UMA EXPERIÊNCIA PROFISSIONAL DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS UTILIZANDO SISTEMAS HÍBRIDOS INTELIGENTES

Cybele Luzana Reis
cybele@cepel.br

Emmanuel Piseces Lopes Passos
emmanuel@ele.puc-rio.br

ICA : Núcleo de Pesquisa em Inteligência Computacional Aplicada
Departamento de Engenharia Elétrica
Rua Marquês de São Vicente, 225
CEP 22.453-900 Rio de Janeiro, R.J., Brasil
ICA@ele.puc-rio.br

RESUMO

Uma das principais etapas da descoberta de conhecimento em Banco de Dados (KDD - Knowledge Discovery & Datamining) é a mineração de dados (DataMining). O presente trabalho investiga o potencial de Sistemas Inteligentes na Descoberta de Conhecimento em bancos de dados, através da aplicação de Algoritmos Genéticos e Redes Neurais, analisando vantagens e desvantagens.

O projeto aqui descrito, apresenta a construção de um sistema híbrido inteligente - Redes Neurais, Algoritmos Genéticos, Estatística, Sistemas de visualização de dados, na solução dos principais problemas associados à descoberta de conhecimento em bancos de dados: classificação, identificação de padrões (regras) e grupos (clusters). O artigo apresenta, ainda, o processo de KDD, os modelos utilizados, os experimentos realizados utilizando um banco de dados real, e uma comparação entre os diversos modelos utilizados.

Palavras Chaves: Mineração de Dados, Descoberta de Conhecimento em Base de Dados, Redes Neurais, Algoritmos Genéticos, Inteligência Artificial.

1. INTRODUÇÃO

Grandes volumes de dados claramente ultrapassam os métodos manuais tradicionais de análise de dados. Tais métodos podem criar relatórios informativos dos dados, mas não podem analisar o

conteúdo destes relatórios para focalizar num conhecimento importante [PIAT91].

Surge portanto, uma grande necessidade em desenvolver novas técnicas e novas ferramentas que possam de forma inteligente e automática ajudar ao homem na análise desta "montanha" de dados para encontrar informações e conhecimentos úteis.

Com isso um grande interesse tem sido demonstrado pelo novo campo de pesquisa, o KDD (descoberta de conhecimentos em banco de dados), passando pelas etapas preparatórias até alcançar a mineração de dados.

Pode-se, portanto, definir KDD (Knowledge Discovery) como a descoberta de novos conhecimentos, sejam estes, padrões, tendências, associações, probabilidades, fatos, que não são óbvios ou fáceis de serem identificados e para isto, utilizam-se diversos tipos de algoritmos como rede neurais [HAYK94] [RUME86], algoritmos genéticos [GOLD89], estatística [ELDE96] e sistemas de visualização de dados [LEE95].

Os objetivos deste artigo são: modelar o problema de classificação e identificação de "clusters" em bancos de dados por Redes Neurais, bem como avaliar o desempenho do modelo Backpropagation em um banco de dados real. Ainda identifica um modelo de algoritmo genético dedicado à classificação e identificação de "clusters" em bancos de dados. E, finalmente, o de construir um sistema híbrido no qual o Algoritmo Genético, a partir das regras evoluídas, é capaz de gerar novos padrões de treinamento para as Redes Neurais, aumentando seu desempenho no aprendizado.

Na seção 2 é explicado o processo de KDD. Na seção 3 é explicada a técnica de redes neurais. Na seção 4 a de algoritmo genético. A seção 5 descreve os experimentos realizados, apresentando os modelos de rede neural e de algoritmo genético utilizados além de outros softwares para gerar regras. A seção 6 descreve a aplicação do Processo KDD na primeira fase do trabalho e a seção 7 a aplicação do Processo KDD na segunda fase do trabalho. E finalmente, na seção 8 são apresentadas as conclusões.

.

2. DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS (KDD)

O KDD (Knowledge Discovery and Data Mining) foi definido como sendo "um processo não trivial de extração de conhecimentos implícitos, previamente não conhecidos e potencialmente úteis, dos dados"**.** Essas descobertas são realizadas, usando-se técnicas não convencionais de consultas e análises da grande massa de dados .

Podemos então dizer que o processo de Descoberta de Conhecimento em Base de Dados (KDD) é um campo de pesquisa multi-disciplinar que inclui aprendizagem de máquina, estatística, tecnologia de banco de dados, sistemas especialistas, e visualização de dados.

