

Busca de Conhecimento com Extração de Regras em Redes Neurais e Algoritmos Genéticos : Caso em Estudo

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

O artigo descreve a aplicação de técnicas inteligentes - Redes Neurais com Extração de Regras e Algoritmos Genéticos, na solução dos principais problemas associados à descoberta de conhecimento em bancos de dados: classificação, identificação de padrões (regras) e de grupos (clusters). O artigo apresenta os modelos utilizados, o estudo de caso com um banco de dados real, e uma comparação entre os resultados obtidos.

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.

Atualmente, utilizam-se diversos tipos de algoritmos para este fim, pode-se citar como exemplo rede neurais, algoritmos genéticos, estatística e sistemas de visualização de dados.

Os objetivos deste artigo são:

- modelar o problema de classificação e identificação de padrões (regras) em bancos de dados por Redes Neurais, avaliar o desempenho do modelo Backpropagation em um banco de dados real e extrair regras de uma rede neural treinada .
- identificar um modelo de algoritmo genético dedicado à classificação e identificação de padrões (regras) em bancos de dados através da evolução de regras de associação.
- desenvolver um sistema capaz de retirar conhecimentos de um banco de dados.
- aplicar o sistema desenvolvido em um banco de dados real, para retirar conhecimentos potencialmente úteis dos dados.

A seção 2 conceitua mineração de dados e apresenta as técnicas utilizadas na mineração: redes

neurais e algoritmo genético. A seção 3 faz uma descrição da extração de regras em redes neurais e apresenta o algoritmo de extração de regras utilizado. A seção 4 mostra como foram implementados o algoritmo genético, a rede neural e o algoritmo de extração de regras de redes neurais utilizados no sistema aqui desenvolvido. A seção 5 descreve os experimentos realizados. E finalmente, na seção 6 são apresentadas as conclusões.

2. Mineração de Dados

A mineração de Dados caracteriza-se pela existência do algoritmo que, diante da tarefa especificada, será capaz de extrair eficientemente conhecimento implícito e útil de um banco de dados. Pode-se dizer que a mineração de dados é a fase que transforma dados em informações.

Para efetuar a mineração de dados é necessário definir a técnica e o algoritmo a serem utilizados.

A

Tabela 2.1 mostra as principais aplicações na Descoberta de Conhecimento num Banco de Dados associadas a mineração e as técnicas mais utilizadas para mineração de dados.

Aplicação	Técnicas
Associação	Estatística e Teoria dos Conjuntos
Classificação	Algoritmos Genéticos, Redes Neurais e Árvores de Decisão
<i>Clustering</i>	Redes Neurais e Estatística
Previsão de Séries Temporais	Redes Neurais, Lógica Nebulosa e Estatística

Tabela 2.1: Algumas tarefas de KDD e suas técnicas de mineração de dados

As vantagens advindas do uso de Técnicas Inteligentes em Mineração de Dados são a capacidade de lidar com dados ruidosos, a maior tolerância a falhas, a facilidade de reconfiguração para executar diferentes atividades e a capacidade de generalização, entre outras.

Atualmente, existe um grande potencial para utilização de Técnicas de Inteligência Computacional em Mineração de Dados: Algoritmos Genéticos e Redes Neurais.

2.1 ALGORITMOS GENÉTICOS

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.

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).

Um algoritmo genético é composto dos seguintes componentes:

- Problema
- Representação
- Decodificação
- Avaliação

- Operadores
- Técnicas
- Parâmetros

O problema tratado por esta técnica, são em geral 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.

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.) .

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.

A avaliação consiste no elo entre o algoritmo genético e o problema, portanto tem-se, $f(\text{cromossoma}) = \text{medida numérica de aptidão}$. 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 pode-se criar operadores específicos ao domínio do problema.

Outro componente de um algoritmo genético são os parâmetros que serão utilizados. Como exemplo de parâmetros podemos citar: o tamanho da população, taxa de crossover, taxa de mutação, número de gerações, etc.

