O Algoritmo *Density-Based Spatial Clustering of Applications With Noise* (DBSCAN) na Clusterização dos Indicadores de Dados Ambientais

Éverton Marangoni Gava[1], Gabriel Felippe[1], Kristian Madeira[1], Maicon Bastos Palhano[1], Merisandra Cortês de Mattos Garcia[1], Paulo João Martins[1], Priscyla Waleska Targino de Azevedo Simões[1]

[1]Grupo de Pesquisa em Inteligência Computacional Aplicada – UNESC – Criciúma – Santa Catarina - Brasil

evertonmgava@yahoo.com.br, gabrielheavy@hotmail.com,{kma, maiconpalhano, mem, pjm, pri}@unesc.net

Abstract. *With the high growth of data bases dimension and the information flow of data generation born the necessity of knowledge extraction from data repositories, so, comes the Knowledge Discovery in Databases (KDD), a process with several levels which the most important level is the data mining technique, able to extract patterns to obtain knowledge. The DBSCAN is an algorithm of data mining able to cluster data by density method and this method was applied in the river system basin of coal region of Criciúma data base to obtain results of water. Using Shell Orion Data Mining Engine and Weka tools to obtain comparison and evaluate generations of results, in which we realized that Araranguá basin have a lower pH.*

Palavras-chave*: KDD, Data Mining, Clustering, DBSCAN.*

Resumo. *Com o crescimento da dimensão das bases de dados e do fluxo de dados gerados, surgiu a necessidade de extração de conhecimento desses grandes repositórios, sendo assim, nasceu o Knowledge Discovery in Databases (KDD), um processo com várias etapas no qual a principal é o data mining, que é capaz de extrair padrões para que com eles seja obtido o conhecimento. O DBSCAN é um algoritmo de data mining que clusteriza dados pelo método de densidade e este foi aplicado em dados referentes a qualidade da água em bacias hidrográficas da região carbonífera de Criciúma. Para isso, utilizaram-se as ferramentas Shell Orion Data Mining Engine e Weka a fim de se comparar e avaliar os resultados gerados, no qual percebeu-se que a bacia de Araranguá tem maior pH.*

Palavras-chave*: KDD, Data Mining, Clusterização, DBSCAN.*

1. Introdução

O progresso nas tecnologias de armazenamento e aquisição de dados resultou em crescimento das bases de dados. Embora sabendo que padrões interessantes e informações potencialmente úteis podem ser extraídos desses repositórios, o volume de

dados torna-se difícil, senão impraticável, a busca por esse conhecimento implícito sem o auxílio de tecnologias computacionais [HAN; KAMBER, 2006].

A fim de facilitar o processo de descoberta de conhecimento, técnicas capazes de extrair informações significativas destes vastos repositórios foram desenvolvidas, sendo que a procura por relações úteis entre os dados ficou conhecida como *Knowledge Discovery in Databases* (KDD), tendo o *data mining* como a principal etapa desse processo, no qual o algoritmo DBSCAN será aplicado na base de dados das bacias hidrográficas da região carbonífera de Criciúma.

2. Data Mining

O *data mining* pode ser definido como o processo de explorar e analisar grandes conjuntos de dados, extraindo informação e conhecimento sob a forma de novas relações e padrões úteis na resolução de problemas de um domínio de aplicação específico.

Considerando que o *data mining* pode ser aplicado em diversos campos de pesquisa, e que os usuários do conhecimento gerado por esse processo podem estar interessados em tipos distintos de padrões, naturalmente existem diversas tarefas de *data mining*, sendo que a escolha de uma determinada tarefa depende do conhecimento que se deseja obter e do tipo de conjunto de dados que se tem a disposição [WITTEN; FRANK; HALL, 2011].

2. Clusterização

A tarefa de clustering analysis de um grupo heterogêneo de dados em vários subgrupos, também chamados de *clusters*, é denominada clusterização [HAN; KAMBER, 2006].

