Algoritmos e técnicas para a mineração de dados

Tobias EVANGELISTA, Alex Sandro Romeo de Souza POLETTO

Fundação Educacional do Município de Assis, Instituto Municipal de Ensino Superior de Assis, São Paulo-SP, Brasil

tobias-evangelista @hotmail.com, apoletto@femanet.com.br

Abstract: In recent years, as a result of the fall in the cost of data storage and the fastest automation of companies and public bodies, a large part of the operations and procedures performed, are recorded computationally and stored in large databases, This information, when accumulated, generate a large volume of data that is mostly unstructured data, that is, data that must be transformed into knowledge, Han (2006) refers to this situation as "rich in data, poor in information". In this research will be presented the steps for the accomplishment of the process of data mining, presenting concepts, techniques, tasks, exposing their particularities, utilities and needs, with emphasis on the model CRISP-DM (cross-Industry Standard Process of Data Mining) (LAROSE, D. T.) (HAND, MANNILA, SMYTH) due to its vast literature found and be the model of greater acceptance according to the ranking of use of the main processes for data mining (KDNuggets).

Keywords: Data mining, Data, algorithm, techniques.

Resumo: Nos últimos anos, em consequência da queda do custo do armazenamento dos dados e a rápida automatização das empresas e órgãos públicos, grande parte das operações e procedimentos realizados, são registradas computacionalmente e armazenadas em grandes bases de dados. Essas informações quando acumuladas, geram um grande volume de dados que em sua maioria são dados não-estruturados, ou seja, dados que deveriam ser transformados em conhecimento, passam despercebidos e são desperdiçados, Han (2006) refere-se a essa situação como "rico em dados, pobre em informação". Nesta pesquisa será apresentado as etapas para a realização do processo de mineração de dados, apresentando conceitos, técnicas, tarefas expondo suas particularidades, utilidades e necessidades, com ênfase no modelo CRISP-DM (Cross-Industry Standard Process of Data Mining) (LAROSE, D. T.) (HAND, MANNILA, SMYTH) devido a sua vasta literatura encontrada e ser o modelo de maior aceitação segundo o *ranking* de uso dos principais processos para mineração de dados(KDNuggets).

Palavras-chave: Mineração de dados, Dados, Algoritmo, Técnicas.

1. Introdução

“Devido a incapacidade do ser humano de interpretar tamanha quantidade de dados, muita informação e conhecimento, possivelmente uteis, podem estar sendo desperdiçados, ficando ocultos dentro das Bases de Dados espalhadas pelo mundo. Em consequência disso, a necessidade de se desenvolver novas ferramentas e técnicas de extração de conhecimento a partir de dados armazenados também vem crescendo e se mostrando cada vez mais indispensável. “(REZENDE, Solange, 2003, P.397).

Neste contexto, surge no final da década de oitenta a Mineração de Dados, “processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados” (FAYYAD et al. 1996), isto é, a exploração e a análise, de grandes quantidades de dados, a fim de descobrir padrões e regras significativos, tornando-se assim uma importante ferramenta de gerenciamento de informação, que deve revelar estruturas de conhecimento, que possam guiar decisões em condições de certeza limitada.

2. Mineração de dados

Existem dois objetivos fundamentais na Mineração de Dados que são: a predição e descrição. A primeira utiliza algumas variáveis que se encontram no banco de dados, com a finalidade de prever valores desconhecidos ou futuros de outras variáveis que sejam de interesse. A descrição, busca por padrões que descrevem os dados, de forma que possam ser interpretáveis pelos usuários (Silva, Gercely, 1972).

Por diversos autores a mineração de dados é considerada parte do processo de Descoberta de Conhecimento em Base de Dados (KDD - Knowledge Discovery in Databases) (Fayyad, Piatetsky-Shapiro e Smyth,1996), para outros mineração de dados e KDD são sinônimos (REZENDE, 2003) (Han et al, 2006). Segundo Berry e Linoff (1997, P. 5) Mineração de dados é a exploração e a análise, por meio automático ou semiautomático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativos.

Para nós, mineração de dados é um processo altamente cooperativo entre homens e maquinas que visa a exploração de grandes bancos de dados, com o objetivo de extrair conhecimento através do reconhecimento de padrões e relacionamentos entre variáveis, conhecimentos esses que possam ser obtidos por técnicas comprovadamente confiáveis e validados pela sua expressividade estatística (Cortes, Porcaro e Lifschitz, 2002).

