IX Brazilian Symposium on GeoInformatics, Campos do Jordão, Brazil, November 25-28, 2007, INPE, p. 277-282.

Extensão do WEKA para Métodos de Agrupamento com Restrição de Contigüidade

Carlos Eduardo R. de Mello, Geraldo Zimbrão da Silva, Jano M. de Souza

Programa de Engenharia de Sistemas e Computação
Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 68.511 – Zip Code: 21945-970 – Rio de Janeiro – RJ – Brazil

{carlosmello,zimbrao,jano}@cos.ufrj.br

Abstract. *This work addresses the shortage of open-source data mining tools that implement spatial data mining methods. Therefore, this work presents the development of a WEKA extension for contiguity-constrained clustering method.*

Resumo. *Este trabalho aponta a pouca disponibilidade de ferramentas de mineração de dados de código-aberto que implementam métodos de mineração de dados espaciais. Portanto, o objetivo deste trabalho, para tentar resolver esse problema, é apresentar o desenvolvimento de uma extensão da ferramenta WEKA para métodos de agrupamento com restrição de contigüidade.*

1. Introdução

O uso de Sistemas de Informação Geográfica e de Bancos de Dados Espaciais permitiu que grandes quantidades de dados fossem coletadas e armazenadas. Entretanto, extrair conhecimento desses dados de maneira manual torna-se inviável, sendo necessário a utilização de métodos de Descoberta de Conhecimento (também conhecido com Mineração de Dados).

Existem vários métodos e ferramentas para Descoberta de Conhecimento em Bases de Dados que tratam dados convencionais. No entanto, além de dados convencionais, os dados geográficos armazenam suas geometrias e relações entre os objetos espaciais. Essa característica faz com que ferramentas e métodos específicos de descoberta de conhecimento em bases de dados espaciais sejam necessários.

Existem várias ferramentas que implementam os algoritmos clássicos de mineração de dados disponíveis no mercado, entretanto, a maioria delas são pagas ou não possuem código-aberto. As ferramentas que implementam algoritmos de mineração de dados em bases de dados espaciais são raras, acarretando a necessidade de desenvolvimento de novas ferramentas ou a implementação dos algoritmos de mineração de dados espaciais em ferramentas de código-aberto existentes.

Uma ferramenta de código-aberto bastante utilizada é o WEKA (*Waikato Environment for Knowledge Analysis*). Esta ferramenta, desenvolvida na Universidade de Waikato de Hamilton, Nova Zelândia, implementa mais de vinte algoritmos diferentes de mineração de dados convencionais. Pelo fato dessa ferramenta ser de código-aberto, muitas iniciativas de extensão têm sido realizadas. Em [Bogorny et al., 2006], temos o desenvolvimento de uma extensão do WEKA para suportar a extração de regras de associação espaciais.

O objetivo deste trabalho é apresentar uma extensão do WEKA que implementa o algoritmo de agrupamento de dados espaciais com restrição de contigüidade, utilizando as funcionalidades já implementadas no WEKA.

Na seção 2, apresentamos uma visão geral de Mineração de Dados. Na seção 3, apresentamos os métodos de agrupamento e o algoritmo de agrupamento com restrição de contigüidade. Em seguida (seção 4), apresentamos a extensão do WEKA e o algoritmo de agrupamento implementado. Finalmente, na seção 5, descrevemos as conclusões desse trabalho e os trabalhos futuros.

2. Mineração de Dados

Mineração de Dados, ou Descoberta de Conhecimento em Bases de Dados, é definida por [Fayyad et al., 1996] como o processo não-trivial de descoberta de padrões válidos, novos, potencialmente úteis e compreensíveis a partir de dados. Segundo [Ester et al.,2000], o processo de Mineração de Dados é interativo e iterativo englobando várias atividades, como as seguintes:

Seleção: seleção do subconjunto de todos os atributos e do subconjunto de todos os dados em que o conhecimento possa ser descoberto;

Redução: redução das dimensões dos atributos ou técnicas de transformação para reduzir o número efetivo de atributos a serem considerados;

Mineração de dados: a aplicação de algoritmos apropriados que, sob um limite aceitável de eficiência computacional, produzem uma enumeração particular de padrões sobre os dados; e

Análise: interpretação e análise dos padrões descobertos com respeito a sua utilidade em uma dada aplicação.