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

- Técnica de inicialização da população (normalmente a população é inicializada randomicamente)
- 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

Para interpretação dos dados utiliza-se porcentagem de validade da regra (quantidade de vezes que ela é verdadeira / quantidade total de registros que ela engloba) e sua abrangência (quanto registros estão de acordo com a regra).

Algoritmos Genéticos têm sido bem sucedidos em diversas áreas de aplicações, principalmente em problemas de otimização. Um problema de otimização caracteriza-se fundamentalmente em encontrar uma ou uma série de boas soluções entre um número muito grande de possíveis soluções (espaço de busca). Em mineração de dados Algoritmos Genéticos têm sido empregados na

classificação, clusterização e extração de regras.

2.2 REDES NEURAIIS

Existem vários tipos de Redes Neurais. Entre elas temos: Perceptron e Backpropagation.

A *rede Perceptron* consiste de uma rede simples com 2 camadas, sendo a camada de entrada e a camada de saída.

A *rede Backpropagation* possui camadas de entrada e saída e um conjunto de camadas intermediárias com nós escondidos. Os nós da entrada são completamente interconectados aos nós da camada escondida, e estes por sua vez aos nós de saídas.

Redes Neurais têm provado serem ferramentas extremamente potentes para a exploração de dados, com a capacidade de descobrir dependências e relações neste mesmo conjunto de dados que eram previamente desconhecidas, realizando bem a tarefa de classificação.

3. Extração de Regras em Redes Neurais

A utilização das redes neurais artificiais em diversos campos demonstram a grande capacidade do paradigma. As principais características que justificam este sucesso são [ANDR95]:

- A maneira direta na qual a rede neural adquire informação/conhecimento sobre um dado domínio do problema, na fase de treinamento.
- A forma compacta na qual a informação/conhecimento adquirido é armazenado dentro da rede treinada e a forma comparativamente fácil e rápida com que este conhecimento pode ser acessado e usado.
- A solução robusta da rede na presença de ruído nos dados de entrada.

Em adição a estas características, uma das mais importantes vantagens da rede é o alto grau de precisão quando uma solução da rede é usada para generalização sobre um conjunto de exemplos previamente desconhecidos do domínio do problema [THRU94].

No entanto, as redes neurais não possuem a capacidade de explicar de uma forma compreensível o processo pelo qual uma dada decisão ou saída gerada pela rede foi atingida, sendo, por isso, consideradas como “caixas pretas” por alguns usuários, que as utilizam, especialmente quando estes não estão familiarizados com os métodos matemáticos empregados em tais sistemas, o que acarreta uma falta de confiança, dos usuários, nos resultados obtidos pelas redes neurais.

Para que se aumente a credibilidade e aceitabilidade das redes neurais artificiais pelos usuários e que estas tenham completa utilidade como ferramentas de aprendizagem e de generalização, a capacidade de explicação deve fazer parte da funcionalidade das redes. Deve-se fornecer ao usuário a capacidade de determinar o conjunto de condições sobre as quais uma unidade de saída da rede está ativa ou não, fornecendo-lhe algum grau de transparência da solução da rede.

3.1. ALGORITMOS DE EXTRAÇÃO DE REGRAS

A maioria das Redes Neurais representam o conhecimento através de um processamento numérico dos pesos e neurônios, de uma forma distribuída e não nítida, tornando-se difícil a extração de conhecimentos e explicações [LOPE93]. Alguns pesquisadores na área de Inteligência Artificial

[TOWE93] e [GEVA95] sugerem que, sem alguma forma de explanação, não se pode explorar todos os recursos das redes neurais.

A aprendizagem e representação de relações e funções não lineares de uma rede neural estão codificadas, de uma forma a princípio não compreensível, nos vetores de pesos. Algoritmos de extração de regras buscam explorar a forma de representação destes pesos, estabelecendo critérios de validação do treinamento de uma rede neural. Esta análise considera três características onde uma rede neural treinada “codifica” um conhecimento específico [GEVA95]:

- a - Na própria topologia da rede;
- b - Na função de ativação associada aos neurônios da camada intermediária e de saída;
- c - No conjunto de pesos que interligam os neurônios.