O processo KDD parte de um banco de dados, seleciona os dados, gerando os dados alvo. Estes dados alvo, passam por uma série de tratamentos, como limpeza, pré-processamento, codificação, enriquecimento, gerando os dados transformados. Estes são submetidos a mineração dos dados, onde são encontrados padrões interessantes. Estes padrões são analisados e geram o conhecimento.

As fases do processo KDD, portanto, inicia com os dados brutos e finaliza com o conhecimento extraído que foi adquirido como resultado das seguintes etapas que basicamente compõem o processo de KDD [FAYY96]:

- Definição do Problema
- Seleção dos Dados
- Limpeza dos Dados

- Pré-processamento dos Dados
- Codificação dos Dados
- Enriquecimento dos Dados
- Mineração dos Dados
- Interpretação dos Resultados
- Relatório

Definição do problema - Para se realizar uma tarefa de KDD é necessário ter um amplo e claro entendimento do problema e dos objetivos, os quais se deseja alcançar. Portanto o problema deve ser cuidadosamente analisado para se fixar as metas à serem atingidas. Isto deve ser realizado visando direcionar as outras etapas para que não ocorra desperdício de recursos em busca de informações desnecessárias.

Seleção dos dados - Com o problema e o objetivo bem definidos o passo seguinte é analisar os dados existentes, sua estrutura, cobertura e qualidade. Escolher os dados necessários para atender as metas previamente fixadas. Se não existir dados que cubram este objetivo, então tais dados deverão ser coletados.

Limpeza de dados - Uma vez que o domínio do problema está perfeitamente delimitado, um trabalho sério é necessário para colocar os dados em forma para análise. Serão retiradas todas as informações inválidas, dados incorretos ou incompletos. Normalmente, esta etapa é realizada através de programas que analisam os dados através de faixas permitidas de valores por campos e críticas sobre os relacionamentos entre campos de registros. Por exemplo, se uma pergunta pode ter valores de zero(0) a dez(10) , e encontra-se valor onze (11) , nesta etapa retirar-se-á o dado incorreto. Certas informações julgadas desnecessárias são removidas. Também os dados são reconfigurados para assegurar um formato consistente.

Pré-processamento dos Dados - deve-se consolidar as informações, por exemplo, a criação de um campo necessário para o processo, ou , pode ser interessante agrupar vários campos num único.

Codificação dos Dados - é a etapa em que se deve ajustar os dados de maneira que eles possam ser utilizados pelo o algoritmo de aprendizado escolhido.

Enriquecimento dos Dados - consiste em conseguir de alguma forma agregar mais informações, que possam ser unidas aos registros existentes, enriquecendo os dados, para que estes contribuam no processo de descoberta de conhecimento.

Mineração dos Dados (Datamining) - nesta etapa devemos achar informações relevantes que estejam implícitas num banco de dados. É a fase de extração do conhecimento propriamente dita.

A primeira fase deste processo é definir o algoritmo que será utilizado. Na tabela 2.1 podemos verificar as técnicas mais utilizadas para este fim de acordo com a aplicação. A segunda fase consiste em implementar o algoritmo escolhido de acordo com o banco de dados. E a última fase é executar o algoritmo escolhido descobrindo conhecimento.

APLICAÇÃO	**ALGORITMO**
Associação	Estatística e Teoria dos Conjuntos
Classificação	Árvore de decisão e Redes Neurais
Clustering	Estatística e Redes Neurais
Modelagem	Regressão Linear ou não-Linear e Redes Neurais

Série Temporais	Modelo ARMA e Redes Neurais
Padrão Seqüenciais	Estatística e Teoria dos Conjuntos

Tabela 2.1 - Algoritmos mais utilizados em Datamining de acordo com a aplicação

Interpretação dos Dados - Deve-se verificar a validade do conhecimento obtido através do processo de mineração de dados. Para validade pode-se utilizar medidas percentuais, como ,por exemplo, a porcentagem de acerto de uma regra. A interpretação do conhecimento, normalmente, é feita por um especialista na área do banco de dados em questão.

Relatório - É a consolidação do conhecimento descoberto. Onde se incorpora este conhecimento na performance do sistema, documentação do conhecimento e em relatórios para as partes interessadas.

3. REDE NEURAL

Entre os algoritmos mais utilizados na fase de mineração de dados, do processo de KDD, está o algoritmo de Rede Neural, que pode ser empregado em diferentes aplicações, como mostra a tabela 2.2.