Embora muitos estudos tenham sido realizados em bancos de dados relacionais (uma visão mais geral pode ser encontrada em [Chen et al.,1996]), ainda há uma grande demanda em outras áreas de aplicação de bancos de dados, incluindo bancos de dados espaciais, bancos de dados temporais, bancos de dados para *multimedia*, etc.

O processo de mineração de dados espaciais é mais complexo que o de dados relacionais, tanto pelo aspecto de eficiência dos algoritmos, como pelo aspecto da complexidade da descoberta de possíveis padrões [Ester et al., 2001]. Uma das razões da complexidade para a descoberta de padrões está no fato que os algoritmos de mineração de dados espaciais devem levar em consideração as relações de vizinhança entre os objetos espaciais para extrair informação útil. Isto é necessário porque as relações de um objeto com os seus vizinhos podem influenciar significativamente o próprio objeto [Ester et al., 2001]. Por outro lado, a eficiência dos algoritmos de mineração de dados espaciais está relacionada à grande quantidade de dados espaciais, à complexidade dos tipos de objetos espaciais e aos métodos de acesso aos dados espaciais [Koperski et al., 1996].

Os principais algoritmos de mineração de dados espaciais cobrem os problemas de classificação, generalização, agrupamento e regras de associação [Ester et al., 2001].

3. Algoritmos de Agrupamento

Nesta seção apresentamos uma visão geral sobre algoritmos (ou métodos) de agrupamento e em seguida descrevemos o método de agrupamento com restrição de contigüidade.

Uma das técnicas mais utilizadas em Mineração de Dados é o Agrupamento. O objetivo dessa técnica é separar objetos ou observações em grupos, onde os objetos mais semelhantes estejam em um mesmo grupo e objetos distintos estejam em grupos diferentes. Portanto, seu objetivo principal é identificar estruturas ou grupos presentes em dados [Ng and Han, 1994].

Os algoritmos de agrupamento podem ser classificados em duas categorias principais: *métodos hierárquicos* ou *métodos de particionamento* [Ng and Han, 2002].

Os *métodos hierárquicos* se dividem em *aglomerativos* ou *divisivos*. Dados *n* objetos para serem agrupados, nos métodos *aglomerativos* começamos o algoritmo com *n* grupos, cada qual formado por um objeto. A cada passo do algoritmo, dois grupos semelhantes são aglomerados transformando-se em um novo grupo. Este processo é repetido até que exista apenas um único grupo contendo todos os *n* objetos. Nos métodos *divisivos*, dados *n* objetos, o algoritmo começa com um único grupo contendo todos os *n* objetos. A cada passo do algoritmo, os grupos formados são divididos de acordo com a similaridade dos objetos contidos neles. Este processo é repetido até que *n* grupos com apenas um objeto sejam formados. Esses dois métodos, *aglomerativos* e *divisivos*, são chamados hierárquicos, pois criam uma relação de hierarquia entre os grupos formados. A partir da visualização da hierarquia desses grupos, o usuário pode decidir com quais grupos deseja trabalhar.

Os *métodos de particionamento* trabalham com um número fixo de grupos. Dados *n* objetos para serem agrupados em *k* grupos, os *métodos de particionamento* tentam encontrar as *k* melhores partições para os *n* objetos. Segundo [Ng and Han, 2002], é muito comum encontrar casos onde os *k* grupos encontrados pelo *método de particionamento* são de melhor qualidade (*i.e.*, mais similares) que os grupos encontrados nos *métodos hierárquicos*. Por isso, os *métodos de particionamento* têm recebido maior atenção da área de análise de grupos. Além disso, muitos *métodos de particionamento* baseados no *k-means* e no *k-medoid* têm sido desenvolvidos.

O *k-means* (ou *k-médias*) é o *método de particionamento* que utiliza o ponto médio da distância entre os objetos no espaço para representar o centro do grupo (ou centróide). Por outro lado, o *k-medoid* utiliza o objeto espacial dentro do grupo mais próximo do ponto médio da distância euclidiana para representar o centro do grupo. O método *k-medoid* é mais robusto que o *k-means* com relação aos *outliers* e os grupos formados por esse independem da ordem com que os objetos são examinados durante sua execução.