A tarefa de extração do conhecimento, de uma rede neural, consiste na interpretação da forma de cooperação existente nas características (a), (b) e (c) de um treinamento.

Vários estudos foram feitos sobre extração de regras. Dentre os algoritmos já propostos para a extração de conhecimento de rede neural estão os denominados “EN” em [BATI97], “SUBSET” e “MofN” em [TOWE93], “RULEX” em [GEVA95], “VIA” em [THRU94], o “LOGIC” e o “PERCEP” em [OLIV97].

Neste trabalho utilizou-se o Algoritmo LOGIC.

3.1.1. ALGORITMO LOGIC [OLIV97]

O algoritmo LOGIC é uma ferramenta de análise de sensibilidade de redes neurais. Está relacionado com a transformação dos neurônios das topologias implementadas em portas lógicas booleanas, sendo cada neurônio da rede feedforward representado por um conjunto de portas lógicas AND (E) e OR (OU), definidas com três estados especiais de entrada. Cada porta lógica AND recebe todas as entradas da rede e cada entrada destas portas representa uma conexão da matriz de pesos da rede. Estas entradas podem assumir três estados especiais denominados : DC - “Diretamente Conectado”, IC - “Inversamente Conectado” ou NI - “Não Importa”. Estes estados surgem naturalmente dos valores de treinamento dos pesos da rede neural após a convergência de treinamento.

Como resultado final da aplicação do LOGIC a uma estrutura treinada de redes neurais, pode-se obter um conjunto de regras do tipo SE ... ENTÃO. Para a obtenção desta estrutura de regra, deve-se associar a cada entrada das portas lógicas AND uma hipótese e a cada saída das portas OU uma conclusão (ou classe). O uso da condição “NOT” estará associado à condição “Inversamente Conectado” das portas lógicas.

4. IMPLEMENTAÇÃO

4.1. IMPLEMENTAÇÃO DO MODELO DE ALGORITMO GENÉTICO

O problema, neste caso, consiste em encontrar a melhor e mais abrangente regra de produção (SE .. ENTÃO) que responda a uma determinada pergunta sobre um banco de dados. Por exemplo, podemos citar perguntas como:

- Quais são as características dos meninos e meninas que foram morar na rua ?
- O que diferencia os meninos e meninas que foram morar na rua, daqueles que continuam

convivendo com suas famílias ?

A representação escolhida foi a de um **gene** para cada campo do banco de dados que não faça parte da resposta da pergunta, estes podendo receber valores que variam do valor mínimo ao máximo do campo em questão. Foi criado um novo valor para cada um dos campo que foi chamado de "don't care", que consiste em dizer que todos os valores possíveis, para o campo específico, valem para a regra. Com a criação do campo don't care deve-se tomar cuidado para não deixar que todos os genes do cromossoma assumam este valor.

Neste caso é importante, para avaliação, levar em conta quantas vezes a regra é verdadeira, portanto, foi proposto dois tipos de avaliação que dependem do problema.

Nos problemas que buscam características de grupo de dados no banco de dados, foi utilizado o seguinte algoritmo para a avaliação:

1. Avaliação do cromossoma = 0
2. Se a regra descrita no cromossoma está correta e a sua resposta está certa para um registro (tupla) do banco de dados então soma-se um ponto na avaliação Senão não se soma ponto.
3. Se o banco de dados não terminou volte ao passo 2.

No caso de problemas que buscam associação entre grupos de dados, verificou-se que a avaliação acima não era suficiente, portanto, foi criada uma avaliação que penalizasse as regras existentes nos grupos de dados, para tal , propõe-se o seguinte algoritmo:

1. Avaliação do cromossoma = 0
2. Se a regra descrita no cromossoma está correta e a sua resposta está certa para um registro (tupla) do banco de dados então soma-se um ponto e pula para o passo 5.
3. Se a regra descrita no cromossoma está correta e a sua resposta está errada para um registro (tupla) do banco de dados então subtrai-se um ponto e pula para o passo 5.
4. Se a regra está errada para um determinado registro então não se atribui ponto.
5. Se o banco de dados não terminou volte ao passo 2.

