3
Market Basket Analysis - MBA

Market basket analysis (MBA) ou, em português, *análise da cesta de compras*, é uma técnica de *data mining* que faz uso de regras de associação para identificar os hábitos de compra dos clientes, fornecendo uma visão da combinação de produtos dentro das cestas de compras dos clientes analisados. Conhecer o perfil de compra do público-alvo é muito importante para aumentar o potencial de recomendação, ajudando a melhorar as vendas.

O termo *cesta*, normalmente, se aplica a uma única ordem de compra. No entanto, a análise admite outras variações como, por exemplo, todas as compras de um determinado cliente. O caso clássico citado na literatura é o da rede de supermercados americana, WalMart. Foi descoberta a relação de compra entre uma marca de fraudas e uma marca de cerveja, quando as compras eram realizadas por homens nas sextas-feiras ao fim do dia. A análise da relação mostrou que as esposas pediam para os maridos trazer as fraudas para o fim-de-semana quando saíssem do trabalho e eles aproveitavam e levavam a cerveja para relaxar durante os dois dias sem trabalho. Tendo esse precioso conhecimento em mãos, o gerente decidiu colocar as fraudas perto das cervejas e com isso as vendas aumentaram ainda mais, pois os homens que não compravam cerveja passaram a comprar.

A entrada de dados para essa técnica é um conjunto de transações correspondendo a compras de vários clientes. As transações são representadas em uma tabela onde cada linha corresponde a uma venda e cada coluna a um produto adquirido. A cesta de compras de um cliente é composta por itens que foram adquiridos na mesma compra, ignorando-se a quantidade e o preço de cada item.

Apesar de a nomenclatura *market basket analysis* evocar uma imagem de carrinhos de compras e supermercados, cabe mencionar que existem outras áreas em que essa abordagem pode ser empregada como, por exemplo, análise de compras em cartão de crédito, análise de padrões em chamadas telefônicas, análise de compra de serviço telefônico, dentre outras aplicações.

PUC-Rio - Certificação Digital Nº 0621292/CA

3.1. Marketing de Relacionamento

No varejo, a maior parte das compras é realizada por impulso. A técnica de *Market Basket Analysis* dá pistas a cerca do que um cliente poderia ter comprado se alguma sugestão interessante lhe fosse feita.

Em um primeiro momento, pode-se considerar que esta técnica tem aplicações na reorganização da localização dos itens dentro de uma loja, bem como promoções para estimular a compra. Entretanto, é possível também comparar resultados entre filiais ou entre grupos de clientes em lugares demográficos distintos ou analisar as compras em função da época do ano. Assim, caso se observe que uma regra vale em uma loja, mas não prevalece em nenhuma outra, é porque há algo de interessante com esta que a torna tão diferente das outras. Investigar essas diferenças pode render algumas sugestões úteis para melhorar as vendas da companhia.

Marketing de relacionamento pode ser definido como "*uma estratégia de marketing que visa construir uma relação duradoura entre cliente e fornecedor, baseada em confiança, colaboração, compromisso, parceria, investimentos e benefícios mútuos*" [20] onde os clientes importantes precisam receber atenção contínua. É importante ter uma relação duradoura e de confiança com os consumidores, conhecendo-os a fim de cativá-los ainda mais, estimulando o consumo de mais e novos produtos e serviços com o objetivo da organização conquistar uma fatia maior do mercado.

Atualmente, as empresas estão se conscientizando da necessidade e da importância de intensificar o foco no marketing de relacionamento para crescer e se manter no mercado [21]. Administrar o relacionamento com o cliente ajuda a empresa a adquirir vantagem competitiva frente à concorrência. Apesar de muito investimento ainda ser feito na conquista de novos clientes, já existe uma percepção maior de que é possível melhorar a rentabilidade vendendo produtos e serviços para os clientes atuais.

O termo *cross-selling* é traduzido na literatura como "venda casada de produtos" e tem por definição a prática de vender produtos ou serviços adicionais para um cliente já conquistado, objetivando o aumento das vendas. Nesse contexto, a técnica de *market basket analysis* baseia-se na chave cliente-item para identificar a cesta de compras com os N itens que aparecem juntos mais frequentemente em transações. A partir do conhecimento das cestas mais frequentes, torna-se fácil partir para o *cross-selling*, sugerindo itens que possam ser de interesse de um cliente com determinado perfil de compra identificado.

