Descoberta de Conhecimento em Big Data Usando Aprendizagem por Quantização Vetorial

FIDELIS, P. V. S.* MACHADO, L. C.S.† BOMFIM Jr, F. C.‡ FERNANDES, S. M.§

Resumo

O trabalho apresenta uma forma não convencional de mineração de dados para a obtenção de conhecimento. A proposta foi a de implementar uma rede neural LVQ que realiza a construção de classes através do seu vetor de pesos que agrupa dados que possuem padrões análogos. O método foi aplicado para determinar os clientes alvo que frequentam uma academia localizada na cidade de Uberaba-MG, desta forma foi possível compreender a característica padrão de seus clientes e, através disso, gerar estratégias de marketing para otimizar a fidelização e a busca de novos clientes.

Palavras-chave: Rede Neural, Big Data, Cluster, LVQ, KDD, Classes.

I Introdução

A manipulação do banco de dados denominado Big Data[1] apresenta grandes desafios quando a retirada de conhecimentos úteis, para serem utilizadas como ferramentas de tomada de decisões, análise de mercado futuro dentre outras aplicações. A proposta de deste trabalho é a de apresentar uma forma não convencional de minerar os dados por meio de uma rede neural LVQ[2], essa rede será um meio para a classificação dos dados em classes (cluster's), que serão avaliados posteriormente para se determinar padrões. A plataforma escolhida para o desenvolvimento do algoritmo foi o Software Matlab®[3] que possui uma linguagem de alto nível para resolução de problemas numéricos. Utilizando uma rede neural artificial com aprendizado em quantização vetorial será realizada a mineração dos dados de um banco de dados de uma academia localizada na cidade de Uberaba-MG. A análise dos dados dos alunos desta academia permitirá a obtenção de conhecimento sobre o seu público alvo para criar a fidelização destes clientes, além de proporcionar uma melhoria nas estratégias de marketing para captação de novos clientes.

*FIDELIS, P. V. S.,Universidade de Uberaba (UNIUBE), Uberaba, Minas Gerais, Brasil, paulavirginia.1994@gmail.com

†MACHADO, L. C.S.,Universidade de Uberaba (UNIUBE), Uberaba, Minas Gerais, Brasil, lorrannacsmachado@gmail.com

‡BOMFIM JUNIOR, F. C., Co-orientador, Universidade de Uberaba (UNIUBE), Uberaba, Minas Gerais, Brasil, florisvaldo.bomfim@uniube.br

§FERNANDES, S. M., Orientador, Universidade de Uberaba (UNIUBE), Uberaba, Minas Gerais, Brasil, gestor.engenhariaproducao@uniube.br

[1]Big Data é o termo que descreve o imenso volume de dados – estruturados e não estruturados – que impactam os negócios no dia a dia.

[2]learning vector quantization

[3]Software interativo de alta performance voltado para o cálculo numérico. O MATLAB integra análise numérica, cálculo com matrizes, processamento de sinais e construção de gráficos em ambiente fácil de usar onde problemas e soluções são expressos somente como eles são escritos matematicamente, ao contrário da programação tradicional.[8]

II Fundamentos Para a Investigação

A Visualização da Informação

A visualização de dados e de informação são úteis para se referir a qualquer representação visual de dados que são:

- Desenhados algoritmicamente (podem ter toques personalizados, mas são renderizados em grande parte com a ajuda de métodos computadorizados);
- Facilidade de regeneração com dados diferentes (a mesma forma pode ser reutilizada para representar conjuntos de dados diferentes com dimensões ou características semelhantes);
- Muitas vezes esteticamente estéril (os dados não estão decorados);
- Relativamente ricos em dados (grandes volumes de dados são bem-vindos e viáveis, em contraste com infográficos).

As visualizações de dados são inicialmente projetadas por um ser humano, então são desenhadas algoritmicamente com gráficos ou diagramação de software. A vantagem dessa abordagem é que é relativamente simples atualizar ou regenerar a visualização com mais ou novos dados. Embora possam mostrar grandes volumes de dados, as visualizações de informações são muitas vezes menos esteticamente ricas do que a infografia[4][9].

