VISUALIZAÇÃO DE AGRUPAMENTOS DE DADOS MULTIDIMENSIONAIS

Renann Rodrigues da Silva (PIBIC/Fundação Araucária/UEPG), José Carlos Ferreira da Rocha (Orientador), e-mail: jrocha@uepg.br.

Universidade Estadual de Ponta Grossa/Departamento de Informática.

Ciências Exatas e da Terra, Metodologias e Técnicas de Computação.

Palavras-chave: Visualização de dados, redes bayesianas, distâncias probabilísticas.

Resumo

O uso de métodos para visualização de dados multivariados objetiva facilitar a detecção de estruturas latentes que não são evidentes na base de dados original. Kontkanen *et al* (2000) utilizaram uma abordagem com distâncias probabilísticas para mensurar a distância entre os casos da base de dados, e usaram estas distâncias para produzir uma visualização. O formalismo das redes bayesianas foi empregado para computar as distribuições posteriores relacionadas a cada caso. A fim de reduzir o tempo de processamento gasto nesta tarefa este trabalho emprega a programação concorrente durante o cômputo de distâncias probabilísticas. Os resultados obtidos indicam que houve um ganho de tempo no processamento da probabilidade *a posteriori.*

Introdução

A disponibilidade de grandes volumes de dados tem motivado o desenvolvimento de métodos e aplicações que tratam da descoberta de conhecimento em bancos de dados e mineração de dados (Fayad, 2001). Contudo, a escolha das técnicas de pré-processamento de dados pode ser dificultada quando a base de dados tem alta dimensionalidade (Ahrens, Geveci e Law, 2006). Neste contexto, o uso de visualizações gráficas na apresentação de dados pode facilitar o processamento dos dados (Etemadpour *et al*, 2015).

Segundo Ward, Grinstein e Keins (2010), a visualização de dados envolve o uso de gráficos e imagens para apresentar dados e comunicar informações explorando a percepção visual humana. Alguns esquemas de visualização exigem que os dados sejam transformados antes que possam ser apresentados. Um procedimento de transformação que tem sido bastante utilizado é o escalonamento multidimensional (EMD) (Shepard, 1962). O objetivo da transformação é capturar estruturas não evidentes em uma matriz de dados empíricos e ao mesmo tempo representá-las em uma forma acessível.

A composição de visualizações alternativas de bases de dados classificadas, isto é, bases de dados em que cada entrada do conjunto de registros fornecidos contém uma indicação da classe a que pertence, pode auxiliar a identificação de agrupamentos de casos, subestruturas dentro de cada classe (Wang et al, 2011). Segundo Kontkanen *et al* (2000) essas subestruturas podem ser identificadas avaliando o quão similar são as predições derivadas a partir dos atributos de entrada para cada caso. Ainda segundo aqueles autores, uma forma efetiva de predizer a similaridade na predição de dois casos é determinar a distância das distribuições posteriores da variável alvo em tais distribuições. O cômputo das distribuições posteriores pode ser efetuado usando o formalismo das redes bayesianas.

De acordo com Chickering (1996), o cômputo da probabilidade *a posteriori* em redes bayesianas é no pior caso, NP-completo. Isto pode acarretar em um alto custo computacional elevado para problema do cálculo das distâncias probabilísticas como proposto por Kontkanen *et al* (2000). Considerando isso, este trabalho apresenta um procedimento que usa computação concorrente no cálculo das distribuições posteriores da variável de classe para cada caso de uma base de dados. O objetivo é reduzir o tempo total de processamento no cômputo das distâncias das distribuições entre pares de registros em uma base de dados. O desempenho do procedimento proposto foi utilizado na avaliado no agrupamento de dados supervisionados. O experimento mostrou que o uso de programação concorrente no cômputo das distribuições a posteriori resultou em um ganho de tempo significativo.

Materiais e Métodos

Uma RB (rede bayesiana) é uma dupla (G, Θ) na qual G é um grafo acíclico e direcionado em que os nós representam elementos de um conjunto de variáveis $\mathbf{V}=\{X_1, X_2 ..., X_n\}$ e os arcos representam dependência direta entre as variáveis (Russell, 1995). O segundo componente, Θ, denota o conjunto de parâmetros da rede. Para cada nó X_i, Θ especifica a distribuição condicional $p(X_i|pais(X_i))$, onde $pais(X_i)$ indica os pais do nó X_i em G. No todo esta estrutura define uma distribuição de probabilidades conjuntas sobre **V**. Uma tarefa de inferência frequente efetuada com redes bayesianas é a atualização de crença. A atualização de crenças é o cálculo das probabilidades posteriores das variáveis aleatórias. Isto é, dada uma coleção de observações $(x_i, \ldots, x_k)$ com evidências sobre as variáveis $X_i, \ldots, X_k$ é uma variável de interesse X_q, o objetivo da atualização de crença é calcular $p(X_q|x_i, \ldots, x_k)$.

Dada uma base de dados **X**, o escalonamento multidimensional é mapeia os dados para um espaço de visualização sem operar diretamente **X**, mas sobre uma matriz de distâncias derivadas de **X** (Shepard, 1972). Os objetos em estudo são então representados por pontos em modelo espacial de tal forma que as características desde objetos são reveladas pela transformação geométrica obtida. Este trabalho considera o emprego do EMD métrico, em que as escalas de mensuração pertencem aos números reais.