Os operadores utilizados, neste trabalho, foram o **crossover** de um ponto [DAVI91], a **mutação** [DAVI91] e a **mutação** forçando **don't care**.

O crossover de um ponto combinará sub-partes de dois indivíduos para gerar dois novos filhos a partir de um ponto de corte escolhido aleatoriamente.

Na mutação escolhe-se uma posição do cromossoma de maneira aleatória e um novo valor para esta posição, aleatoriamente, de acordo com seus limites.

Na mutação forçando don't care escolhe-se uma posição do cromossoma de maneira aleatória e nesta posição é colocado o valor do don't care .

Os parâmetros utilizados foram a taxa de crossover, taxa de mutação, taxa mutação forçando don't care, número de indivíduos em uma geração e total de indivíduos. A taxa de crossover é a porcentagem em que crossover será realizado, aqueles indivíduos que não farão crossover terão seus pais repetidos na próxima geração. A taxa de mutação é a porcentagem que será realizada a mutação, os indivíduos que não farão mutação serão repetidos. A taxa de mutação forçando don't care consiste na porcentagem das mutações as quais será colocado o valor don't care, por exemplo, se esta taxa for 0,5 então serão, aproximadamente, 50% das mutações realizadas forçando o don't care e o resto mutação normal.

As técnicas utilizadas foram a de elitismo (repetir o melhor indivíduo da geração anterior na próxima) e a normalização linear(criar aptidões partindo de um valor constante, crescendo linearmente, a partir dos cromossomas em ordem decrescente de avaliação). Para a seleção foi utilizado o método da

roleta.

Para interpretação dos dados utiliza-se porcentagem de validade da regra (quantidade de vezes que ela é verdadeira / quantidade total de registros que ela engloba) e sua abrangência (quanto registros estão de acordo com a regra).

O algoritmo utilizado [JUNI97] foi desenvolvido na linguagem Delphi 2.0, no Laboratório do ICA do Departamento de Elétrica da PUC-RIO.

4.2. IMPLEMENTAÇÃO DA REDE NEURAL

Neste trabalho, as redes neurais foram treinadas com técnica de aprendizado supervisionado, utilizando o algoritmo “backpropagation”.

A topologia de rede aqui utilizada foi a do tipo “feedforward” totalmente interconectada. Todos os modelos de rede utilizados foram de três camadas. Uma camada de entrada, uma camada escondida contendo dois neurônios, e uma camada de saída com um neurônio.

O treinamento foi realizado utilizando a ferramenta do Matlab 4.2. Em alguns testes utilizou-se a função **trainbpm** que treina a rede com até três camadas. [NEUR94]

Em outros testes, utilizou-se a função **trainlm** que utiliza a Otimização Levenberg-Marquardt para fazer com que o tempo de treinamento seja o menor possível. [NEUR94]

A função de avaliação utilizada foi a função tangente hiperbólica.

Entre os diversos experimentos realizados, variou-se a topologia da rede, as entradas utilizadas, e os valores dos parâmetros da rede (como a taxa de aprendizado e a taxa de mutação).

Para a aplicação do algoritmo LOGIC foi utilizada as matrizes de pesos e os vetores de “bias”, resultantes do treinamento de uma rede neural, onde as entradas e os estados de ativação dos neurônios possuíam valores [-1 e +1].

5. Experimento realizado

A população das grandes cidades brasileiras vem assistindo, desde a década de 80, a ocupação das ruas por uma grande quantidade de crianças e adolescentes que buscam, aí, meios de sobrevivência. Embora inúmeros trabalhos vem sendo realizados no intuito de solucionar o problema desses grupos, que passaram a ser chamados de “meninos e meninas de rua”, modestos são os resultados até então obtidos, tornando indispensável a produção de novos conhecimentos que possam orientar as práticas e políticas públicas dirigidas à ressocialização dos meninos e meninas de rua. [ALVE94]

Este trabalho, portanto, desenvolve um sistema de descoberta de conhecimento em banco de dados, usado para descobrir fatos importantes, sobre os meninos e meninas que vivem nas ruas, fatos ainda não conhecidos, que possa de algum modo contribuir na solução deste problema.