Um *cluster* pode ser entendido como uma coleção de registros que são similares entre si e dissimilares de objetos em outros grupos, onde objetos pertencentes a um dado *cluster* devem compartilhar um conjunto de propriedades comuns, sendo que essas propriedades não são compartilhadas com objetos de outros *clusters* [FAYADD; PIATETSKY-SHAPIRO,SMYTH, 1996].

Além de permitir a estruturação e fornecer uma melhor compreensão do conjunto de dados original, os resultados da tarefa de clusterização também podem ser utilizados por outras técnicas de *data mining*, que realizariam seu trabalho nos *clusters* encontrados.

3. Métodos de Clusterização

Existem diversos métodos de Clusterização, dentre os quais podem ser destacados os métodos hierárquicos, de particionamento, baseados em grade, em modelos e em densidade [HAN; KAMBER, 2006; JAIN; MARTY; FLYNN, 1999].

Métodos tradicionais de clusterização, como os de particionamento, geralmente enfrentam dificuldades para encontrar agrupamentos com formatos arbitrários e não retornam bons resultados quando a base de dados em questão está contaminada com *outliers* (dados que destoam do padrão geral da distribuição). Outro ponto a ser considerado em alguns destes métodos, é que o usuário tem a necessidade de informar

previamente o número de *clusters* que serão gerados, o que na maioria das vezes não é uma tarefa simples [ESTER et al, 1996; HAN; KAMBER, 2006].

Considerando-se isso foi proposto o método de clusterização baseado em densidade [KRIEGER et al, 2011].

Métodos baseados em densidade são usados para a detecção de agrupamentos com formatos arbitrários em conjuntos de dados contendo *outliers*. Para algoritmos que adotam essa abordagem, *clusters* são regiões com alta densidade de pontos no espaço dos dados, separadas de outras regiões densas, por regiões de baixa densidade (que representam ruídos). Por sua vez essas regiões de alta densidade podem conter formato arbitrário no espaço de dados [ANKERST et al, 1999; KRIEGEL et al, 2011].

Algoritmos que encontram clusters baseados em densidade não necessitam que seja informado de maneira prévia o número de grupos a serem formados, e não fazem suposições sobre a variância ou a distribuição interna dos objetos nos possíveis grupos que possam vir a existir no conjunto de dados. Essas propriedades permitem que sejam encontrados agrupamentos baseados nas propriedades dos dados, o que consequentemente impõe uma estruturação menos rígida aos objetos [KRIEGEL et al, 2011].

2. O Algoritmo DBSCAN

No ano de 1996, Martin Ester, Hans-Peter Kriegel, Jörg Sander e Xiaowei Xu publicaram o artigo intitulado *A Density Based Spatial Clustering of Applications With Noise*. Nesse artigo foi apresentado o algoritmo DBSCAN, que pode ser aplicado a grandes conjuntos de dados que possuem *outliers*, ao mesmo tempo em que encontra *clusters* com diversos formatos com eficiência aceitável.

O DBSCAN encontra agrupamentos baseando-se na vizinhança dos objetos, onde a densidade associada a um ponto é obtida por meio da contagem do número de pontos vizinhos em uma determinada região ao redor desse ponto [ERTÖZ, STEINBACH; KUMAR, 2003]. Esse algoritmo possui a capacidade de encontrar clusters considerando as propriedades dos dados, pois não requer que seja informado antecipadamente o número de *clusters*, permitindo a formação de grupos com formatos arbitrários. Em contrapartida são necessários outros dois parâmetros de entrada para o algoritmo. Outras características importantes do algoritmo são a capacidade de identificar *outliers*, e a possibilidade de poder trabalhar com diversas funções de distância [ANKERST et al, 1999; METZ, 2006; ESTER et al, 1996].