Assim, entende-se que o objetivo principal da mineração de dados é descobrir relacionamentos entre dados, fornecendo auxilio para que seja feito previsões

de tendências futuras, baseadas no passado, desta forma, facilitar a tomada de decisões.

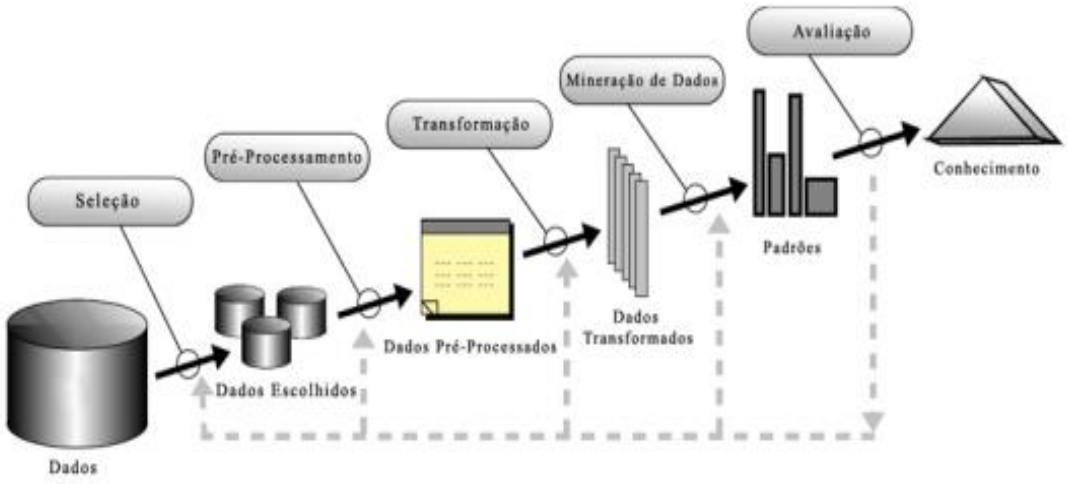

Figura 1: Figura representando o processo de KDD (Fayyad,1996).

3. Tarefas

As técnicas para realização do processo podem ser aplicadas a tarefas, usadas para solucionar um problema de descoberta que precisa ser solucionado.

3.1 Classificação

Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes, o objetivo é descobrir um relacionamento entre um atributo meta (cujo valor será previsto) e um conjunto de atributos de previsão. É usado para classificar pedidos de créditos, esclarecer pedidos de seguros fraudulentos ou até para identificar a melhor forma de tratamento de um paciente etc.

3.2 Estimativa (ou Regressão)

Usada para definir um valor para alguma variável contínua desconhecida, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais. É uma das tarefas mais utilizadas e importantes para o processo de mineração de dados, pode ser usado para estimar o número de filhos ou a renda total de uma família; estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de diagnósticos médicos; Prever a demanda de um consumidor para um novo produto etc.

3.3 Associação

A tarefa de associação consiste em identificar quais atributos estão relacionados. Apresentam a forma: SE atributo X ENTÃO atributo Y. É uma das tarefas mais conhecidas e utilizadas devido seus resultados obtidos, é bastante utilizado em “cestas de compras” para identificar que produtos são levados juntos pelos consumidores; determinar os casos onde um novo medicamento pode apresentar efeitos colaterais; identificar os usuários de planos que respondem bem a oferta de novos serviços etc.

3.4 Segmentação (ou Clustering)

A tarefa de agrupamento visa identificar e aproximar os registros similares. Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados. Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares. Costuma-se utilizar esta tarefa para agrupar clientes por região do país; agrupar clientes com comportamento de compra similar; agrupar seções de usuários Web para prever comportamento futuro de usuário etc.

3.5 Predição

A tarefa de predição é similar às tarefas de classificação e estimação, porém ela visa descobrir o valor futuro de um determinado atributo. Pode ser usada para predizer o vencedor do campeonato baseando-se na comparação das estatísticas dos times; predizer o percentual que será aumentado de tráfego na rede se a velocidade aumentar, entre outras necessidades.

3.6 Descrição

É a tarefa utilizada para descrever os padrões e tendências revelados pelos dados. A descrição geralmente oferece uma possível interpretação para os resultados obtidos. É utilizado para tabular o significado e desvios padrão para todos os itens de dados; derivar regras de síntese etc.