Para atender tal objetivo foi necessário colher dados reais, para formação do banco de dados, obtidos por meio de questionários, respondidos pelos próprios meninos. Como a intenção era de **buscar fatos sobre a vida dos meninos e meninas** que vivem nas ruas, procurando **descobrir os motivos** que os levaram a optar por esta vida, decidiu-se também para fins de comparação, obter dados de meninos e meninas que continuam vinculados às suas famílias, e também às escolas e/ ou a um trabalho, para verificar se existe alguns fatos importantes que fazem com que meninos e meninas, de mesma idade, mesma situação financeira e familiar optam por viverem nas ruas, enquanto outras não. Pois se conseguindo identificar estes motivos, pode-se tentar solucionar o problema, evitando assim que tantas

crianças continuem a ir viver nas ruas.

Esses fatos foram, então, organizados e estruturados em forma de regras de produção (forma de representação do conhecimento) dando-nos um tipo de "conhecimento" sobre os "meninos e meninas de rua".

Foram colhidos 82 dados de meninas e meninos que vivem nas ruas e mais 211 dados de meninas e meninos que vivem com suas famílias. A faixa etária pesquisada foi de 8 á 18 anos.

Analisado o problema dos meninos e meninas de rua, fixou-se, como metas para este estudo, responder as seguintes perguntas:

- Quais são as características dos meninos e meninas que foram morar na rua ?
- O que diferencia os meninos e meninas que foram morar na rua, daqueles que continuam convivendo com suas famílias ?
- Quais as características dos meninos que não moram na rua ?
- O que diferencia o menino que continua vivendo com sua família, do menino que vai morar na rua?

Para realização deste experimentos foi utilizado o banco de dados contendo oitenta e nove (89) campos e noventa e três (293) registros.

Na escolha das informações necessárias para atender às metas fixadas, foi selecionado 20 dos 89 campos do banco de dados, foram escolhidos os campos que eram considerados relevantes ao processo. Para responder cada pergunta utilizou-se os dados relevantes ao problema. Nas perguntas que buscam características utiliza-se os dados de quem se deseja as características. Nas perguntas associativas utiliza-se os dados que se deseja associar.

Estes dados foram então preparados (limpos, processados e codificados) para serem utilizados em cada um dos processos de mineração de dados escolhidos.

5.1. MINERAÇÃO DOS DADOS

Neste caso os algoritmos escolhidos foram Algoritmo Genético e Redes Neurais.

5.1.1. Algoritmos Genéticos

Com o objetivo de encontrar a melhor e mais abrangente regra de produção para responder as questões propostas, utilizou-se o algoritmo genético com os seguintes parametros: população de 50 indivíduos por geração, 0,6 para taxa de crossover, 0,05 para taxa de mutação e 0,8 para mutação forçando don't care (em relação ao percentual da mutação). A população foi inicializada com valores do banco de dados.

Os resultados encontrados para as perguntas propostas estão descritos abaixo:

Quais as características dos meninos que vão morar na rua ?

- **Responsáveis trabalhavam.** Enquadram-se 73 de 82, portanto 89%. Avaliação: 73.
- **Apanhavam dos pais.** Enquadram-se 59 de 82, portanto 71.95%. Avaliação: 59.

O que diferencia o menino que vai morar na rua, do que continua vivendo com sua família?

- **Pais usavam drogas.** Enquadram-se 17 de 82, portanto 20.73%. Das pessoas que se enquadram

nesta regra 19, 89.47 % foram morar na rua.

- **Responsáveis trabalhavam, pais usavam drogas.** Enquadram-se 15 de 82, portanto 18.29%. Das pessoas que se enquadram nesta regra 16, 93.75 % foram morar na rua.
- **Responsáveis trabalhavam, pais usavam drogas e trabalhavam.** Enquadram-se 10 de 82, portanto 12.2%. Das pessoas que se enquadram nesta regra 11, 90.9% foram morar na rua.