3.1. Métodos de Agrupamento com Restrição de Contigüidade

A maioria dos algoritmos de agrupamento de dados espaciais utiliza a distância física entre os objetos espaciais para calcular a similaridade entre os objetos. Em [Ng and Han, 2002], Ng e Han apresentam como agrupar objetos espaciais de geometria poligonal convexa utilizando o algoritmo CLARANS e três maneiras diferentes de calcular a distância entre dois polígonos convexos.

Os métodos de agrupamento com restrição de contigüidade, além dos dados convencionais, também utilizam a informação espacial para restringir os grupos de objetos formados. Em [Gordon, 1995], são definidas duas abordagens para incorporar a informação de proximidade dos objetos espaciais aos métodos de agrupamento: distância geográfica ou grafo (ou matriz) de contigüidade.

Na primeira abordagem, uma medida de distância entre os objetos espaciais é adotada para descobrir a proximidade entre os objetos. Um método de agrupamento clássico pode ser adaptado para considerar, além dos dados convencionais (os atributos dos objetos), a distância física entre os objetos. Alternativamente, esta abordagem pode ser realizada em dois estágios. No primeiro, um algoritmo de agrupamento tradicional pode ser aplicado aos dados convencionais. No segundo estágio, é realizada uma reavaliação dos grupos formados, levando em consideração a distância entre os objetos espaciais para a realocação de objetos.

Na segunda abordagem, a informação topológica dos objetos espaciais é representada através de dispositivos auxiliares como grafo ou matriz. Nessa representação, os nós dos grafos são os objetos espaciais e as arestas são as relações de vizinhança entre eles. Uma aresta *a* ligando os nós *A* e *B* em um grafo indica que os objetos espaciais *A* e *B* são vizinhos. Essa abordagem também pode ser realizada em dois estágios. No primeiro estágio, um algoritmo de agrupamento clássico é aplicado aos dados convencionais. No segundo estágio, os grupos formados são reavaliados utilizando as informações de vizinhança entre objetos contidos no grafo (ou matriz), que foram carregadas a partir um banco de dados espacial.

Em [Neves et al., 2002], é apresentado o método de agrupamento com restrição de contigüidade por árvore geradora mínima. Esse método começa com a construção de um grafo, onde regiões são os nós do grafo e as arestas são as relações de vizinhança entre as regiões. Em seguida, são dados pesos para as arestas desse grafo. O valor dos pesos é a medida de dissimilaridade entre os vetores dos atributos das regiões que são representadas no grafo. A partir do grafo das regiões com os pesos das arestas, é executado um algoritmo para encontrar uma árvore geradora mínima para o grafo. Então, a árvore geradora mínima sofre uma poda. As arestas mais caras da árvore são retiradas, formando sub-árvores desconectadas. Essas sub-árvores desconectadas representam os grupos de regiões formados. Á medida que as arestas mais caras da árvore geradora mínima são retiradas, novos grupos são formados.

Segundo [Neves et al., 2002], o método de agrupamento com restrição de contigüidade por árvore geradora mínima funciona de maneira eficiente para a detecção de regiões. A ferramenta SKATER desenvolvida na Universidade Federal de Minas Gerais implementa esse método [SKATER] [Neves et al., 2002].

4. Extensão do WEKA

Nesta seção vamos apresentar a nossa extensão do WEKA, que implementa o algoritmo de agrupamento com restrição de contigüidade por árvore geradora mínima.

O WEKA possui implementados vários métodos de agrupamento para dados convencionais, entretanto, não encontramos na literatura nenhuma iniciativa de implementação de métodos de agrupamento espaciais. O objetivo do presente trabalho é implementar uma extensão do WEKA para o método de agrupamento com restrição de contigüidade por árvore geradora mínima.

Parte da implementação do trabalho realizado em [Bogorny et al., 2006] foi utilizada em nossa extensão. Nesse trabalho foi desenvolvida uma extensão do WEKA para extração de regras de associação espaciais. Para isso, foi implementado um algoritmo de pré-processamento das informações espaciais, que armazena os dados da topologia dos objetos espaciais em uma tabela do banco de dados e em arquivos.

Através do resultado do pré-processamento descrito em [Bogorny et al., 2006], informações da topologia de vizinhança entre as regiões são extraídas do banco de dados. Essas informações são carregadas em uma matriz de contigüidade, onde as dimensões dessa matriz são iguais ao número de regiões a serem agrupadas. Para cada par de regiões vizinhas é calculado o valor da dissimilaridade entre elas e armazenado na matriz.