Os dois parâmetros de entrada que o DBSCAN necessita são:

a) **raio de ε-vizinhança de um ponto**: determina o raio de vizinhança ε para cada ponto da base de dados. Dado o parâmetro ε, o algoritmo DBSCAN verifica a quantidade de pontos contidos no raio ε para cada ponto da base de dados, e se essa quantidade exceder certo número, um *cluster* é formado;
b) **número mínimo de pontos (η)**: parâmetro que especifica o número mínimo de pontos, no dado raio de ε-vizinhança, que um ponto precisa possuir para ser considerado um ponto central e conseqüentemente, de acordo com as definições de *cluster* baseado em densidade, iniciar a formação de um *cluster*.

Definidos os parâmetros de entrada e a base de dados a ser clusterizada, o primeiro passo do DBSCAN consiste em construir uma estrutura denominada matriz de dissimilaridade. Em uma matriz de dissimilaridade é possível representar a distância entre pares de objetos. Essa matriz sempre será quadrada e de tamanho *n* x *n*, onde *n* representa a quantidade de objetos que serão clusterizados.

$$\begin{bmatrix} 0 & & & \\ dist(2,1) & 0 & & \\ dist(3,1) & dist(3,2) & 0 & \\ dist(4,1) & dist(4,2) & dist(4,3) & 0 \end{bmatrix}$$

Essa matriz de dissimilaridade é descoberta empregando-se uma medida que é responsável pelo cálculo de distância para todos os pares de objetos da base de dados. As medidas de distancia mais comumente utilizadas são a distância euclidiana (equação 1) e a *Manhattan* (equação 2):

$$d(xi, xj) = \sqrt{\sum_{l=1}^{d} (xil - xjl)^2}$$

$$d(xi, xj) = \sum_{l=1}^{d} |(xil - xjl)|$$

Sabendo que atributos diferentes podem ser medidos em escalas distintas, caso for utilizada diretamente uma medida de distância como a euclidiana, por exemplo, atributos com escalas maiores irão se sobrepor a atributos medidos em escalas menores, tornando a clusterização tendenciosa [HAN; KAMBER, 2006]. Para normalização das escalas dos atributos, utilizou-se a normalização MIN-MAX:

$$Zi = \frac{Vi - \min vi}{\max vi - \min vi}(n\_\max - n\_\min) + n\_\min$$

Tendo a matriz de dissimilaridade montada, o próximo passo executado pelo algoritmo consiste em verificar a ε-vizinhança de cada ponto da base de dados *D* com a finalidade de identificar possíveis pontos centrais para iniciar a formação dos agrupamentos. Um objeto x_j está na ε-vizinhança de um objeto x_i se:

$$Xj \in N\varepsilon(xi) = \{xi, xj \in D | dist(xi, xj) \leq \varepsilon\}$$

A matriz de dissimilaridade contendo as distâncias entre os pares de objetos da base de dados é consultada e se a distância entre um ponto x_i e um ponto x_j for menor ou igual a o dado parâmetro ε, ou seja, $(x_i, x_j) \leq \varepsilon$ e $dist_{ij}\, x\, x$, então o ponto x_j está na ε-vizinhança do ponto x_i.

Verificados quais os objetos que estão contidos na vizinhança de x_i, o algoritmo precisa verificar a cardinalidade (*Card*) desse ponto com relação aos objetos vizinhos, com a finalidade de definir a condição do ponto.

Caso a cardinalidade de x_i com relação ao raio de ε-vizinhança seja igual ou exceda o parâmetro η, então esse ponto é considerado um ponto central e os objetos contidos em sua ε-vizinhança são então diretamente alcançáveis por densidade. Para que um ponto x_j

seja diretamente alcançável por densidade a partir de um ponto x_i, o objeto x_i deve satisfazer as condições:

$$xj \in N\varepsilon(xi)$$
$$Card(N\varepsilon(xi)) \geq \eta$$

Se as duas condições forem satisfeitas, o algoritmo irá criar um cluster C_i contendo todos os pontos diretamente alcançáveis por densidade a partir de um determinado objeto x_i:

$$Ci = N\varepsilon(xi)$$

A partir do momento em que o DBSCAN forma um *cluster* C_i, todo o objeto nesse grupo tem recursivamente a sua ε-vizinhança recuperada permitindo assim que novos pontos possam ser adicionados ao *cluster*. O processo de consulta de vizinhos próximos é repetido até que todos os pontos em C_i sejam verificados. Quando isso ocorrer, o algoritmo encerra o crescimento de C_i e visita o próximo ponto não visitado da base de dados. O processo é repetido até que todos os registros do conjunto de objetos tenham sido visitados e classificados.