Rede Neural é uma técnica de máquina de aprendizagem, que foi inspirada na funcionalidade e estrutura do cérebro humano, imitando a forma de aprendizado e de reconhecimento de padrões.

Uma rede neural consiste em um conjunto de nós: nós de entrada que recebem os sinais de entrada; nós de saída que fornecem o sinal de saída e um ilimitado número de camadas intermediárias, contendo nós internos.

A representação do conhecimento de uma rede neural artificial está distribuída nas conexões e o aprendizado é feito diretamente do conjunto de dados, alterando-se os valores associados às conexões.

Existem várias arquiteturas de redes neurais, que utilizam diferentes ligações estratégicas para realizar uma tarefa.

A rede neural possui 2 estágios:

- O estágio de codificação - na qual a Rede Neural é treinada para realizar uma certa tarefa.
- O estágio de decodificação - na qual a Rede é usada para classificar exemplos, fazer previsões ou executar a tarefa que foi aprendida.

Duas das características mais importantes das redes neurais são as capacidades de generalização e de aprendizado. A generalização permite que essas estruturas sejam capazes de gerar saídas adequadas mesmo quando os valores apresentados são diferentes daqueles utilizados durante o treinamento.

As técnicas de aprendizado das redes neurais dividem-se em dois grandes grupos:

- aprendizado supervisionado - utiliza a apresentação de um conjunto de treinamento, formado pelo conjunto de vetores de entrada e pelo conjunto de saídas desejadas correspondentes. No processo de treinamento os padrões são apresentados repetidamente à rede até que o erro esteja abaixo de uma certa tolerância preestabelecida.
- aprendizado não-supervisionado - neste caso não há vetores de saída preestabelecidos no processo de treinamento. O próprio algoritmo de treinamento modifica os pesos da rede buscando formar diversas classes de vetores de entradas semelhantes.

4. ALGORITMO GENÉTICO

Outro algoritmo muito utilizado na fase de mineração de dados é o algoritmo genético, principalmente em classificação, clusterização e extração de regras.

A técnica de algoritmo genético foi baseada na reprodução genética e seleção natural. Pode-se, então, definir algoritmo genético como uma técnica de busca baseada na seleção natural e reprodução genética, que combina a sobrevivência do mais apto e troca aleatória de informações. [DAVI96]

Esta técnica tem como operações básicas a seleção (privilegia os indivíduos mais aptos), reprodução (indivíduos são reproduzidos com base na aptidão), crossover (troca de genes) e mutação (troca aleatória de gene).

Exemplo de aplicações existentes com esta técnica são a otimização de distribuição de gás, planejamento dos Jogos Olímpicos de Atlanta, aplicações de modelagem econômica , entre outras.

Uma definição importante para algoritmo genético é a de espaço de busca que consiste no conjunto de todas as possíveis configurações da estrutura que representa a solução do problema.

Um algoritmo genético é composto dos seguintes componentes:

- Problema
- Representação
- Decodificação
- Avaliação
- Operadores
- Técnicas
- Parâmetros

Os problemas tratados por esta técnica devem ser problemas complexos de otimização, que consistem em problemas com muitas variáveis, parâmetros e opções, grandes espaços de busca, com otimização combinatorial complicados, problemas mal estruturados com condições e/ou com restrições difíceis de serem modeladas matematicamente [DHAR97].

A estrutura da representação de um cromossoma deve ser orientada na estrutura do problema, deve descrever todo o espaço de busca relevante ao problema, permitir a expressão natural do problema e estar em sintonia com os operadores (crossover, mutação, etc.) . Um exemplo de representação é a binária, que consiste em representar um cromossoma através de 0 e 1.

Na decodificação deve-se construir uma solução para o problema a partir de uma solução, portanto, decodifica os cromossomas em soluções. Um exemplo é decodificar um cromossoma binário num número real.

Um outro componente de algoritmo genético é a avaliação que consiste no elo entre o algoritmo genético e o problema, então temos, f(cromossoma) = medida numérica de aptidão. A aptidão é extremamente importante porque as chance da seleção para reprodução são proporcionais a ela. O algoritmo genético tem como objetivo maximizar ou minimizar (dependendo do problema) a avaliação. Portanto, ele encontra o melhor cromossoma que realize esta tarefa.

Os operadores atuam no processo de criação de novos indivíduos (descendentes), os mais comuns são o crossover (combinação de sub-partes do indivíduos) e a mutação (variações para manter "agitada" a variedade genética). Para cada problema podemos criar operadores específicos ao domínio do problema. Um exemplo, é o operador de inversão.