Quais as características dos meninos que não moram na rua ?

- **Não apanha da policia.** Enquadram-se 207 de 211, portanto 98,1%. Avaliação:207.
- **Responsáveis trabalham, estuda, não apanha da policia, não trabalha.** Enquadram-se 163 de 211, portanto 77,25%. Avaliação:163.

O que diferencia o menino que continua vivendo com sua família, do menino que vai morar na rua?

- **Não saiu de casa, estuda, não apanha da policia.** Enquadram-se 204 de 211, portanto 96,68%. Do total de 293 pessoas, 204 se enquadram nesta regra, portanto 69.62%. Das pessoas que se enquadram nesta regra 100 % não foram morar na rua. Avaliação:204.
- **Não saiu de casa, responsáveis trabalham, estuda, não apanha da policia, apanha dos pais, não trabalha.** Enquadram-se 112 de 211, portanto 53,08%. Do total de 293 pessoas, 120 se enquadram nesta regra, portanto 40.95%. Das pessoas que se enquadram nesta regra 100 % não foram morar na rua. Avaliação:112.

5.1.2. Redes Neurais

Na realização dos testes, foram utilizados para o treinamento 132 dados, e os 171 registros restantes foram usados na generalização.

Foram realizados experimentos utilizando Redes Neurais “backpropagation”, para encontrar as melhores regras, e verificar a consistência dos resultados encontrados pelo algoritmo genético.

O algoritmo de extração de regra LOGIC foi aplicado às redes treinadas, de modo a gerar regras que pudessem ser comparadas as regras geradas pelo algoritmo genético.

Neste caso as entradas da rede seriam os campos relevantes do banco de dados dos meninos, os campos que na aplicação do algoritmo genético, indicaram porque os meninos foram morar na rua. E a saída desejada “target” seria portanto o campo “Mora na rua ou não”. Os dados escolhidos e sua codificação estão na tabela 5.1.

Nº	ENTRADAS	SE PRESENTE	SE AUSENTE
1	Saiu de casa com < 8 anos	1	-1
2	Saiu de casa com > 8 anos	1	-1
3	Morava com mãe e padrasto	1	-1
4	Morava com outros, não com mãe e padrasto	1	-1
5	Responsáveis trabalhavam	1	-1
6	Diferente de responsáveis trabalhavam	1	-1
7	Apanhava dos pais	1	-1
8	Não apanhava dos pais	1	-1

9	Pais usam ou drogas ou cigarros ou bebidas	1	-1
10	Pais não usam drogas ou cigarros ou bebidas	1	-1
11	Trabalha	1	-1
12	Não trabalha	1	-1

Tabela 5.1: Dados selecionados e codificados para a aplicação do algoritmo LOGIC

A rede foi treinada no Matlab 2.0, usando a função **trainbpm**, com os parâmetros: erro=0.001, lr=0.05, momento=0.95 e com 1000 épocas, e a função de avaliação tangente hiperbólica.

À partir dos resultados concluiu-se que a resposta as perguntas a seguir seriam:

“O que diferencia o menino que vai morar na rua, do que continua vivendo com sua família?”

- **Saiu de casa com > 8 anos, apanhava dos pais e pais usavam drogas ou cigarros ou bebidas.** Enquadram-se nesta regra 33 de 82, portanto 40.24%. Do total de 293 pessoas, 78 se enquadram nesta regra, portanto 26.62%. Das pessoas que se enquadram nesta regra 42.30 % foram morar na rua.
- **Apanhava dos pais, pais usavam drogas ou cigarros ou bebidas, e trabalhavam.** Enquadram-se nesta regra 21 de 82, portanto 25.6%. Do total de 293 pessoas, 29 se enquadram nesta regra, portanto 9.89%. Das pessoas que se enquadram nesta regra 72.41 % foram morar na rua.
- **Apanhava dos pais, pais usavam drogas ou cigarros ou bebidas.** Enquadram-se nesta regra 40 de 82, portanto 48.78%. Do total de 293 pessoas, 85 se enquadram nesta regra, portanto 29.01%. Das pessoas que se enquadram nesta regra 47.05 % foram morar na rua.