O algoritmo distingue três tipos de objetos em um conjunto de dados:

a) **pontos centrais:** são pontos que estão no interior de uma região densa, onde existem pelo menos η pontos no raio ε desse objeto. A cardinalidade desses pontos em relação ao parâmetro de ε-vizinhança deve ser de no mínimo η pontos;
b) **pontos de borda:** estão na fronteira de uma região densa, ou seja, sã pontos que estão na ε-vizinhança de algum ponto central, porém não são pontos centrais, pois a cardinalidade desses pontos em relação ao raio ε não excede η;
c) ***outliers*:** esses pontos não são centrais e nem de borda e assim não são conectados por densidade a nenhum outro ponto, não pertencendo a nenhum *cluster*.

Caso um ponto for classificado como *outlier* pelo algoritmo, posteriormente ele pode estar na ε-vizinhança de outro ponto não visitado ainda pelo DBSCAN. Sendo assim, essa classificação pode ser removida caso o objeto seja diretamente alcançável por densidade a partir de um ponto central ainda não visitado.

3. Base de dados

A base de dados utilizada para avaliação do algoritmo DBSCAN apresenta indicadores ambientais da qualidade dos recursos hídricos da bacia do rio Araranguá, rio Tubarão e rio Urussanga, sendo composta por 1723 registros com 20 atributos, sendo ela de domínio público com dados compreendendo o período de 2002 até 2009.

Baseado no conhecimento das etapas do KDD e da necessidade de uma base de dados limpa de *outliers* foi feita uma preparação específica na base dados, onde foram eliminados atributos com valores nulos e corrigidos os atributos com valores incorretos, além disso, foram atualizados novos campos com dados atuais.

3.1. Utilização das ferramentas Shell Orion e Weka

A ferramenta Shell Orion Data Mining Engine encontra-se em desenvolvimento por um grupo de pesquisa em inteligência computacional de uma Universidade do Brasil e tem diversos algoritmos de *data mining* implementandos, entre eles o DBSCAN.

A ferramenta Weka, desenvolvida na Universidade de Waikato, Nova Zelândia, possui muitos algoritmos de *data mining* e como a Shell Orion, também possui o DBSCAN, que é o algoritmo utilizado nesta pesquisa.

Sendo assim, ambas as ferramentas foram testadas e utilizadas com a base de dados das bacias hidrográficas, os resultados foram comparados e todos os parâmetros das duas ferramentas foram utilizados de maneira compatível com o objetivo de não gerar resultados que fujam do padrão.

Para a seleção dos parâmetros de entrada do algoritmo DBSCAN o seguinte critério foi adotado:

a) os atributos de entrada desejados foram selecionados;
b) para o parâmetro η foram selecionados valores entre 4 e 10;
c) a função *k-dist* foi calculada iterativamente com *k* valendo η em cada iteração, ou seja η-*dist*;
d) o parâmetro ε foi definido iterativamente analisando-se o gráfico *k-dist*;
e) utilizou-se a medida de distância euclidiana normalizada.

4. Resultados obtidos

Os atributos de entrada selecionados foram os índices de pH e a concentração de ferro presentes nas amostras obtidas ao longo da bacia do rio Araranguá. O objetivo desse teste foi separar os registros coletados em dois grupos, ou seja, pontos de coleta em que a água está contaminada e pontos de coleta em que a água não está contaminada. Na Tabela 1 estão descritos os resultados gerados pelo algoritmo utilizando a distância euclidiana com anormalização dos dados.