Existem várias técnicas que podem ser utilizadas num algoritmo genético, por exemplo:

- Técnica de inicialização da população (normalmente a população é inicializada aleatoriamente)
- Técnicas de seleção (a técnica normalmente utilizada é a da roleta, a qual seleciona um indivíduo aleatoriamente proporcionando maiores chances aos indivíduos mais aptos)
- Técnicas de reprodução
- Técnicas de Eliminação da população antiga
- Interpolação de parâmetros
- Técnicas de elitismo (como manter a população antiga)
- Técnicas de seleção de operadores
- Técnicas de Normalização da aptidão

O último componente de um algoritmo genético é o conjunto de parâmetros que será utilizado. Como exemplo de parâmetros podemos citar: o tamanho da população, taxa de crossover, taxa de mutação, número de gerações, etc.

Uma característica importante de algoritmo genético é ser um processo estocástico (desempenho (respostas) vária a cada rodada) e ser robusto (bom desempenho em várias classes de problema). As principais limitações de algoritmo genético diz respeito a problemas que tem dificuldade na representação de uma solução (difícil encontrar um cromossoma) e aqueles que tem dificuldade de determinar uma função de avaliação.

5. EXPERIMENTOS

O trabalho aqui apresentado é uma aplicação da metodologia de Classificação em um banco de dados convencional de empresas (clientes atuais, suas filiais, ex-clientes e clientes potenciais) em 8 classes previamente definidas.

Este trabalho foi dividido em duas etapas. Na primeira os dados a serem classificados estavam em um banco de dados com 58.651 empresas. Destes dados 1.113 empresas já estavam classificadas em 8 “clusters”. E à partir desta classificação é que se devia classificar as outras empresas.

Estes 1.113 dados foram utilizados simultaneamente pelo algoritmo genético para se gerar regras e pela rede neural para gerar um modelo de classificação.

Na segunda etapa do trabalho utilizou-se 50 tabelas, contendo novos dados de 48.150 empresas. Destas devia se usar apenas 30.793, que eram dados de varejo. Entre estas 30.793 empresas estavam 4.072 que na fase anterior tinham sido classificadas, com um índice de acerto de 100%. Estas 4.072 empresas com os novos dados é que foram usadas para gerar novas regras, desta vez utilizando Algoritmo Genético e mais os softwares WIZRULE e WIZWHY. E foram também usadas para criar um novo modelo de rede neural. Esta segunda fase é descrita na seção 7. Na preparação dos dados foi utilizado o software Crystal Info, da Seagate Software.

5.1. MODELO DE REDE NEURAL UTILIZADO

O modelo de rede Neural utilizado foi treinado com a técnica de aprendizado supervisionado, utilizando o algoritmo backpropagation .

Para classificar os dados em 8 “cluster”, foram treinadas 8 redes, sendo que a saída de cada rede era o campo que representava um dos “clusters”.

Para a construção do modelo de rede neural foi usado o software PREDICT da empresa Neural

Works, que possui cinco fases:

- Seleção do conjunto de treinamento/teste
- Analise dos dados
- Criação do conjunto dos dados transformados
- Seleção das variáveis de entrada
- Treinamento da rede neural

Seleção do conjunto de treinamento/teste - O conjunto de treinamento e teste é definido previamente. No caso definiu-se que 70% dos dados seriam usados para o treinamento e os outros 30% para o teste.

Análise dos dados - O PREDICT faz uma análise dos dados, retirando os campos que não possui informação útil para o modelo.

Criação do conjunto dos dados transformados - Depois de analisar os dados o PREDICT cria uma ou mais transformação para muitos campos dependendo do nível da análise e do tipo do campo.

Seleção de variáveis - Se você tem um grande número de variáveis de entrada, existem muitos subgrupos que podem resolver bem o seu problema. Para procurar o melhor destes subgrupos que podem solucionar o problema o PREDICT utiliza Algoritmos Genéticos.

Algoritmo Genético é baseado na evolução biológica. O algoritmo Genético possui uma população de indivíduos que muda de uma geração para outra, normalmente combinando características de dois indivíduos “pais” criando o indivíduo “filho”. A cada indivíduo é medida uma aptidão e o conceito da “sobrevivência do mais adaptado” é implementada através da seleção dos pais mais adaptados mais freqüentemente que dos pais menos adaptados.