Quais as características dos meninos que vão morar na rua ?

- **Saiu de casa com > de 8 anos AND Responsáveis trabalhavam.** Enquadram-se 60 dos 82 meninos que foram morar na rua, portanto 73.17%. Avaliação = 60
- **Não morava com mãe e padrasto AND Responsáveis trabalhavam.** Enquadram-se 72 dos 82 meninos que foram morar na rua, portanto 87.80%. Avaliação = 72

Após este experimento modificou-se os campos pais usavam drogas ou cigarros ou bebidas, para pais usavam drogas. E o campo pais não usavam drogas ou cigarros ou bebidas, para pais não usavam drogas. E obteve-se os seguintes resultados:

O que diferencia o menino que vai morar na rua, do que continua vivendo com sua família?

- **Responsáveis trabalhavam AND Apanhava dos pais AND Pais usam drogas.** Enquadram-se 14 dos 82 meninos que foram morar na rua, portanto 17.07%. Das crianças que se enquadram nesta regra (13), 93.33% foram morar na rua. Avaliação = 13
- **Saiu de casa com mais de 8 anos AND Responsáveis trabalhavam AND Apanhava dos pais AND Pais usam drogas.** Enquadram-se 13 dos 82 meninos que foram morar na rua, portanto 15.85%. Das crianças que se enquadram nesta regra (12), 92.85% foram morar na rua. Avaliação=12

Quais as características dos meninos que vão morar na rua ?

- **Saiu de casa com mais de 8 anos AND Responsáveis trabalhavam** . Enquadram-se 60 dos 82 meninos que foram morar na rua, portanto 73.17%. Avaliação = 60

7. Conclusões

O estudo realizado, neste trabalho, sobre a aplicação de redes Neurais e Algoritmos Genéticos, mostrou o bom desempenho destes métodos na extração de regras de bancos de dados.

O método utilizando algoritmo genético para encontrar a melhor regra funciona corretamente conforme comprovado pelos experimentos realizados.

Foi verificada a necessidade do especialista para analisar o interesse de uma regra encontrada pelo método. Caso a regra encontrada não tenha interesse, deve-se excluir a informação do banco de dados para uma nova avaliação.

Quanto ao desempenho do algoritmo genético, foi verificado que a utilização da mutação forçando don't care é de extrema importância para o desempenho do processo.

Do ponto de vista da inicialização, quando se usa valores do próprio banco de dados na criação dos cromossomos iniciais, no lugar de inicializar randomicamente, acelera a convergência. Isto se deve a utilização de material genético existente no banco de dados diminuindo a possibilidade de busca errada (regras que nunca iriam existir no banco de dados).

O algoritmo LOGIC, de extração de regras de redes neurais, conseguiu explicar de forma compreensível a saída gerada pela rede neural treinada. Mostrando uma coerência quando comparadas as regras geradas pelo algoritmo genético. Demonstrando, portanto, que esses algoritmos funcionam corretamente, e que a utilização de métodos de extração de regras de redes neurais fornecem uma grande facilidade de compreensão dos resultados da rede neural.

Como o tempo disponível não foi suficiente para se conseguir uma quantidade de dados que pudesse dar um resultado satisfatório, os resultados devem ser levados em consideração com as restrições normais de carência de dados.

Bibliografia

[ADRI97] - Adriaans, P., Zantinge, D., “Data Mining”, Addison-Wesley, 1996.

[ALVE94] - Alves-Mazzotti, A.J., “Do trabalho à rua: uma análise das representações sociais produzidas por meninos trabalhadores e meninos de rua.” Laboratório do Imaginário Social e Educação, UFRJ, 1994.

[ANDR95] - Andrews, R., Diederich, J., Tickle, A. B., “A Survey And Critique Of Techniques for Extracting Rules from Trained Artificial Neural Networks”, Neurocomputing Research Centre, 1995.