4. Técnicas

Não existe uma técnica que resolva todos os problemas da mineração de dados, várias técnicas servem para diferentes objetivos, com vantagens e desvantagens. A escolha da técnica esta intimamente ligada com o tipo de conhecimento que se deseja obter ou o tipo de dados em que ela será aplicada.

4.1 Árvores de Decisão

Árvore de decisão é um modelo preditivo que pode ser visualizado na forma de uma árvore, cada ramo da árvore é uma questão de classificação e cada folha é uma partição do conjunto de dados com sua classificação. A forma de realizar essa técnica é simples, utiliza-se um tipo de algoritmo de aprendizado de máquina baseado na abordagem de dividir para conquistar, um problema complexo é decomposto em subproblemas mais simples e recursivamente a mesma estratégia é aplicada a cada subproblema.

Uma árvore de decisão é um modelo de função discreta no qual é determinado o valor de uma variável, com base neste valor é executada alguma ação (Silva, Gercely,1972).

Para realizar a arvore de decisão é utilizado tarefas de classificação e associação, os algoritmos mais comuns utilizados são: Apriori e c4.5.

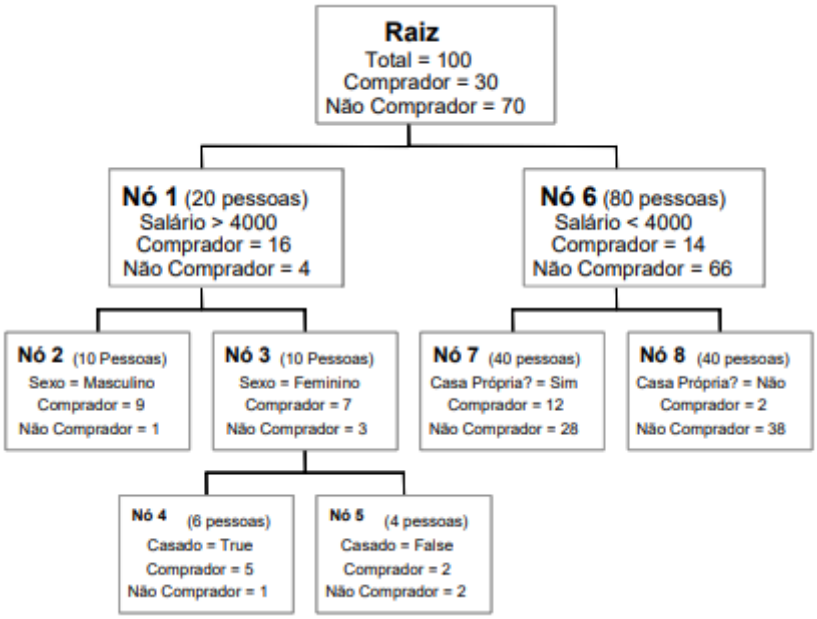

Figura 2: Exemplo de visualização de Arvore de decisão (Cortes, Porcaro, LIFSCHITZ, 2002)

4.1.1 Algoritmo Apriori

é o algoritmo mais utilizado para descobrir regras de associação. Para isto, o algoritmo Apriori executa múltiplas passagens sobre o banco de dados de transações, e é capaz de trabalhar com um número grande de atributos, obtendo como resultado, várias alternativas combinatórias entre eles, a partir da realização de buscas sucessivas em toda a base de dados e, apesar disso, os autores apontam o ótimo desempenho em termos de processamento desse algoritmo (LIBRELOTTO).

4.1.2 Algoritmo C4.5

Ele constrói árvores de decisão a partir de um dado conjunto de exemplos, sendo a árvore resultante usada para classificar amostras futuras. Após a construção de uma árvore de decisão é importante avaliá-la. Esta avaliação é realizada através da utilização de dados que não tenham sido usados no treinamento. Esta estratégia permite estimar como a árvore generaliza os dados e se adapta a novas situações, podendo, também, se estimar a proporção de erros e acertos ocorridos na construção da árvore.

Para que não haja a criação de todas as arvores de decisões possíveis, ele se fundamenta no atributo mais informativo, escolhido entre todos os atributos ainda não considerados no caminho até a raiz, o algoritmo seleciona como o mais informativo o que gerar maior ganho de informação, isto é, a resultante da diferença do valor da informação do atributo categórico e do valor da informação (entropia) do atributo em questão.