PUC-Rio - Certificação Digital Nº 0621292/CA

Em resumo, para auxiliar nas tomadas de decisão em aplicações de marketing, o MBA é uma técnica poderosa que suporta a implementação de estratégias de *cross-seling*.

3.2. Regras de Associação

A análise de compras e registros de produtos tipicamente usa *regras de associação* representando os padrões de relacionamento encontrados entre os itens de dados do conjunto analisado. Em bases de dados onde os registros são transações, estas regras são conhecidas como *regras de associação transacionais*, enquanto no caso de bases de dados onde os registros representam clientes, contas, produtos, serviços e outros, as mesmas envolvem múltiplos atributos e, por isso, são chamadas de *regras de associação multidimensional*.

No âmbito da venda casada de produtos, ou *cross-selling,* as regras de associação permitem que uma loja possa recomendar o produto B para um cliente comprando o produto A, uma vez que ela conhece a regra, por exemplo, de que 30% dos seus clientes que compram A também compram B. Assim, o cliente é incentivado a comprar mais produtos, que eventualmente possam interessá-lo, baseado em características de consumo de compras anteriores no sistema,. Assim, não só a quantidade de vendas é maximizada, mas também a quantidade de vendas de determinado produto.

Regras de associação foram introduzidas em [25] da seguinte forma. Sejam $I = \{i_1, i_2, ..., i_m\}$ um conjunto de m itens distintos e D uma base de dados formada por um conjunto de transações, onde cada transação T é composta por um conjunto de itens tal que $T \subseteq I$. Uma regra de associação é uma expressão do tipo $X \rightarrow Y$ onde $X \subseteq I$, $Y \subseteq I$, $X \neq \emptyset$, $Y \neq \emptyset$ e $X \cap Y = \emptyset$, ou seja, tanto o antecedente (X), quanto o conseqüente (Y) de uma regra de associação podem ser formados por conjuntos contendo um ou mais itens, mas não podem ser vazios e são conjuntos independentes já que não possuem itens em comum. Por exemplo:

{Cerveja, Faldas} → {Leite}
{Cerveja, Leite} → {Fraldas}
{Fraldas, Leite} → {Cerveja}
{Cerveja} → {Fraldas, Leite}
{Leite} → {Cerveja, Fraldas}
{Fraldas} → {Cerveja, Leite}

PUC-Rio - Certificação Digital Nº 0621292/CA

A importância de uma regra de associação pode ser medida em termos de *suporte* e *confiança*. O *suporte* de uma regra determina com qual frequência uma regra é aplicada a um conjunto de dados, ou seja, a probabilidade do primeiro termo da implicação ser verdade. Já a *confiança* de uma regra determina o quão frequente os itens em *Y* aparecem nas transações que contém *X*. No exemplo anterior, a probabilidade de um cliente comprar cerveja e fraldas (ou seja, o termo *X* ser verdadeiro) é referida como o suporte, enquanto a probabilidade condicional de um cliente comprar leite, dado que comprou cerveja e fralda, é referida como a confiança.

O problema da mineração de regras de associação consiste em encontrar todas as regras de associação que possuam suporte e confiança maiores ou iguais, respectivamente, a um suporte mínimo (SupMin) e uma confiança mínima (ConfMin), especificados pelo usuário.

De fato, por muito tempo, a busca por regras de associação foi de interesse exclusivo de aplicações que lidassem com informação de cestos de compra (*market baskets*), o que levou esta técnica a ser comumente chamada de *market basket analysis* (MBA).

No entanto, podemos ver que regras de associação podem ser extraídas de qualquer base de dados onde existam relacionamentos implícitos entre os diferentes atributos. A utilidade destas regras não está somente na geração de novo conhecimento, mas também na confirmação de regras de negócio que são utilizadas, mas que nunca foram comprovadas. Sendo assim, regras de associação aplicam-se a diversas áreas de negócio como, por exemplo, estudo dos acessos a computadores, busca de novos clientes, recenseamento de população e análise de informação médica, dentre outros.

