8º FÓRUM ENSINO · PESQUISA EXTENSÃO · GESTÃO
FEPEG
UNIVERSIDADE: SABERES E PRÁTICAS INOVADORAS
Trabalhos científicos · Apresentações artísticas e culturais · Debates · Minicursos e Palestras
REALIZAÇÃO: Unimontes
APOIO: FAPEMIG FADENOR
24 a 27 setembro
Campus Universitário Professor Darcy Ribeiro
www.fepeg.unimontes.br

Comparação da Eficiência dos Algoritmos de *K-means* e *LVQ – Learning Vector Quantization* na Categorização de Dados

Jessica Medeiros Queirós, Huggo Ferreira Silva, Ana Cláudia de Moura Laurentino

Introdução

A categorização de dados é uma das técnicas mais utilizadas no processo de mineração de dados. O objetivo principal de categorizar dados é extrair relações lógicas entre dados aparentemente desconexos para produção de informações e extrair conclusões sobre a distribuição de dados de um determinado conjunto.

O processo de categorização dos dados é feito normalmente através de algoritmos específicos que filtram os dados apresentando como resultado as informações que ele inicialmente foi treinado para apresentar. O treinamento desse algoritmo pode ser supervisionado ou não-supervisionado, de acordo com a sua especificidade.

São exemplos de algoritmos amplamente utilizados para a categorização de dados o K-means e o LVQ, ambos são conhecidos como algoritmos de clusterização, nomenclatura que segundo (Cluto, 2010) é dada àqueles algoritmos que trabalham com a divisão do conjunto de dados em grupos significativos, chamados de clusters, e buscam maximizar a similaridade intra-cluster e minimizar a similaridade inter-cluster. E serão estes algoritmos os adotados para composição deste trabalho.

Embora ambos utilizem o mesmo método de aprendizado, o fato de algoritmos usarem o método de aprendizado de máquina não-supervisionado não implica que ele obtém o mesmo resultado ao final de um teste com os mesmos dados. Ao final desse trabalho mostraremos que utilizando as mesmas técnicas e os mesmos dados os algoritmos em questão têm resultados diferentes no final.

Material e métodos

A. Aprendizagem Não-Supervisionada

No paradigma de aprendizado não-supervisionado, como o próprio nome sugere, não se tem um agente externo auxiliando no aprendizado da rede. A rede tem acesso, somente, aos padrões de entrada e estas devem conter dados redundantes para que seja possível identificar um padrão, não há uma resposta desejada para ser comparada com a resposta da rede, como é feito no aprendizado supervisionado.

As redes baseadas no aprendizado não-supervisionado tentam descobrir características semelhantes nos dados de entrada e a partir destas similaridades é que os pesos da rede vão sendo ajustados. À medida que a rede vai estabelecendo uma harmonia com as regularidades estatísticas dos dados de entrada, ela desenvolve a habilidade de criar representações internas para codificar as características de entrada e criar novas classes automaticamente.

A aprendizagem não supervisionada é dividida em 6 etapas: Seleção de atributos; Medida de proximidade; Critério de agrupamento; Algoritmo de agrupamento; Verificação dos resultados; e Interpretação dos resultados.

Primeiramente os atributos devem ser adequadamente selecionados de forma a codificar a maior quantidade possível de informações relacionada a tarefa de interesse e devem ter também uma redundância mínima entre eles. Medida de Proximidade é a medida para quantificar quão similar ou dissimilar são dois vetores de atributos. É ideal que todos os atributos contribuam de maneira igual no cálculo da medida de proximidade, um atributo não pode ser dominante sobre o outro, por isso é importante normalizar os dados.

O Critério de Agrupamento depende da interpretação que o especialista dá ao termo sensível com base no tipo de cluster que são esperados. Um cluster compacto de vetores de atributos, por exemplo, pode ser sensível de acordo com um critério enquanto outro cluster alongado, pode ser sensível de acordo com outro critério.

Tendo adotado uma medida de proximidade e um critério de agrupamento devemos escolher um algoritmo de clusterização que revele a estrutura agrupada do conjunto de dados. Uma vez obtidos os resultados do algoritmo de agrupamento, devemos verificar se o resultado está correto. Essa verificação geralmente é feita através de testes apropriados. Em geral, os resultados da clusterização devem ser integrados com outras evidências experimentais e análises para chegar as conclusões corretas.

B. K- means

Desenvolvido por J.B. MacQueen em 1967, o K-means é um algoritmo de para agrupamento de dados tentando fornecer classificações de acordo com os próprios dados, isso é, utilizando de aprendizagem não supervisionada. Tal

classificação é feita por similaridades de grupo e um novo objeto é classificado de acordo no grupo que tiver maior similaridade.

O algoritmo de clusterização K-means poder ser também chamado de K-médias. Segundo Jain et. al. (1999) o algoritmo K-means tornou-se popular por sua fácil implementação e seu baixo custo, pois tem ordem de complexidade O(n), onde n é o número de padrões.