[4]Gênero jornalístico que utiliza recursos gráfico-visuais para apresentação sucinta e atraente de determinadas informações.

A.1 Valores Distintos

Um fator a considerar ao escolher uma propriedade visual é a quantidade de valores distintos que o seu leitor poderá perceber, diferenciar e possivelmente lembrar. Por exemplo, há muitas cores no mundo, mas não podemos dizer-lhes separadas se forem muito parecidas. Podemos diferenciar mais facilmente um grande número de formas, uma grande quantidade de posições e um número infinito de números. Ao escolher uma propriedade visual, selecione uma que tenha vários valores diferenciáveis úteis e uma ordem semelhante a dos seus dados. Com isso temos quatro categorias de gráficos que são: Comparação, Composição, Distribuição e Relação.

B KDD - Knowleadge Discovery Database

O KDD é o conhecimento descoberto utilizando uma base de dados, é o processo de metamorfosear dados em conhecimento. Todo o processo KDD é apresentado na figura 1 que representa cada etapa da técnica de subtração de conhecimento da base de dados, possuindo várias etapas interligadas sequencialmente que são: seleção, pré-processamento, transformação, *Data mining* e interpretação enquanto que *Data mining*[5] é empregado somente para o estágio de descoberta do processo.[7]

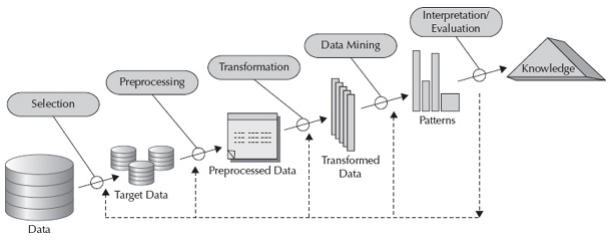

Figura 1: Processo KDD [7]

Etapas do processo

1. Limpeza dos dados: etapa onde são eliminados ruídos e dados inconsistentes.

2. Integração dos dados: etapa onde diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados.

3. Seleção: etapa onde são selecionados os atributos que interessam ao usuário. Por exemplo, o usuário pode decidir quais informações, como endereço e telefone, não são relevantes para determinar se um cliente é um bom comprador ou não.

4. Transformação dos dados: etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (por exemplo, através de operações de agregação).

5. Mineração: etapa essencial do processo consistindo na aplicação de técnicas inteligentes a fim de se extrair os padrões de interesse.

6. Avaliação ou Pós-processamento: etapa onde são identificados os padrões interessantes de acordo com algum critério do usuário.

7. Visualização dos Resultados: etapa onde são utilizadas técnicas de representação de conhecimento a fim de apresentar ao usuário.

B.1 Análise de Clusters (Agrupamentos)

Diferentemente da classificação e predição onde os dados de treinamento estão devidamente classificados e as etiquetas das classes são conhecidas, a análise de *clusters* trabalha sobre dados onde as etiquetas das classes não estão definidas. A tarefa consiste em identificar agrupamentos de objetos, agrupamentos estes que identificam uma classe. Por exemplo, poderíamos aplicar análise de *clusters* sobre o banco de dados de um supermercado a fim de identificar grupos homogêneos de clientes, por exemplo, clientes aglutinados em determinados pontos da cidade costumam vir ao supermercado aos domingos, enquanto clientes aglutinados em outros pontos da cidade costumam fazer suas compras às segundas-feira.[7]

C Redes Neurais Artificiais

Inteligência Artificial é um algoritmo com a capacidade de processar informações e, a partir delas, cumprir tarefas cognitivas como perceber, aprender, melhorar seu desempenho, classificar, tomar decisões e agir de acordo com as condições externas. Haykin (1999, p. 59) descreve que uma IA deve ser capaz de armazenar conhecimento, aplicar o conhecimento armazenado para resolver problemas, e adquirir novo conhecimento através da experiência. Existem vários tipos de IAs que se diferenciam pela suas formas de processar e classificar informações. A Rede Neural Artificial, que será utilizada neste projeto, baseia-se no funcionamento do cérebro biológico. Formado por cadeias de neurônios artificiais, elas se adaptam conforme o ambiente, sendo capazes de generalizar e organizar os dados obtidos na aprendizagem de acordo com padrões detectados.