A descoberta de padrões através da análise das regras de associação é um importante para suporte à tomada de decisão. Com essas informações, um gestor pode identificar novas oportunidades de negócio, conhecer melhor seus clientes e seus perfis de compra, identificar produtos que influenciam na venda de outros, além de diversas outras informações que podem aumentar sua competitividade no mercado.

3.3. Algoritmos de MBA

Os primeiros algoritmos a serem utilizados na descoberta de regras de associação foram o *artificial immune system* (AIS) [23][24] e o *set-oriented mining* (SETM) [22]. Na maioria dos estudos mais recentes o algoritmo Apriori tem sido bastante utilizado, bem como suas variações. Nas subseções a seguir alguns desses algoritmos são explorados.

PUC-Rio - Certificação Digital Nº 0621292/CA

3.3.1.
Artificial Immune System

A técnica utilizada pelo *artificial immune system* (AIS) gera e conta conjuntos de itens à medida que são lidas as transações da base de dados. Para cada transação, determinam-se quais dos maiores conjuntos encontrados na transação anterior também se encontram na transação corrente e novos conjuntos de itens são gerados estendendo-se esses conjuntos com itens deste registro. A desvantagem é que este método gera e conta desnecessariamente conjuntos de itens que são considerados pequenos. A figura a seguir ilustra o seu funcionamento.

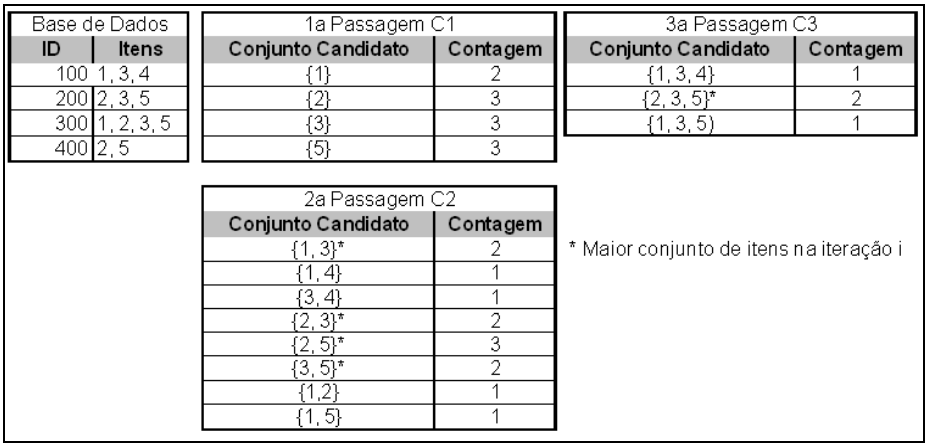

Figura 3 Exemplo de iteração do algoritmo AIS

A primeira passagem do algoritmo na base de dados faz a busca por cestas de apenas um item contando quantas vezes estes se repetem nas transações, ignorando aqueles que ocorrem apenas uma vez. A segunda passagem na base busca por cestas com dois itens, contando quantas vezes estes aparecem juntos nos registros varridos. Neste exemplo, a última passada encontra cestas de três itens, que são a maior cesta possível com frequencia maior que um. Assim, o algoritmo tem como solução o conjunto {2,3,5}.

3.3.2.
Set-oriented Mining

Assim como o AIS, o *set-oriented mining* (SETM) gera os conjuntos de itens candidatos no momento em que está passando pelas transações da base de dados. Entretanto, a contagem só é realizada no final da varredura da base. O identificador da transação é guardado junto com o conjunto candidato em uma estrutura sequencial. Ao fim da leitura de todos os registros, o tamanho do conjunto de itens mais frequente é determinado pela agregação da estrutura sequencial. Além de ter a mesma desvantagem do AIS, para cada conjunto candidato, existem vários valores de tamanho calculados. A figura a seguir ilustra os passos do algoritmo.