No caso da seleção de variáveis de entrada, os indivíduos são os conjuntos de variáveis de entrada. A aptidão de um indivíduo é medida pelo sucesso da rede que utiliza como entrada este conjunto de variáveis. O algoritmo começa com um pequeno conjunto de variáveis, e os grupos de variáveis bem sucedidos permanecem na população e são usados pelo algoritmo para selecionar conjuntos de variáveis maiores se necessário. O algoritmo dá preferencia a conjuntos de variáveis menores que a maiores.

Treinamento da rede neural - A rede Neural é construída incrementalmente pela adição de nós na camada escondida da rede. Cada nó escondido ou par de nós escondidos tem seus pesos treinados de várias inicializações. Cada inicialização é referida como sendo um novo candidato. O melhor candidato é estabelecido na rede, e então todos os pesos dos nó(s) de saída da rede são retreinados.

5.2. O MODELO DE ALGORITMO GENÉTICO UTILIZADO

Para gerar as regras utilizou-se o Algoritmo Genético desenvolvido no Laboratório do ICA do Departamento de Elétrica da PUC-RIO.

A modelagem do Algoritmo Genético consistiu fundamentalmente na definição de uma representação dos cromossomas, da função de avaliação e dos operadores genéticos. Em mineração de dados por regras de associação é necessário considerar-se atributos quantitativos e categóricos. Atributos quantitativos representam variáveis contínuas (faixa de valores) e atributos categóricos variáveis discretas. Na representação definida, cada cromossoma representa uma regra e cada gene corresponde a um atributo do BD, que pode ser quantitativo ou categórico conforme a aplicação

[AGRA93]. A função de avaliação associa um valor numérico à regra encontrada, refletindo assim uma medida da qualidade desta solução. O problema de Mineração de Dados com AG é um problema de otimização onde a função de avaliação possui a meta de encontrar as melhores regras de associação possíveis. Os operadores genéticos utilizados (*crossover* e mutação) buscam recombinar as cláusulas das regras, de modo a procurar obter novas regras com maior acurácia e abrangência dentre as já encontradas.

5.3. OUTROS SOFTWARES UTILIZADOS

Para gerar as regras utilizou-se também dois outros softwares chamados WIZRULE e WIZWHY, ambos da WizSoft Inc. Sendo que o software WIZRULE descreve o banco de dados, e o software WIZWHY gera as regras, do banco de dados. Este softwares foram utilizados apenas na segunda fase do projeto.

6. APLICAÇÃO DO PROCESSO DE KDD NA PRIMEIRA FASE DO TRABALHO

Definição do problema - Aplicação da metodologia de Classificação em um banco de dados convencional de empresas (clientes atuais, suas filiais, ex-clientes e clientes potenciais). As empresas deveriam ser classificadas em 8 grupos, de acordo com a classificação das 1.113 empresas que já estavam classificadas.

Seleção dos dados - Entre os dados existentes foram escolhidos aqueles que eram considerados relevantes ao processo. Foram selecionados 19 dos 103 campos do banco de dados.

Limpeza dos dados - Inicialmente, foi convertido os dados de formato Excel para um banco de dados em formato DBF. Foram eliminados, através de comandos SQL, os dados incompletos(com valores nulo) ou incorretos do banco de dados, visto através do software de visualização de dados, o Crystal Info.

Pré-processamento dos dados - Neste caso o pré-processamento foi utilizado para gerar os 8 campos de saída, gerando um campo para cada “cluster”, onde o valor do campo seria 1 onde a classificação do registro seria aquele “cluster” e 0 quando fosse outro “cluster” qualquer. Cada uma desta saída foi utilizada por uma da 8 redes treinadas.

Codificação dos dados - Em alguns campos numéricos, foi criado faixas de valores, como por exemplo, no campo receita, número de empregados e dispersão.

Mineração dos dados - Uma vez que os dados já foram transformados e preparados para a mineração dos dados em questão, inicio-se o processo de extração de conhecimentos.

Neste caso para gerar regras foi utilizado um software de Algoritmo Genético desenvolvido pelo laboratório do ICA. E utilizou-se o PREDICT para gerar um modelo de redes neurais para classificação.

- Experimentos Com Algoritmo Genético :

Para gerar as regras utilizou-se os dados do banco de dados com 1.113 registros, que já estavam classificados em 8 “clusters”. Várias regras foram conseguidas, e usadas para classificar o conjunto total dos dados, conseguindo classificar com índice de acerto de 100% 10.070 empresas.