O algoritmo c4.5 é comumente utilizado, pois ele trabalha com valores indisponíveis, com valores contínuos, podar árvores de decisão e derivar regras.

4.2 Regras de indução

A regra de indução é uma técnica altamente automatizada e, considerada a melhor técnica para expor todas as possibilidades de padrões existentes em um banco de dados (Prass, Fernando).

A Regra de Indução consiste em uma expressão condicional do tipo: se [condição] então [consequência]. Após conseguir formular a regra, é construído uma tabela com o percentual de precisão (frequência em que a regra está correta) e de cobertura (frequência em que a regra pode ser aplicada), quanto maior o percentual, melhor é a regra.

É normalmente realizado tarefas de associação e classificação para a realização dessa técnica, seu algoritmo mais utilizado é o CN2 desenvolvido por Clark e Nibbet.

4.3 Redes Neurais

As Redes Neurais Artificiais são técnicas que procuram reproduzir de maneira simplificada as conexões do sistema biológico neural. Estruturalmente, consistem em um número de elementos interconectados, chamados neurônios, organizados em camadas que aprendem pela modificação de suas conexões. Normalmente, tem-se uma camada de entrada ligada a uma ou mais camadas intermediárias que são ligadas a uma camada de saída (BERRY e LINOFF, 1997).

A função básica de cada neurônio é avaliar valores de entrada, calcular o total para valores de entrada combinados, comparar o total com um valor limiar e

determinar o valor de saída, ou seja, a principal característica das redes neurais é dada pela capacidade de aprender com base na exposição de exemplos, até que a rede consiga aprender como resolver o problema, melhorando desta forma seu desempenho. Utiliza-se das tarefas de classificação e segmentação para a sua realização, seu algoritmo mais utilizado é o Backpropagation (realiza o aprendizado pela correção de erros).

4.3.1 Algoritmo Backpropagation

A vantagem principal de utilizar o algoritmo backpropagation é que ele trabalha com multicamadas e resolve problemas “não-linearmente separáveis” e alguns algoritmos não resolvem, ele se separa em duas fases:

O passo para frente (forward pass), onde nossas entradas são passadas através da rede e as previsões de saída obtidas (essa etapa também é conhecida como fase de propagação).

O passo para trás (backward pass), onde calculamos o gradiente da função de perda na camada final (ou seja, camada de previsão) da rede e usamos esse gradiente para aplicar recursivamente a regra da cadeia (chain rule) para atualizar os pesos em nossa rede (etapa também conhecida como fase de atualização de pesos ou retro-propagação).

4.4 Algoritmos Genéticos

Algoritmos genéticos são aqueles que simulam o processo de seleção natural proposto por Charles Darwin em 1859. De acordo com a teoria de Darwin, pode-se dizer que os organismos são equivalentes às estruturas de dados, enquanto os cromossomos são equivalentes às cadeias de bits, surgindo mais de um conjunto de considerações inteiramente diferentes que podem ser usados numa mesma solução do problema (Silva, Gercely). É muito difícil conseguir uma solução matemática ótima para um problema, porem pode-se encontrar soluções que sejam aceitáveis, os algoritmos genéticos são utilizados em casos em que o problema tenha variáveis e restrições. Os algoritmos genéticos utilizam tarefas de classificação e segmentação.

4.5 Classificação Bayesiana

É uma técnica estatística baseada no teorema de Thomas Bayes. Segundo o teorema de Bayes, é possível encontrar a probabilidade de um certo evento ocorrer, dada a probabilidade de um outro evento que já ocorreu. Seu uso mais comum é para estimar custos com um novo cliente etc.

5. Etapas do processo da extração de conhecimento de Bases de Dados

A extração de conhecimento a partir de grande quantidade de dados e vista como um processo interativo e iterativo, e não como um sistema de análise automática, sendo centrado na interação entre usuários, especialistas do domínio e responsáveis pela aplicação (REZENDE, 2003). Há inúmeros processos que definem e padronizam as fases e atividades da mineração de dados, porem todas contém a mesma estrutura. Fayyad, Piatetsky-Shapiro & Smyth (1996) propuseram inicialmente a divisão do processo em nove etapas, para Olson et al. (2008) consiste de seis etapas, apesar da diversidade no número de etapas todos são organizados de forma cíclica e unidirecional, podendo ir voltar entre as etapas. Nesta pesquisa será apresentado o processo CRISP-DM (Cross-Industry Standard Processo of Data Mining), segundo Olson et al. (2008) (apresentado por Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas Cássio Oliveira Camilo João Carlos da Silva).