A começo, de todos os padrões de entrada, k deles são selecionados, seja aleatoriamente ou conforme alguma heurística. A definição do valor de k é conforme a quantidade de grupos planeja-se classificar. Cada um dos k padrões escolhidos se tornam, inicialmente, o centro, ou centróide, de cada de seu respectivo grupo. Após isso, cada um dos outros padrões deve ser ligado ao centroide mais próximo.

Uma vez feito isso, é definido o novo centróide para cada grupo através do cálculo de centroides. Tal cálculo é feito a partir da localização do ponto médio de cada grupo que se tornará o novo centroide. Isso é, dado um conjunto de padrões $V=\{x_1, x_2, ..., x_n\}$ de um grupo *a*, o novo centroide desse grupo será dado por $c(a)= \frac{\sum V}{n_a}$, sendo que n_a é o número de elementos no grupo *a*.

Após isso, haverá novamente a associação dos padrões com seus centroides mais próximos para, novamente, haver o cálculo de centroides. Essa repetição continua até que não haja alterações em duas interações seguidas.

C. LVQ - Learning Vector Quantization

O Algoritmo LVQ foi desenvolvido por Teuvo Kohonen em 1991. A ideia principal do algoritmo é encontrar agrupamentos naturais em um conjunto de dados. O algoritmo consiste na utilização de vetores nomeados, e cada um destes corresponde a uma classe, às quais possuem um vetor de pesos associados, também chamado de vetor de referência.

O treinamento deste algoritmo é feito de maneira não supervisionada, o que significa que este modelo de rede extrai, por si mesma, os padrões e relações existentes dentro do conjunto de dados inserido, ou seja o processo de treinamento é realizado sem a existência da saída esperada para verificação do processo de inferência realizado.

Cada item do vetor de entrada de dados para o treinamento é inserido na rede. O vetor de peso mais próximo, chamado também de neurônio vencedor, é definido, o neurônio então é atualizado. Após essa atualização se o neurônio se aproximar do item do vetor de entrada, significa que estes são da mesma classe, se os mesmos se afastarem, então pertencem a classes diferentes.

Algoritmo - LVQ

Passo 1. Inicializar os pesos e os parâmetros
Passo 2. Fazer durante várias iterações
Passo 3. Para cada padrão de treinamento
Passo 4. Definir neurônio vencedor
Passo 5. Atualizar os pesos do vencedor e dos vizinhos
Passo 6. Reduzir taxa de aprendizagem
Passo 7. Até o erro ser menor que um valor alvo

Desenvolvimento

A metodologia utilizada para o desenvolvimento deste trabalho foi a implementação dos algoritmos utilizando a ferramenta Matlab, e para comparação da eficiência dos mesmos, estes algoritmos foram testados com um conjunto de dados gerados de forma aleatória pela ferramenta Weka.

Após a aplicação do conjunto de dados a acertividade de cada algoritmo foi calculada através da contagem de elementos que foram erroneamente classificados, através da análise dos dados em cada cluster.

A base de dados utilizada é constituída de 95 padrões, cujos valores foram gerados aleatoriamente seguindo as regras:

- Os padrões deviam ser agrupados em quatro grupos;
- Cada grupo deve ter, no mínimo, um padrão e, no máximo, cinquenta padrões.

Conclusões

Considerando que o Kohonen dedica um neurônio para a identificação de um tipo de padrão, ele se torna mais sensível que o K-means no processo de clusterização. Isso se torna ainda mais claro quando se é levado em conta que o K-means é altamente dependente do espaçamento entre dois grupos de padrões. Logo, por mais que o K-means demonstre um desempenho eficiente, principalmente quando há o grande afastamento de grupos, a utilização do Kohonen é mais

recomendável quando há uma alta necessidade de diminuição do erro, pois esse é capaz de se adequar mais à situação. Em uma situação em que não haja tanta necessidade de adequação, o K-means se torna mais interessante por sua facilidade de entendimento, implementação e utilização.

Referências

As referências (limitadas a 10) deverão ser escritas com a fonte Times New Roman (tamanho 07, espaçamento simples, alinhadas à esquerda).

[1] BROWNLEE, Jason. **Clever Algorithms: Nature-Inspired Programming Recipes.** Disponível em <http://www.cleveralgorithms.com/nature-inspired/neural/lvq.html> Acesso em: 24 jun 2014

[2] Disponível em <http://geneura.ugr.es/g-lvq/section3_2.html> Acesso em: 24 jun 2014

[3] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. ACM Comput. Surv., 31(3):264–323

Tabela 1. Os resultados obtidos através da aplicação da base supracitada.

Comparação entre os Algoritmos	K- means	Kohonen
Padrões	95	95
Número de Interações	6	9
Acerto	57	91
Aproveitamento	60,00%	95,79%
Agrupamento	4 centróides	4 neurônios