- Experimentos com Rede Neural Algoritmo Backpropagation:

Foram realizados experimentos utilizando Rede Neural com o algoritmo de backpropagtion. Para o treinamento da rede neural utilizou um banco de dados com 1.113 registros, que já estava classificados em 8 “clusters”. Foram treinadas 8 redes, uma para cada “cluster”. Mas os resultados não foram muito satisfatório.

Então uma nova rede foi treinada, também com o algoritmo de backpropagation, mas os dados utilizados foram os 10.070, que foram classificadas em 8 (clusters) classes, via algoritmos genéticos, com 100% de certeza na classificação. Foram implementadas 8 redes, com 19 entradas, 20 neurônios na camada escondida e uma saída. Sendo que cada uma destas redes tinha como saída um dos “clusters”.

Para a construção do modelo de rede foi usado o software PREDICT.

Interpretação dos resultados - Os resultados obtidos com o modelo de rede neural treinada com dados das 1.113 não foram satisfatórios para classificar o restante dos dados, ou seja as 58.651 empresas. Pois ela conseguiu apenas 60% de acerto.

Com as regras conseguidas com o Algoritmo genético se conseguiu classificar :
10.070 empresas com índice de certeza de 100%, 843 empresas com índice de certeza de 90%, 3.124 empresas com índice de certeza de 80%, 5.523 empresas com índice de certeza de 75%, 7.011 empresas com índice de certeza de 60%, 3.077 empresas com índice de certeza de 56% e o restante 28.003 tinham menos de 56% de certeza.

O modelo de rede neural treinada com os dados das 10.070 empresas, que foram classificadas pelo Algoritmo Genético, com um índice de classificação de 100% de certeza, conseguiu um acerto de 90%. Este modelo então foi usado para classificar as empresas que com a utilização do Algoritmo Genético tiveram um índice de acerto menor que 75%. As empresas que com a utilização do Algoritmo Genético tiveram um índice de acerto maior de 75% permaneceram com a classificação dada pelo Algoritmo Genético.

7. APLICAÇÃO DO PROCESSO DE KDD NA SEGUNDA FASE DO TRABALHO

Definição do problema - Aplicação da metodologia de Classificação de Cadastro, em um banco de dados convencional de empresas (clientes atuais, suas filiais, ex-clientes e clientes potenciais) classificando essas empresas nos diversos segmentos de Telecomunicações, gerando, a partir de regras de classificação, a identificação de seus respectivos padrões de comportamentos (perfis), tendo em vista o oferecimento de soluções que mais se adeqüem às suas necessidades.

Seleção dos dados - A partir dos dados das 50 tabelas contendo informações diversas das empresas que deviam ser classificadas, criou-se um banco de dados, com os dados que julgou-se necessários para atender a meta previamente fixada, que era a de classificar estas empresas em diversos seguimentos. Escolhendo assim os campos necessários. Os registros deviam ser os do varejo, portanto dos 48 mil registros, apenas 30.793 foram selecionados.

Limpeza dos dados - Com os dados já selecionados, a tarefa seguinte era a de retirar as informações inválidas, dados incorretos ou incompletos. Através de comandos SQL, os dados incompletos foram substituídos por valores nulos.
Utilizando uma ferramenta de visualização de dados, o Crystal Info, comandos SQL, e uma ferramenta

de estatística, o SPSS, verificou-se que alguns campos possuíam apenas valores zeros, ou muitos "missings", ou apenas outro valor qualquer. Esses campos foram removidos do banco de dados. Além desses tinham outros campos julgados desnecessários que também foram removidos, como por exemplo o campo de dispersão, que foi julgado desnecessários.

Pré-processamento dos dados - Neste caso o pré-processamento foi utilizado apenas para gerar os 8 campos de saída, gerando um campo para cada "cluster", onde o valor do campo seria 1 onde a classificação do registro seria aquele "cluster" e 0 quando fosse outro "cluster" qualquer. Cada uma desta saída foi utilizada por uma da 8 redes treinadas.

Codificação dos dados - Alguns campos como o dos estados foram transformados para campos numéricos.

Enriquecimento dos dados - Algumas informações consideradas importantes para a classificação das empresas não estavam presentes nestes dados, como por exemplo a receita anual das empresas. Mas como tínhamos tais informações no banco de dados do trabalho da fase anterior, agregamos tais informações aos nossos dados.

Mineração dos dados - Uma vez os que dados já foram transformados e preparados para a mineração dos dados em questão, iniciou-se o processo de extração de conhecimentos.