PUC-Rio - Certificação Digital Nº 0621292/CA

Base de Dados

ID	Itens
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

1a Passagem C1

Conjunto Candidato	Contagem
{1}	2
{2}	3
{3}	3
{5}	3

3a Passagem C3

Conjunto Candidato	ID
{1, 3, 4}	100
{2, 3, 5}*	200
{1, 3, 5)	300
{2, 3, 5}	300

2a Passagem C3

Conjunto Candidato	ID
{1, 3}	100
{1, 4}	100
{3, 4}	100
{2, 3}	200
{2, 5}	200
{3, 5}	200
{1, 2}	300
{1, 3}	300
{1, 5}	300
{2, 3}	300
{2, 5}	300
{3, 5}	300
{2, 5}	400

Figura 4 Exemplo de iteração do algoritmo SETM

No SETM, a primeira passagem na base de dados resulta na frequencia de ocorrência de cada item nas transações, assim como no exemplo anterior baseado no AIS. Entretanto, a segunda varredura na base tem como resultado uma lista com as combinações de pares de produtos onde cada combinação recebe o identificador da transação de onde foi extraída. Assim, teremos combinações repetidas com identificadores diferentes como a cesta {1,3} que neste exemplo ocorre na transação 100 e na 300. A última passagem na base revela as cestas de três itens e a solução encontrada é aquela que aparece em mais transações. Com isso, também temos o conjunto de itens {2,3,5} como resultado da aplicação do SETM nesse conjunto de registros.

3.3.3. Algoritmo Apriori

Este algoritmo foi a primeira ferramenta para descoberta de regras de associação em bases de dados com grandes volumes. Várias modificações foram propostas para melhorar sua eficiência e algoritmos similares foram propostos [26][27][28], introduzindo regras mais expressivas.

Primeiramente, o algoritmo identifica os itens que fazem parte de cada uma das transações. Em seguida, o algoritmo determina as regras de associação entre estes itens, selecionando as associações que ocorrem com mais frequência (ou maiores) no conjunto de transações em questão. O maior conjunto de itens candidato da passagem anterior é levado para próxima iteração, gerando outros conjuntos com tamanho maior que 1. Por fim, o algoritmo elimina os conjuntos gerados que tem um subconjunto que não é o maior.

Com base nos conjuntos de itens mais frequentes as regras de associação que atendem aos valores mínimos de suporte e confiança são geradas. Como resultado,

PUC-Rio - Certificação Digital Nº 0621292/CA

todos os conjuntos de itens frequentes são descobertos, produzindo todas as regras de associação que respeitam esses limites.

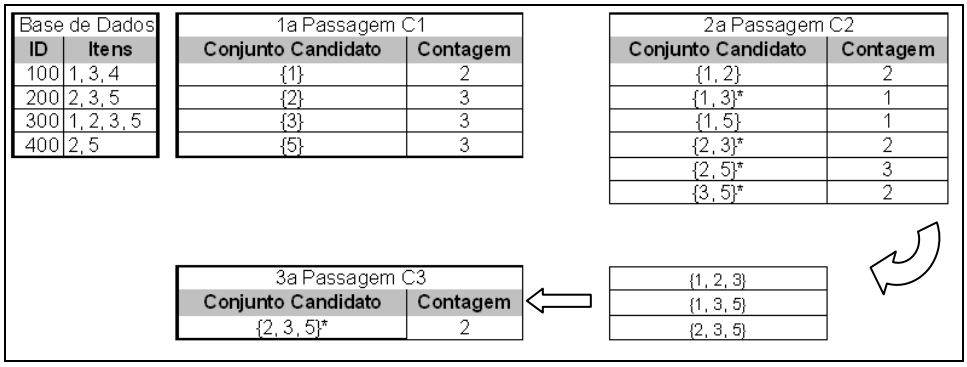

Figura 5 Exemplo de iteração do algoritmo Apriori

Assim como nos exemplos anteriores, a primeira passada do algoritmo Apriori na base de dados resulta na listagem da frequencia dos itens nas transações (cestas de um item apenas). Na passada seguinte, a lista gerada contem as cestas de dois itens com as respectivas frequencias. Destas cestas apenas as mais freqüentes serão consideradas e, assim, as cestas com três itens já são preparadas para o último passo, que irá somente então contabilizar as cestas mais frequentes relacionadas na etapa anterior. Com isso, o resultado obtido é a cesta de itens {2,3,5}.