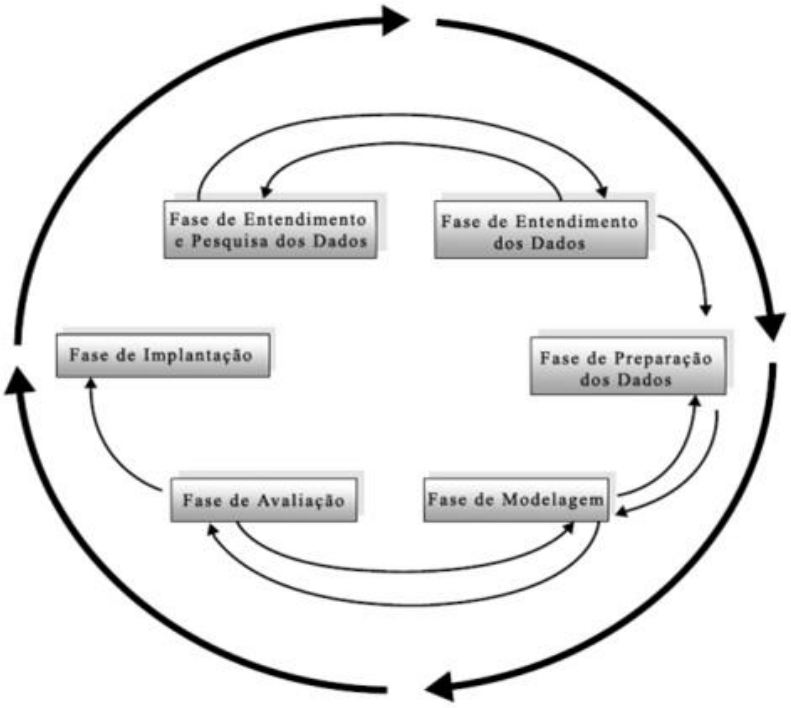

Figura 3: Figura representando o processo CRISP (Larose, D. T apud Camilo, 2009)

5.1 Entendimento da Camada de Negócio

O processo de mineração inicia-se com o entendimento da camada de negócio, isto é, o domínio da aplicação, considerando informações como o objetivo da aplicação e sua base de dados. Os conjuntos de dados resultantes dessa seleção são, então, pré-processados, ou seja, recebem um tratamento para poderem ser submetidos aos métodos e ferramentas na etapa de extração de padrões (REZENDE, 2003).

5.2 Entendimento dos dados

Segundo Olson *et al*. (2008) após definir os objetivos, é necessário conhecer os dados tendo em vista:

- Descrever de forma clara o problema;
- Identificar os dados relevantes para o problema em questão;
- Certificar-se de que as variáveis relevantes para o projeto não são interdependentes.

No final desta é utilizado as técnicas de agrupamento e exploração visual.

5.3 Pré-Processamento dos dados

Em razão da grande quantidade de origens e formatos que os dados podem ter, é necessário preparar os dados para que eles fiquem adequados para a mineração, transformando-os em apenas uma fonte de dados no formato atributo-valor (REZENDE,2003, P. 408). Nesta etapa envolve limpeza dos dados, Integração dos dados, reduzir dados e preencher valores vazios.