A figura 2 representa um neurônio artificial do modelo MCP, proposto por McCullock e Pitts em 1943.[6] De acordo com o modelo, os vetores xn são as entradas do neurônio, estas obtidas a partir dos dados externos de uma certa aplicação. Cada vetor possui um peso sináptico w

[5] É uma expressão inglesa ligada à informática cuja tradução é mineração de dados. Consiste em uma funcionalidade que agrega e organiza dados, encontrando neles padrões, associações, mudanças e anomalias relevantes.[8]

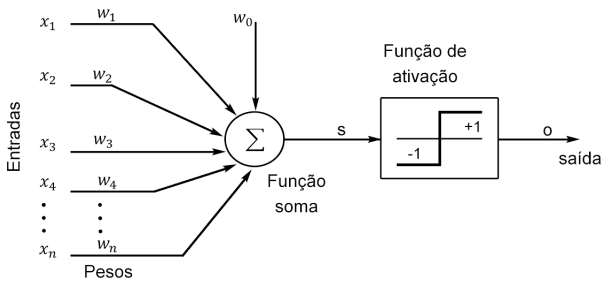

Figura 2: Perceptron [6]

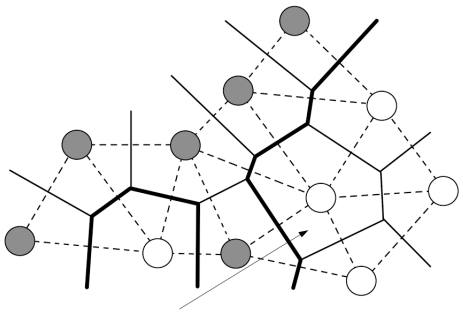

Figura 3: Cluster's de uma rede LVQ

que mede a relevância de tal entrada na saída do neurônio. Com os valores de x(n) e w(n) definidos, é feita a soma ponderada das entradas, sendo esta a saída linear u. Aplicando uma função de ativação, geralmente uma função degrau unipolar (ou função de *threshold*) obtém-se a saída do neurônio y.Dessa forma, a saída do neurônio pode ser expressa (Equação 1) como:

$$\begin{aligned} y_i = 0 \text{ se } u < 0 \\ y_i = 1 \text{ se } u >= 0 \end{aligned} \tag{1}$$

Assim como existem vários tipos de IA, existem vários tipos de classificações de redes neurais, que são baseadas em suas arquiteturas e funcionamentos. Neste projeto, utilizaremos uma rede neural LVQ.

D Rede Neural LVQ

A rede LVQ foi idealizada por Teuvo Kohonen[5], sendo uma versão supervisionada dos mapas auto-organizáveis, tendo seu treinamento baseado em competição. É um método de classificação de padrões no qual cada unidade de saída representa uma classe em particular, o vetor pesos para a unidade de saída é chamado de *codebook*.[2]

D.1 Treinamento

O treinamento de uma rede LVQ é, de forma competitiva, análogos aos usados nas rede SOM[6], sendo que os vetores de peso dos neurônios estão representando os respectivos vetores quantizadores de classes, conforme figura 3. Assim para a utilização dessa topologia, as diversas classes associadas à representação do processo devem ser conhecidas. A figura 4 demonstra uma rede LVQ composta de ordem n entradas e n_1 neurônios, os quais estão representando todas as classes envolvidas com o referido problema de classificação de padrões a ser mapeado.

Conforme observado, a arquitetura de Kohonen não possui conexões laterais entre neurônios, sendo que este aspecto implica que aqueles neurônios vizinhos ao vencedor não vão ter os seus pesos ajustados.[1] Existem dois tipos de algoritmo de treinamento denominados $LVQ-1$ e $LVQ-2$ que são utilizados para o ajuste de peso do neurônio vencedor.

