19%	74%
Tortas	Bolos	78%	1,46	3%	4%	53%
Condicionador de cabelo	Champô	79%	1,76	17%	21%	45%
Vermutes	Cervejas com álcool	62%	1,66	2%	3%	38%
Frutas congeladas	Natas/cantilly	60%	1,51	3%	4%	40%
Fraldas	Iogurtes infantis	58%	2,18	7%	12%	27%
Esparregado	Legumes transformados	55%	1,78	2%	4%	31%
Carne de coelho embalada	Carne de bovino embalada	54%	1,87	3%	5%	29%
Frutos vermelhos	Legumes embalados	53%	1,88	3%	5%	29%
Frango do campo	Legumes aromáticos	53%	1,48	3%	6%	36%