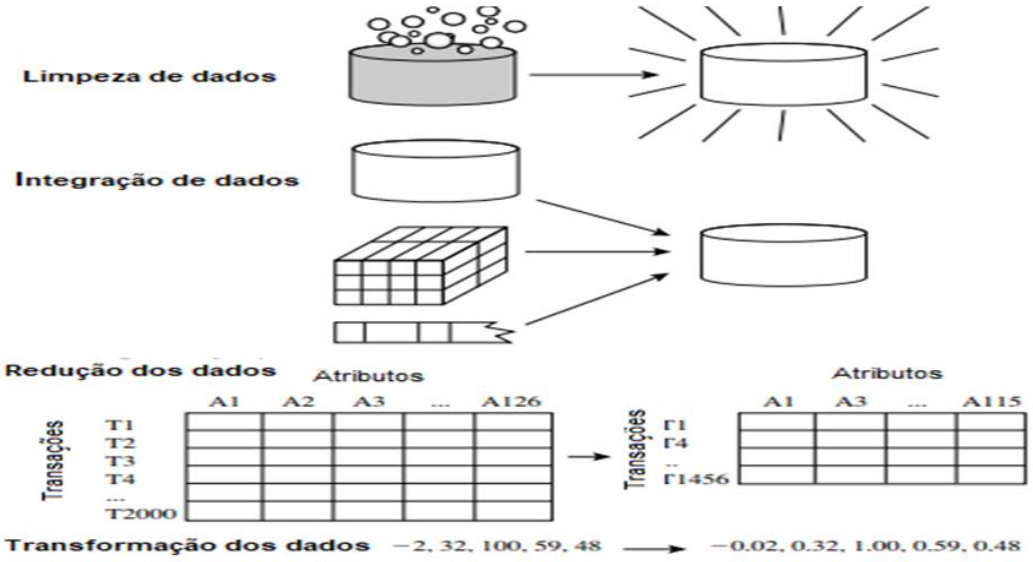

Figura 3 Etapa Pré-processamento de dados. Han, Kamber; 2006 apud (Camilo, Silva;2009)

5.3.1 Limpeza dos dados

Na maioria dos casos, os dados são encontrados com diversas inconsistências: registros incompletos, valores errados e dados inconsistentes. O objetivo dessa etapa é eliminar estes problemas para que eles não interfiram no resultado dos algoritmos utilizados (Camilo,2009).

As tarefas usadas nesta etapa vão desde a remoção do registro com problemas, passando pela atribuição de valores padrões, até a aplicação de tarefas de agrupamento para auxiliar na descoberta dos melhores valores.

5.3.2 Integração dos dados

Normalmente os dados que serão minerados são obtidos de diversas fontes: banco de dados, arquivos de texto, planilhas, vídeos, entre outras. Surge então, a necessidade de integrar esses dados, em uma única fonte consistente (Camilo,2009). Para essa etapa é necessário a analise aprofundada dos dados, observando redundâncias, dependência de variáveis e conflito de valores.

5.3.3 Transformação dos dados

Os algoritmos utilizados divergem na forma de trabalhar, alguns utilizam valores numéricos enquanto outros valores categóricos. Nesta etapa cabe transformar valores numéricos em categóricos ou categóricos em numéricos, dependendo da sua necessidade.

Para essa etapa dependendo do objetivo que se pretende alcançar é utilizadas diversas técnicas e tarefas para sua realização como: suavização (remove valores errados dos dados), agrupamento (agrupa valores em faixas sumarizadas), generalização (converte valores muito específicos para valores mais genéricos), normalização (colocar as variáveis em uma mesma escala) e a criação de novos atributos (gerados a partir de outros já existentes) (Camilo,2009).

5.3.4 Redução dos dados

A mineração de dados utiliza uma grande quantidade de dados, em alguns casos esses valores são tão grandes que torna o processo de análise impraticável. Nestes casos, são aplicadas técnicas e tarefas de redução de dados, assim diminuindo a massa original de dados em uma massa menor, sem perder a qualidade dos dados.

Essa etapa irá permitir que os algoritmos tenham maior eficiência, mantendo a qualidade do resultado. Nesta etapa se utiliza da criação de estruturas otimizadas para os dados (cubos de dados), seleção de um subconjunto dos atributos, redução da dimensionalidade e discretização. Dentre as diversas técnicas, a PCA - Principal Components Analysis (Smith, I. Lindsay apud Camilo, 2009), desempenha um papel muito importante na redução da dimensionalidade, outra técnica muito utilizada é a Discretização Baseada na Entropia (Han, Kamber apud Camilo, 2009).

5.4 Modelagem

Nesta etapa as técnicas (Algoritmos) de mineração de dados serão aplicadas. A escolha da(s) técnica(s) depende dos objetivos desejados.

Figura 4: Processo de comparação com algumas técnicas (MCCUE apud Camilo, Silva,2009)

5.5 Avaliação

Considerada uma etapa crítica do processo de mineração, nesta etapa é necessária a participação de especialistas nos dados, conhecedores do negócio e tomadores de decisão. Diversas ferramentas gráficas são utilizadas para a visualização e análise dos resultados (modelos) (Camilo e Silva,2009, P.5).