[BATI97] - Batista, G. E. A. P. A., Milaré, C. R., Monard, M. C., “Descrição da implementação Prolog de uma Ferramenta para extração de Conhecimento de Redes Neurais”, Relatórios Técnicos, Instituto de Ciências e Matemática de São Carlos, Universidade de São Paulo, 1997.

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

[CHEN96] - Chen, M. - S., Han, J., Yu, P. S., “Data Mining: Na Overview from a Database perspective”, IEEE Transactions on knowledge and data engineering, vol. 8, no. 6, Dezembro, 1996.

- [**DATA96**] - Database Programming & Design (USA), Special Edition - Data Mining, Volume 9, Number 9, September 1996.
- [**DAVI91**] - Davis, L., "Handbook of Genetic Algorithms", Van Nostrand Reinhold, 1991.
- [**FAYY96**] - Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. - "Advances in Knowledge Discovery and Data Mining", AAAI Press, The MIT Press, 1996.
- [**GEVA95**] - Geva, S., Andrews, R., "Inserting and Extracting Knowledge from Constrained Error Back Propagation MLP". Proc. 5th Australian Conference on Neural Networks, Sydney, 1995.
- [**GOLD89**] - Goldberg, D. E., "Genetic Algorithms in Search, Optimization and Machine Learning", Addison Wesley Publishing Company, Inc, 1989.
- [**GORM88**] - Gorman, R.P., Sejnowski, T.J., "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets, Neural Networks", vol.1, pp.75-89, 1988.
- [**JAEG96**] - Jaeger, M., Mannila, H., Weydert, E., "Data Mining as Selective Theory Extraction in Probabilistic Logic", MPI Informatik, Im Stadtwald, D-66123 Saarbrücken, Germany.
- [**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.
- [**LOPE93**] - Lopes, H. S., Nassar, S. M., Pozo, A. R., Lima, W. C., "Arquitetura híbridas Integrando Paradigmas Conexionistas e Simbolistas". I Simpósio Brasileiro de Automação Inteligente-SBAI, Rio Claro, pp:10-18, 8-10 Setembro, 1993.
- [**MICH92**] - Michalewicz, Z., Genetic Algorithms + Data Structures = Evolution Programs. Second, Extended Edition. Springer-Verlag Berlin Heidelberg, 1992
- [**MICH95**] - Michalewicz, Z., "Genetic Algorithms + Data Structures = Evolution Programs", Springer series Artificial Intelligence – 3ª edição, 1995.
- [**MITC6**] - Mitchel, M., "An Introduction to Genetic Algorithms", MIT Press, 1996.
- [**NEUR94**] - "Neural Network Toolbox - For Use with MATLAB - User's Guide", The MathWorks Inc., versão 3, Howard Demuth Mark Beale, 1994.
- [**OLIV97**] - Oliveira, J. P., Uliana, P.B., Lima, W.C., "Algoritmos de Extração de Regras de Redes Neurais Artificiais", III Congresso Brasileiro de Redes Neurais, IV Escola de Redes Neurais, Florianópolis, julho de 1997.
- [**REIS97**] - Reis, C.L., Busca de conhecimentos em Bases de Dados, 1997, PUC-RIO.
- [**RUME86**] - Rumelhart, D.E., Hilton, G.E., Williams, R.J. - "Learning Representation by Back-Propagation Errors Nature", vol.323, pp. 533-536, October, 1986.
- [**THRU94**] - Thrun, S.B., "Extracting Provably Correct Rules From Artificial Neural Network". Technical Report IAI-TR-93-5 Institut für Informatik III Universität Bonn, 1994.
- [**TOWE93**] - Towell, G. G., Shavlik, J. W., "Extracting Refined Rules from Knowledge-Based Neural Networks". Machine Learning, Vol 131, pp:71-101, 1993.
- [**USAM96**] - Usama, M. F., Gregory, P. S., Padhraic, S. E Ramasamy, U., Advances in Knowledge Discovery and Data Mining, AAAI Press, The MIT Press, 1996.
- [**ZURA92**] - Zurada, J. M., "Introduction to Artificial Neural Systems", West Publishing Company, 1992.