Neste caso para gerar regras foi utilizado um software de Algoritmo Genético desenvolvido pelo laboratório do ICA, e dois outros softwares chamados WIZRULE e WIZWHY. Para a classificação por redes neurais foi usado o software PREDICT. O modelo de rede neural utilizado foi o Backpropagation.

- Experimentos com Algoritmo Genético e outros Softwares para a geração de Regras:

Para gerar as regras utilizou-se o Algoritmo Genético e os softwares WIZRULE e WIZWHY. Os dados utilizados foram os dados das 4.072 empresas que tinham sido classificadas na fase anterior do trabalho. Conseguiu-se descobrir 1.073 regras que classificaram com índice de acerto de 100%, 9.674 empresas. Entre essas empresas 1.555 empresas já faziam parte das empresas que na fase anterior foram classificadas com índice de acerto de 100%, portanto com estas regras pode-se classificar mais 8.119 empresas com índice de acerto também de 100%.

Portanto das 30.793 empresas que deviam ser classificadas, sobraram 18.602 empresas que foram classificadas pela tecnologia de Redes Neurais.

- Experimentos com Rede Neural Algoritmo Backpropagation:

Foram realizados experimentos utilizando Rede Neural com o algoritmo de backpropagtion. Para o treinamento da rede neural utilizou um banco de dados com 4.072 registros, que já tinham sido classificadas em 8 (clusters) classes, via algoritmos genéticos, e que tinham 100% de certeza da classificação. Para conseguir uma melhor classificação via redes neurais foram implementadas 9 redes, com 19 entradas, 50 neurônios na camada escondida e uma saída. Sendo que 8 destas redes era uma para cada "cluster", e uma rede cuja saída era 1 (um) quando o "cluster" era 3, 4 ou 6 e 0 (zero) quando o "cluster" era 1, 2, 5, 7 ou 8.

Para a construção do modelo de rede foi usado o software PREDICT.

Interpretação dos dados - Das 30.793 empresas que deviam ser classificadas, 4.072 empresas que tinham na classificação anterior um índice de 100% de acerto, foi mantida com esta classificação. Sobrando 26.721 empresas. Destas através de 1.073 regras descobertas por Algoritmo Genéticos foram "clusterizados" 9.674 empresas com índice de 100% de acerto. No entanto 1.555 empresas, já faziam

parte do banco de 4.072 empresas. Portanto apenas mais 8.119 empresas foram classificadas.

Finalmente, as 18.602 empresas restantes foram classificadas pela tecnologia de Redes Neurais. O modelo de Rede Neural utilizado para classificar estes dados foi treinado com os dados das 4.072 empresas, que tinham índice de acerto de 100%. O índice de acerto conseguido pela rede foi de 83,86%

8. CONCLUSÕES

Como vimos nem sempre o algoritmo escolhido é o melhor para se resolver o problema, ou não é totalmente bom, como no caso do algoritmo genético, que conseguiu com as regras descobertas classificar, uma boa parte dos dados com índice de certeza de 100%, mas não conseguiu classificar todos os dados. Sendo necessário a utilização do modelo de redes neurais para conseguir ter um bom índice de classificação nos dados restantes.

Mas podemos dizer que o método utilizando algoritmo genético para encontrar a melhor regra funciona corretamente conforme os resultados encontrados nos experimentos realizados. E que com as regras descobertas pode se classificar uma grande parte dos dados com um alto índice de acerto.

Com a utilização de Redes Neurais podemos também conseguir um bom índice de classificação.

Mas verificou-se que com a utilização conjunta de redes neurais e algoritmos genéticos pode se conseguiu um melhor índice de classificação para todos os dados.

BIBLIOGRAFIA

[ADRI97] Adriaans, P., Zantinge, D., "Data Mining**",** Addison-Wesley, 1996.

[AGRA93] Agrawal R, Imielinski T. and Swami A., "Mining Association rules between sets of itens in large databases". Proc. 1993 Int. Conf. Management of Data (SIGMOD-93), 207-216. May/93.

[AGRA94] Agrawal R, Srikant R., "Fast Algorithms for Mining Association Rules in Large Databases", VLDB-94, Santiago, Chile, Sept. 1994.

[AGRA95] Agrawal R, Mannila H., Srikant R, Toivonen H. and Verkamo A. I., "Fast Discovery of Association Rules, In Knowledge Discovery in Databases", Volume II, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), AAAI/MIT Press, 1995.