[6] self-organization maps

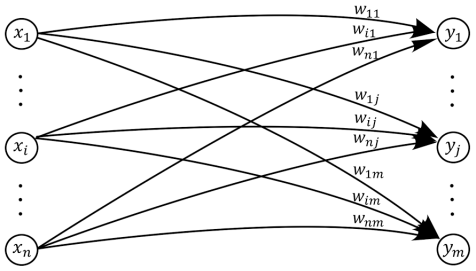

Figura 4: Estrutura da Rede LVQ

O algoritmo $LVQ-1$ ajusta os pesos apenas do neurônio vencedor, já o algoritmo $LVQ-2$ ajusta os pesos do neurônio vencedor e do vice. Para a realização deste trabalho foi adotado apenas o algoritmo de treinamento $LVQ-1$.

E Algoritmo de treinamento LVQ-1

Como foi dito, esse algoritmo apenas realiza o ajuste do neurônio vencedor, neste caso considera-se que cada vetor de entrada $x^{(k)}$, pertence somente a uma das classes j previamente conhecidas, pois o mecanismo de aprendizagem é feito de forma supervisionada.[1] Os dois passos principais do algoritmo consiste na obtenção do neurônio vencedor (neurônio com menor distância Euclidiana) e ajuste do peso do mesmo. Em relação à obtenção do vencedor, aquele que obtiver maior proximidade com uma determinada amostra $x^{(k)}$, será declarado o vitorioso, sendo a medida de proximidade a norma euclidiana[7] entre esses dois parâmetros (Equação 2).

$$dist_j^{(k)} = \sqrt{\sum_{i=1}^{n}(x_i^{(k)} - w_i^{(j)})^2}, onde(j = 1, ..., n_1) \tag{2}$$

Sendo $dist_j^{(K)}$ a distância entre o vetor de entrada representado à k-ésima amostra $x^{(k)}$ em relação ao vetor de peso do j-ésimo neurônio $w^{(j)}$. Os ajuste de pesos são realizados

[7] Em matemática, distância euclidiana (ou distância métrica) é a distância entre dois pontos, que pode ser provada pela aplicação repetida do teorema de Pitágoras. Aplicando essa fórmula como distância, o espaço euclidiano torna-se um espaço métrico.

de acordo com as condições estabelecidas pela equação 3 apresentada abaixo.

$$\begin{aligned} &\text{Se } x^{(k)} \in C^{(j)} \\ \text{Então: } & w^{(j)} = w^{(j)} + \eta.(x^{(k)} - w^{(j)}) \\ \text{Senão: } & w^{(j)} = w^{(j)} - \eta.(x^{(k)} - w^{(j)}) \end{aligned} \tag{3}$$

Onde o parâmetro η define a taxa de aprendizagem.

A figura 5 ilustra os mecanismo de ajustes dos pesos do neurônio vencedor.

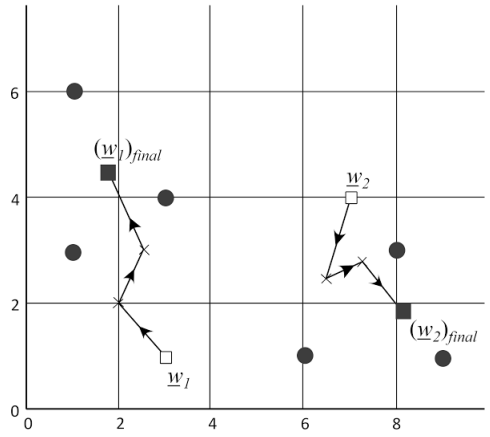

Figura 5: Representação de Aprendizagem

O algorítimo 1 de treinamento LVQ-1 e apresentado abaixo demonstrando todas as suas etapas de treinamento. Onde o algorítimo so vai cessar o laço quando o peso neural e o número de época for igual ao limite estabelecido ou não houver alterações nos pesos neurais.