Nesta etapa os resultados dos modelos devem ser avaliados no contexto dos objetivos definidos na primeira etapa, frequentemente revertendo para as etapas anteriores do processo. Ganhar a compreensão do negócio é um procedimento iterativo em mineração de dados, onde os resultados de várias visualizações, estatísticas((matriz de confusão, índice de correção e incorreção de instâncias mineradas, estatística kappa, erro médio absoluto, erro relativo médio, precisão, F-measure) e ferramentas de inteligência artificial (cross validation, use training set, percentage split) (Camilo e Silva,2009, P.5) mostram ao usuário novas relações que fornecem uma compreensão mais profunda das operações organizacionais (Olson et al.,2008, P.10).

5.6 Distribuição

Após executado o modelo com os dados reais e completos é necessário que os envolvidos conheçam os resultados.

6 Considerações Finais

A Mineração de Dados está sendo utilizada pelas organizações de uma forma crescente, devido ao acúmulo de informações armazenadas nas bases de

dados, e a mineração de dados tem a possibilidade de obter conhecimento útil e interessante, o qual poderá ser utilizado como base concreta, auxiliando na tomada de decisão (). Ao utilizar o processo de mineração de dados, as empresas podem descrever características do passado, como também predizer tendências futuras, desta forma conseguir uma vantagem competitiva diferenciada das demais empresas.

O presente artigo apresentou uma introdução ao processo de mineração de dados, destacando suas tarefas e técnicas utilizadas durante o processo CRISP-DM (Cross-Industry Standard Processo of Data Mining), o mais utilizado para realizar a mineração de dados, este artigo teve caráter introdutório a cerca do tema, por ele ser um assunto complexo e extenso, que necessita de auxilio de muitos profissionais de diversas áreas, para quem busca um maior conhecimento sobre o assunto sugiro como leitura as referências bibliográficas utilizadas.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R. & SRIKANT, R. **Fast algorithms for mining association rules. Proc. of the 20th Int'l Conference on Very Large Databases**. Santiago, Chile, set. 1994. Disponível na Internet.

Baranauska, josé augusto; Monard, Maria Carolina. **Reviewing some machine learning concepts and methods.** São Carlos: USP,2000.

BERRY, M.J.A.; LINOFF, G. **Data mining techniques**. New York: John Wiley & Sons, Inc. 1997.

BERSON, Alex; SMITH, Stephen; THEARLING, Kurt. **Building Data Mining Applications** for CRM. USA, New York: MacGrawHill, 1999.

Camilo, Oliveira Cássio; Silva, Carlos da João. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.** Instituto de Informática Universidade Federal de Goiás-ufg,2009.

CORTÊS, Sérgio da C.; PORCARO, Rosa M.; LIFSCHITZ, Sérgio. **Mineração de Dados Funcionalidades, técnicas e abordagens**. 2002, 34f. Artigo-Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P**. From Data Mining to Knowledge Discovery in Databases**. American Association for Artificial Intelligence, 1996.

Garcia, C Simone; Alvares, O. Luis. **Árvores de decisão – algoritmos ID3 e C4.5.**

HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques.** Elsevier, 2006.

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

LIBRELOTTO, Solange Rubert ; MOZZAQUATRO, Patricia Mariotto2**. ANÁLISE DOS ALGORITMOS DE MINERAÇÃO J48 E APRIORI APLICADOS NA DETECÇÃO DE INDICADORES DA QUALIDADE DE VIDA E SAÚDE.** Revista interdisciplinar de ensino, pesquisa e extensão vol 1. N°103.

MCCUE, C**. Data Mining and Predictive Analysis - Intelligence Gathering and Crime Analysis**. Elsevier, 2007.

OLSON, D. L; DELEN, D. **Advanced Data Mining Techniques**. Springer, 2008.

Prass, S. Fernando. **KDD – UMA VISAL GERAL DO PROCESSO.**

REZENDE, Solange O. et al. **Mineração de dados. Sistemas inteligentes: fundamentos e aplicações,** v. 1, p. 307-335, 2003.

Romão, Wesley; Niederauer , A. P. Carlos; Martins, Alejandro; Tcholakian , Aran, Pacheco, C. S. Roberto; Barcia, M. Ricardo. **EXTRAÇÃO DE REGRAS DE ASSOCIAÇÃO EM C&T: O ALGORITMO APRIORI,** Programa de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina, Centro Tecnológico.

SILVA, Gercely da S**. Estudo de Técnicas e Utilização de Mineração de Dados em uma Base de Dados da Saúde Pública.** 8f. Resumo do Trabalho de Conclusão de Curso Universidade Luterana do Brasil, Canoas, 1972.