3.4. Implementação baseada em Grafo

O objetivo da abordagem baseada em grafos é possibilitar a manipulação de informações de bases de dados de grandes volumes.

Nesta seção serão descritos os passos básicos do algoritmo baseado em grafos, que chamaremos de Graph-based Market Basket Analysis (GMB). Inicialmente serão introduzidos conceitos de grafos importantes para o entendimento do algoritmo. Em seguida, será abordado o princípio da implementação do algoritmo.

3.4.1. Conceitos de Grafos

Uma maneira de condensar as informações das relações entre itens é criar uma estrutura de grafo que as represente. Um grafo é definido por um par $G = (V,E)$, onde V é um conjunto finito de vértices e E representa um conjunto de arestas que ligam os vértices, de forma que $E \subseteq [V]^2$. No caso do *market basket analysis*, cada vértice representa um item e a aresta representa a relação encontrada entre dois itens. A frequência com que dois itens aparecem juntos em uma mesma transação, ou registro, é representada por valores associados às arestas, resultando em um grafo com

PUC-Rio - Certificação Digital Nº 0621292/CA

ponderação nas arestas. O objetivo desse tipo de representação é realizar a análise do ponto de vista das relações entre os itens e não da relação entre as transações como alguns estudos de associação abordam [18].

Definimos como $G_1=(V_1, E_1)$ um subgrafo de $G=(V,E)$ onde $V_1 \subseteq V$ e $E_1 \subseteq E$. O subgrafo G_1 é considerado *completo* se houver uma aresta para cada par de vértices, ou seja, cada vértice é adjacente a todos os outros vértices. Definimos como *clique* de *G* um subgrafo completo de G. O *clique máximo* é o maior clique encontrado no grafo, ou seja, nenhum outro clique o contém. Sendo assim, o *problema do clique máximo* consiste em determinar o tamanho do maior clique de G. Entende-se por tamanho a cardinalidade do conjunto de vértices do subgrafo em questão, sendo representado por $\omega(G)$.

Para exemplificar o problema do clique máximo, a figura a seguir apresenta um grafo *G*, onde $V=\{1,2,3,4,5\}$ e $E=\{(1,2),(1,3),(1,4),(1,5),(2,3),(3,4),(4,5)\}$. Os nós destacados em vermelho são os nós que formam o clique máximo desse conjunto de vértices, que possui tamanho 4.

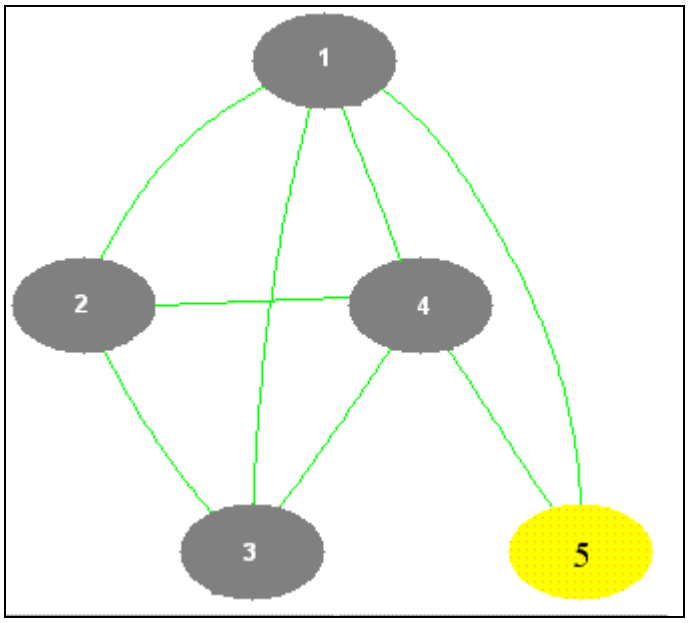

Figura 6 Exemplo de *clique* máximo

Se um peso for atribuído a cada aresta, então o subgrafo é conhecido como um *subgrafo ponderado* e o seu peso é definido pela soma dos pesos das arestas.