Algoritmo 1: LVQ-1 -Fase de treinamento[1]

início
  Obter o conjunto de amostra de treinamento $x^{(k)}$;
  Associar cada amostra $x^{(k)}$ com sua respectiva classe;
  Iniciar os vetores de peso de cada neurônio;
  Normatizar os vetores de amostras e pesos;
  Especificar a taxa de aprendizagem η;
  Iniciar o contador do número de epocas;
  repita
    Calcular distância euclidiana;
    Declarar o vencedor com menor distância;
    Ajustar o vetor de peso do vencedor(equação 3);
    Normatizar vetores de peso;
    $epoca = epoca + 1$;//
  até *epoca>=n ou não mudança nos vetores de peso*;
fim

Quando o processo de treinamento é finalizado usamos o segundo algorítimo para classificar cada dado do Big Data em sua respectiva classe. Se observarmos o algoritmo 2 vemos que o mesmo trabalha com os pesos neurais encontrado e os dados de entrada.

Algoritmo 2: LVQ-1 - Fase de operação[1]

Entrada: x,w
Resultado: Classificação
início
  Apresentar a amostra x a ser classifica e normatizada;
  Assumir os vetores de peso já ajustados no algoritmo 1;
  Calcular distância euclidiana entre x e $w^{(k)}$;
  Declarar o neurônio vencedor;
  Associar a amostra a classe;
fim

III Desenvolvimento

Primeiramente todos os dados forma convertidos para um range de 0 à 1, e logo após foram processados pela rede neural. O todo o software desenvolvido foi usando o Matlab®, para a determinação da menor distância euclidiana foi usado o seguinte código (algoritmo 3) abaixo:

Algoritmo 3: Determinação do menor valor

```
for i=1:nn
 total=0;
  for j=1:ne
 total=(entrada(j)-w(i,j))^2+total;
  end
 total=sqrt(total);
 w(i,ne+1)=total;
end
minww=min(w(:,ne+1));
```

Assim que foi determinada a menor distância euclidiana, e feita uma varredura em todas os dados para localizar o elemento de menor distância fazendo assim os ajuste do peso neural (algoritmo 4).

IV Resultados dos Eventos Simulados

Foi escolhido que a rede neural deve dividir o Big Data(4000 dados) em 10 classes, com uma taxa de aprendizagem η de 0.05 e com número de treinamento máximo de 70 treinamento, para fazer a análise, após o treinamento realizado pelo algorítimo 1 para obtermos os pesos neurais para serem utilizados.

Após realizado a divisão aplicou-se o algoritmo 2, realizando assim a mineração dos dados os quais obtemos as

Algoritmo 4: Atualização de pesos neurais

```
flag=0;
for i=1:nn
 if minww==w(i,ne+1)
  if flag==0
   for j=1:ne
 w(i,j)=w(i,j)+0.05*(entrada(j)-w(i,j));
 dados(ponteiro,ne+1)=i;
 flag=1;
  end
 end
end
```

seguintes informações alocadas em 5 classes demonstradas nas tabelas de 1 até 5.

Tabela 1: Característica da classe 1

1ª Classe	
Mulheres:	100%
Homens:	0%
Faixa etária:	criança
Distância:	0→6Km
Turno:	Noite

Tabela 2: Característica da classe 2

2ª Classe	
Mulheres:	70%
Homens:	30%
Faixa etária:	criança, adolescente e adulto(maioria)
Distância:	0→9Km
Turno:	Manhã(maioria), Tarde e Noite.

Tabela 3: Característica da classe 3

3ª Classe	
Mulheres:	100%
Homens:	0%
Faixa etária:	criança(maioria), adolescente e adulto
Distância:	0→5Km
Turno:	Manhã(maioria), Tarde e Noite.

Analisando as classes encontradas podemos definir 5 padrões de clientes para essa empresa que são:

1. criança do sexo feminino que moram em uma faixa de até 6km no turno noturno.

Tabela 4: Característica da classe 4

4ª Classe	
Mulheres:	0%
Homens:	100%
Faixa etária:	criança, adolescente e adulto(maioria)
Distância:	0→6Km
Turno:	Tarde e Noite(maioria).