Tabela 1- Resultados obtidos utilizando a distância euclidiana normalizada

Cluster	Quantidade de elementos	Porcentagem	Situação do ponto de coleta
1	132	19,44%	não-contaminado
2	538	79,23%	contaminado
Outliers	9	01,33%	-

Empregando-se a distância euclidiana normalizada o DBSCAN identificou que em 79,23% das amostras coletadas na bacia do rio Araranguá os índices de pH estão geralmente abaixo de 5.0 e a concentração de ferro presente nas águas se encontra em índices geralmente maiores que 20 mg/L. Baixos valores de pH e altas concentrações de ferro e outros sulfatos demonstram a degradação dos rios da bacia por atividades ligadas a extração de carvão e fazem com que os seus recursos hídricos se apresentem de maneira imprópria para o consumo humano e uso em geral [ALEXANDRE; KREBS, 1996].

Através da mineração de dados utilizando as ferramentas Shell Orion e Weka com o algoritmo DBSCAN, foi possível obter diversos resultados, dentre eles o

conhecimento de que a bacia de Araranguá possui concentração maior de valores de coleta com o pH menor, como é possível visualizar nos gráficos gerados pelos algoritmos.

Em ambas as ferramentas o algoritmo DBSCAN gerou três *clusters*, um representando a bacia de Araranguá, outro representando a bacia de Tubarão e outro a de Urussanga, observa-se um gráfico de X por Y, onde Y representa a bacia e X representa os valores dos pH's, cada ponto no gráfico é atribuído de uma forma no qual ele esteja no seu respectivo cluster e ao mesmo tempo vinculado com o seu pH, com isso pode-se observar as concentrações de pH nas suas devidas bacias. Os resultados encontrados pelas duas ferramentas foram idênticos.

Diante dos resultados obtidos pode-se confirmar a aplicabilidade do algoritmo para a tarefa de clusterização. A possibilidade de encontrar *outliers* entre os dados e a formação de *clusters* com formatos arbitrários, tornam o algoritmo uma opção interessante na clusterização de bases de dados com essas características. Também não existe a necessidade de ser informada a quantidade de c*lusters* desejados, o que faz com que o algoritmo imponha uma estruturação menos rígida aos dados.

5. Considerações finais

O artigo apresentou todos os processos necessários para que seja feito uma extração de conhecimento de uma base de dados, sendo assim o KDD e suas etapas foram estudadas, em especial o processo principal conhecido como *data mining* e o algoritmo DBSCAN que utiliza o método de densidade foi aplicado.

Entende-se que com base no *data mining* e nas suas respectivas tarefas e métodos, especificamente pelo DBSCAN, observa-se uma forma de contribuição desses métodos computacionais para diversas áreas de conhecimento.

Conclui-se que os resultados do DBSCAN foram satisfatórios em ambas as ferramentas, possibilitando que novas minerações e tarefas de data mining sejam aplicadas nesta base de dados dos indicadores ambientais da qualidade dos recursos hídricos das bacias hidrográficas da região carbonífera.

Referências

ANKERST, Mihael; BREUNIG, Markus M.; KRIEGEL, Hans-Peter; SANDER,Jörg. OPTICS: Ordering Points To Identify the Clustering Structure. In Proc. ACM SIGMOD'99 Int. Conf. on Management of Data (SIGMOD'99), pages 49-60, Philadelphia, 1999.

ERTÖZ, Levent; STEINBACH, Michael; KUMAR, Vipin. Finding Cluster of Diferent Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In Proceedings of Second SIAM International Conference on Data Mining, Arlington, 2003.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to discovery knowledge in databases. AI Magazine, 3(17): 37-54. 1996.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. Data mining: uma guia prático: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005.

HAN, Jiawei. KAMBER, Micheline. Data mining: concepts and techniques. 2. ed. San

Francisco: Morgan Kaufmann, 2006.

J. Han e M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

LAROSE, Daniel T. Discovering knowledge in data: an introduction to data mining.

Hoboken: Wiley-Interscience, 2005.

REZENDE, Solange Oliveira (coordenadora). Sistemas Inteligentes: fundamentos e aplicações. Barueri, SP: Manole, 2003. 525p.