No escopo de *market basket analysis*, achar o clique máximo significa achar o conjunto de itens que mais aparecem juntos nas relações analisadas. Nas bases usadas, os registros não são transações e, por isso, o objetivo da aplicação do algoritmo é encontrar as associações existentes entre os atributos (variáveis) de cada registro. O

PUC-Rio - Certificação Digital Nº 0621292/CA

peso das arestas do clique ponderado indica a frequência com que as variáveis representadas pelos vértices ligados por ele aparecem juntas em um registro.

O GMB evita assim várias passadas na base de dados, típicas de algoritmos como o Apriori

Após o término da execução do GMB, o analista de negócios pode começar a por em prática a estratégia de *cross-selling,* pois tem em mãos as relações e as respectivas frequencias entre as variáveis selecionadas para estudo.

3.4.2. Montando a Matriz de Adjacências

O primeiro passo do algoritmo é montar a matriz de adjacências que irá representar o grafo ponderado, considerando apenas os atributos selecionados pelo usuário como o objetivo do estudo.

A matriz de adjacências foi implementada como uma lista encadeada de objetos representando os produtos (atributos do registro) da transação e cada produto possui a lista de relacionamentos com outros produtos e a quantidade de vezes que esses produtos aparecem juntos nos registros. Sendo assim, os atributos são os vértices do grafo e as relações entre os atributos são representadas pelas arestas que os ligam.

Para cada registro da base de dados, o algoritmo percorre atributo por atributo montando as relações entre os mesmos. Sendo assim, a cada atributo a_k, é feita a verificação se este já está representado na matriz de adjacências. Se não estiver, então uma nova entrada é criada para representá-lo. O próximo passo é então percorrer todos os atributos a_i restantes na transação, preenchendo a lista de relacionamentos entre esses produtos e, caso uma nova relação de a_k com a_i seja identificada, a_i é adicionado à lista de relações de a_k com contador de frequencia inicializado com 1. No caso dessa relação ter sido identificada em uma transação anterior, o contador é acrescido de 1.

3.4.3. Buscando a *Clique* de Tamanho Máximo

O grupo de interesse comum ou a cesta de produtos mais recorrente nas transações é representada pelo clique máximo do grafo. Em francês, a expressão *la clique* é definida como o grupo de indivíduos que partilham interesses em comum. Encontrar o clique de tamanho máximo significa encontrar o maior grupo de interesses em comum possível. Quando o grafo em questão é um grafo ponderado, o clique com maior peso corresponde ao grupo de interesses em comum que se repete mais frequentemente.

PUC-Rio - Certificação Digital Nº 0621292/CA

Na seção 3.4.2 foi apresentado como o conjunto de registros com seus atributos podem ser transformados em uma matriz de adjacências representando um grafo. Neste sentido, podemos dizer que o problema da cesta de compras pode ser transformado no problema do clique máximo (PCM).

Convém observar que encontrar o clique máximo em um grafo é um problema NP-difícil. A seguir descrevemos então um algoritmo simples para computar o clique máximo.

O clique máximo é inicializado como sendo o clique vazio e seu peso é definido com valor negativo. Além disso, uma lista que representa o clique "candidato" e seu respectivo peso também são inicializados da mesma maneira.

A cada passada *n* na lista de adjacências, o atributo a_n é adicionado ao clique candidato e, para cada atributo a_i restante, verifica-se a existência de relacionamento com todos os atributos já adicionados ao clique candidato. Se houver relacionamento entre os atributos, ou seja, se a_i está presente na lista de relacionamentos de todos os atributos incluídos no clique candidato, significa que são vértices ligados por uma aresta. Sendo assim, a_i também é adicionado ao clique candidato e o peso do mesmo é atualizado. Caso a_i não se relacione com algum dos outros nós, então a_i não faz parte do clique e o processamento passa para o próximo atributo.

Com o término do processamento do clique candidato, seu peso é comparado ao peso do clique máximo e, se for maior, então o clique candidato passa a ser a solução até que outro clique de maior peso seja encontrado ou até que todo vértice da lista de adjacências tenha sido visitado. As funções FindTheMaximalClique e Find_Clique transcritas no Apêndice A realizam esse processamento no protótipo desenvolvido.

PUC-Rio - Certificação Digital Nº 0621292/CA