Tabela 5: Característica da classe 5

5ª Classe	
Mulheres:	0%
Homens:	100%
Faixa etária:	criança, adolescente e adulto(maioria)
Distância:	0→9Km
Turno:	Manhã, Tarde

2. Mulheres que moram em uma faixa de até 9 km no turno da manhã.
3. criança do sexo feminino que moram em uma faixa de até 5 km no turno da manhã.
4. Homens que moram até 6Km no período da noite.
5. Homens que moram até 9Km no período da tarde e noite.

Após analise das tabelas podemos motar o seguinte figura 6 usando técnicas de visualização da informação.

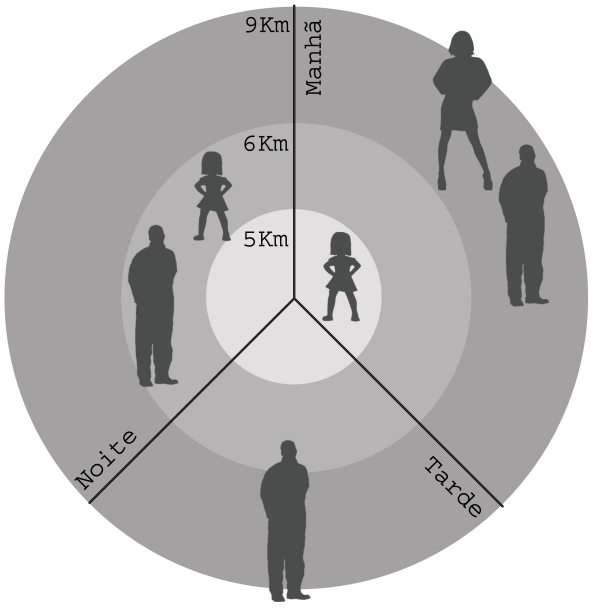

Figura 6: Gráfico de apresentação

Desta forma a empresa deve focar sua propaganda em

uma região até 9km de sua localização com horários de curso que atenda a demanda deste clientes.

V Conclusão

Com base nos resultados encontrado, concluímos que é possível utilizar uma rede neural LVQ como ferramenta de mineração de dados, a qual consegui dividir-los em grupos que otimizaram a retirada do conhecimento do grande grupo de dados usados na mineração. Os resultados foram repassados para empresa para a tomada das ações necessárias.

Referências

[1] da Silva, I.N., Spatti, D.H. e Flauzino, R.A. (2010) "Redes Neurais Artificiais Para Engenharia e Ciências Aplicadas", Artliber Editora Ltda., ISBN: 9788588098534.

[2] Braga, A.P., de Carvalho, A.P.L.F. e Ludermir, T.B. (2007) "Redes Neurais Artificiais – Teoria e Aplicações", Editora LTC, 2a. edição, ISBN: 9788521615644.

[3] Chauvin, Y. e Rumelhart, D.E. (1995) "Backpropagation: Theory, Architectures, and Applications", Lawrence Erlbaum Associates, ISBN: 080581258X.

[4] Arbib, M.A. (ed.) (2002) "The Handbook of Brain Theory and Neural Networks", The MIT Press, 2nd. edition, ISBN: 0262011972.

[5] Kohonen,T., The Self-Organizing Map, Proceedings of the IEEE, vol.78, no. 9, pp. 1464-1480, September, 1990.

[6] MCCULLOCH W. and PITTS W.. A Logical Calculus of the Ideas Immanent in Nervous Activity, Bulletin of Mathematical Biophysics, 5, p. 115-133, 1943.

[7] S. de Amo: Curso de Data Mining, Programa de Mestrado em Ciência da Computação, Universidade Federal de Uberlândia, 2003. http://www.deamo.prof.ufu.br/CursoDM.html

[8] WIKIPEDIA, Matlab, Disponível em:<https://pt.wikipedia.org/wiki/MATLAB>, Acesso em: 12/10/2017.

[9] ILIINSKY, N. and STEELE, J.. Designing Data Visualizations,,Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2011.