O cálculo da dissimilaridade entre os objetos espaciais é realizado através da função de dissimilaridade já implementada no WEKA. A medida adotada para isso é a distância euclidiana entre os vetores formados pelos dados dos atributos das regiões. Portanto, quanto mais semelhantes são as regiões, mais próximo de zero será o valor da dissimilaridade entre elas.

A matriz de contigüidade preenchida pode ser interpretada como um grafo. Portanto, desenvolvemos um algoritmo para encontrar a árvore geradora mínima (AGM) a partir da matriz. Com isso, as arestas de maior custo que geram ciclos são eliminadas do grafo.

A AGM resultante representa as relações de vizinhança mais fortes, isto é, conectam as regiões mais similares. Assim, cada aresta podada da árvore gera duas sub-árvores. O critério utilizado para a poda das arestas consiste em eliminar as arestas de maior peso em ordem descrente, separando as regiões vizinhas menos similares. Cada árvore da floresta representa um grupo de regiões vizinhas e similares.

A implementação do algoritmo foi realizada através da classe chamada *ClusterContiguityConstraintTree*. Além disso, esta classe também é responsável por carregar as informações da topologia que estão na tabela gerada pelo pré-processamento dos dados espaciais e as informações dos dados contidos no arquivo *ARFF* do WEKA.

Portanto, nossa extensão do WEKA baseou-se em utilizar uma série de funcionalidades já existentes dentro da própria ferramenta e de parte da implementação realizada por [Bogorny et al., 2006], contemplando o método de agrupamento com restrição de contigüidade por árvore geradora mínima.

5. Conclusões e Trabalhos Futuros

Neste trabalho apontamos o problema da pouca disponibilidade de ferramentas mineração de dados em código-aberto que implementam métodos de mineração de dados espaciais. A principal contribuição deste trabalho, para tentar resolver esse problema, foi o desenvolvimento de uma extensão do WEKA que implementa o método de agrupamento de dados espaciais com restrição de contigüidade por árvore geradora mínima. Como trabalhos futuros, estamos realizando experimentos e avaliando os resultados obtidos de nossa implementação.

6. Agradecimentos

Agradecemos ao CNPQ pelo apoio financeiro.

Referências

Bogorny, V., Palma, A., Engel, P. and Alvares, L.O. (2006) "Weka-GDPM: Integrating Classical Data Mining Toolkit to Geographic Information Systems.", In: SBBD Workshop on Data Mining Algorithms and Aplications(WAAMD'06), pp.9-16, Florianopolis, Brazil, October 16-20.

Chen, M., Han, J. and Yu, P. S. (1996) "Data Mining: An Overview from Database Perspective", IEEE Transactions on Knowledge and Data Eng., 8(6):866-883, December.

Ester, M., Frommelt, A., Kriegel, H. P. and Sander, J. (2000) "Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support.", Data Mining and Knowledge Discovery Vol. 4, No. 2, July, pp. 193-216.

Ester, M., Kriegel, H. P. and Sander, J. (2001) "Algorithms and Applications for Spatial Data Mining", invited chapter for Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) "Knowledge discovery and data mining toward a unifying framework.", In Proceeding of The Second Int. Conference on Knowledge Discovery and Data Mining, pages 82-88.

Gordon, A.D. (1996) "A survey of constrained classification.", Computational Statistics & Data Analysis, v. 21, p. 17-29.

Koperski, K., Adhikary, J. and Han, J. (1996) "Spatial data mining: Progress and challenges survey paper", In Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada.

Neves, C.M., Câmara, G., Assunção, R.M. and Freitas, C.C. (2002) "Procedimentos Automáticos e Semi-automáticos de Regionalização por Árvore Geradora Mínima.", In: Simpósio Brasileiro de Geoinformática, GeoInfo.

Ng, R. and Han, J. (1994) "Efficient and Effective Clustering Methods for Spatial Data Mining", Proceedings of 20th International Conference on Very Large DataBases, pp 144-155.

Ng, R. and Han, J. (2002) "CLARANS: A Method for Clustering Objects for Spatial Data Mining", IEEE Trans. Knowledge & Data Engineering , 14, 5, pp 1003-1016, September.

SKATER, http://www.est.ufmg.br/leste/skater.htm.