[AGRA97] Srikant R, Vu Q., Agrawal R, "Mining Association Rules with Item Constraints", Proc. Of 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.

[BIGU96] Bigus, J. P., "Data Mining with Neural Network - Solving Business Problems from Application Development to Decision Support", Mc Graw-Hill, 1996.

[BYTE95] Byte (USA), Volume 20, Number 10, October 1995.

[CRYS97] Crystal Info, Desktop User's Guide – Seagate Software, Version 6. Seagate Crystal Info. Canada, 1997.

[DHAR97] Dhar V., Stein R., "Seven Methods for Transforming Corporate Data into Business Intelligence", Prentice-Hall, 1997.

[DASG97] Dasgupta D., Michalewicz Z., "Evolutionary Algoritms in Engineering Applications", (Eds.) Springer, 1997.

[DATA96] "Database Programming & Design **"**(USA**)**, Special Edition - Data Mining, Volume 9,

Number 9, September 1996.
[DAVI91] Davis, L., "Handbook of Genetic Algoritms", Van Nostrand Reinhold, 1991.
[DAVI96] Davis, L., (Ed.) "Handbook of Genetic Algorithms". Int. Thomson Comp. Press 1996.
[ELDE96] Elder IV J. F. and Pregibon D., "A statistical perspective on knowledge discovery in data bases". In: U. M. Fayyad et al. (Ed.) Advances in Knowledge Discovery and Data Mining, 83-113. AAAI/MIT Press, 1996.
[FAYY96] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. - "Advances in Knowledge Discovery and Data Mining", AAAI Press, The MIT Press, 1996.
[GOLD89] Goldberg , D. E., "Genetic Algorithms in Search, Optimization and Machine Learning", Addison Wesley Publishing Company, Inc, 1989.
[HAYK94] Haykin, S., Neural Networks: "A Comprehensive Foundation", Macmillan College Publishing Company, New York, NY, 1994.
[JUNI97] Junior, V. R., Fukuda, F. H., Barros, J. L. A., Reis, C. L., "Busca de Conhecimento em Banco de Dados utilizando técnicas inteligentes", DEE, PUC-RIO, Anexo aos Anais do III Congresso Brasileiro de Redes Neurais, Florianópolis, julho de 1997.
[LANG95] Langlay P. and Simon H. A., "Applications of Machine Learning and Rule Induction". Comm. of the ACM 38(11), Nov./95, 55-64.
[LANG96] Langley P., "Elements of Machine Learning". Morgan Kaufmann, 1996.
[LEE 95] Lee H-Y., Ong H-L. and Quek L-H., "Exploiting visualization in knowledge discovery". Proc. 1st Int. Conf. Knowledge Discovery and Data Mining (KDD-95), 198-203. AAAI, 1995.
[PEI 95] Pei M., Goodman E. D., Punch W. F. III , 1995. "Pattern Discovery from data using Genetic Algorithms". Michigan State University's Center for Microbial Ecology and Beijing Natural Science Foundation of China. GARAGe – Genetic Algorithms Research and Applications Group.
[PIAT91] Piatetsky-Shapiro G., "Knowledge discovery in real databases: A report on the IJCAI-89 Workshop". AI Magazine, Vol. 11, No. 5, Jan. 1991, Special issue, 68-70.
[PRED96] Manual do PREDICT (predman) – NeuralWare, Inc. –USA, 1996.
[REIS97] Reis, C. L., "Busca de conhecimentos em Bases de Dados", Dissertação de Mestrado. Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 1997.
[RUME86] Rumelhart D. and McClelland. (Eds.) "Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge", MA: MIT Press, 1986.
[SHAF96] Shafer J., Agrawal R., Mehta M.: "SPRINT: A Scalable Parallel Classifier for Data Mining", VLDB-96, Bombay, India, September 1996.
[SHAV90] Shavlik J. W. and Diettrich. T. G., (Eds.) "Readings in Machine Learning". San Mateo, CA: Morgan Kaufmann, 1990.
[SRIK95] Srikant R. and Agrawal R.. "Mining generalized association rules". Proc. 21st Very Large Databases (VLDB) Conf., Zurich, Switzerland, 1995.
[ZURA92] Zurada, J. M., "Introduction to Artificial Neural Systems", West Publishing Company, 1992.
[WIZW96] WizWhy Help Contents – WizSoft Inc, 1996.
[WIZR97] WizRule Help Contents – WizSoft Inc, 1997.