Kontkanen *et al* (2000) propõe o uso de redes bayesianas para determinar a distância probabilísticas usadas no EMD. A ideia básica é executar o procedimento de atualização de crença para calcular a probabilidade posterior da classe de cada caso, determinar as distâncias das distribuições entre cada par de pontos e então elaborar uma visualização que minimize a expressão:

$$\text{Minimize} \sum_{i-0}^{n} \sum_{j=i+i}^{N} \left(d(wi,wj) - d'(wi',wj')\right)^2$$

Nesta expressão, *d(wi, wj)* indica a distância entre dois vetores, *wi* e *wj* no espaço original e *d'(wi', wj')* a distância equivalente entre os vetores no espaço transformado. Em particular, os autores consideram o problema da visualização de dados supervisionados com o objetivo de detectar estruturas latentes que divirjam da classificação original. Kontkanen *et al* (2000) sugerem o uso da distância baseada na hipótese MAP e da divergência de Kullback-Leibler (Gelman, Carlin, Stern, & Rubin, 1995) como métricas para o EMD. É importante destacar que outros escores possíveis são: a distância de Bhattacharyya e a distância de Hellinger (Yang e Le Cam, 2000).

Apesar de existirem diversos algoritmos eficientes para o cômputo da atualização de crença em redes bayesianas (Russell, 1995), a necessidade de

executá-los para cada par de pontos do conjunto de dados pode dificultar o emprego do procedimento proposto por Kontkanen *et al* (2000) no escalonamento multidimensional. Em termos práticos, a utilização desta abordagem para o processo de visualização proposto, quando repetidas execuções do mesmo processo se faz necessário, o uso de inferências *a posteriori* pode ser inviável. Principalmente, quando se considera que estas distâncias devem ser calculadas para *n*(n-1)* pares de registros; assumindo que a base de dados *n* casos.

A programação concorrente é um paradigma de programação para a construção de softwares que fazem o uso da execução concorrente de várias tarefas computacionais interativas, que podem ser implementados como programas separados ou como um conjunto de threads (Hartley, 1998). A intenção é tirar vantagem de situações em que um programa com múltiplos threads é executado em computadores com múltiplos processadores e memória compartilhada. Neste caso, a execução pode ocorrer de forma que um ou mais segmentos de threads sejam executados simultaneamente em diferentes unidades de processamento, gerando assim um ganho de tempo na tarefa como um todo.

Resultados e Discussão

A fim de disponibilizar um mecanismo eficiente para o cômputo da atualização de crença quanto do cálculo das distâncias probabilísticas para o EMD, O desenvolvimento explorou a técnica de programação orientada a objetos. Neste contexto, a classe chamada “Utils” localizada dentro do pacote EMDc, implementa as funções para o cálculo das distâncias probabilísticas entre os pontos dados, usando a divergência Kullback-leibler, distância Hellinger e Bhattacharyya. A rotina responsável pelo cálculo da matriz que armazena as distâncias entre os dados e que usa computação concorrente é descrita na Figura1. Este procedimento cria um thread, que por sua vez recebe uma linha como parâmetro, tornando-se responsável pelo cálculo da distância desta linha para as demais linhas da lista. Ao final do procedimento é retornada como resposta uma matriz contendo as distâncias de cada par de pontos de dados.

Algoritmo 1: MATRIZ DE DISTANCIAS

```
Entrada: Lista de probabilidades
1 início
2     para cada linha ∈ Lista faça
3         Obj ← Distancia(linha, Lista)
4         NovaThread ← Obj.calcularDistancias()
5         novaLista ← NovaThread.getObj().getListaDistancias()
6         resultado ← novaLista
7     fim
8 fim
9 retorna resultado // matriz com as distancias entre as linhas
```

Figura1 – Pseudocódigo do procedimento concorrente.

O protótipo foi testado no cômputo das distâncias empregadas para realizar a tarefa agrupamento sobre a base de dados *Breast Cancer Wisconsin*. O algoritmo usado no agrupamento de dados foi o médias-K (Russell, 1995). Os resultados do agrupamento podem ser visualizados nos gráficos da Figura2. O experimento foi repetido trinta vezes e observou-se na média um ganho 44% no desempenho em termos de tempo de computação gasto no cálculo das distâncias, em relação ao procedimento sequencial. Esta diferença foi verificada com o teste T de Student para alfa 0,05.

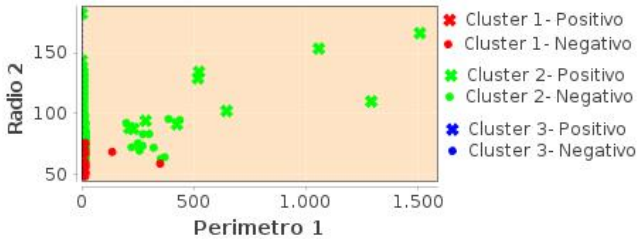


Figura 2 – Gráfico de visualização gerado pela aplicação.

Conclusões

Este trabalho apresentou um procedimento concorrente que calcula as distâncias entre distribuições de probabilidades da variável de classe entre os casos de uma base de dados supervisionada. O procedimento permite o uso de diferentes medidas de divergência no cálculo destas distâncias. Os resultados indicaram que o procedimento proposto tem um desempenho significativamente mais rápido do que aquele que usa processamento sequencial.

Agradecimentos

Ao CNPq, à UEPG e à Fundação Araucária pela oportunidade e apoio financeiro.

Referências

CHICKERING, David Maxwell. Learning Bayesian networks is NP-complete. **Learning from data: Artificial intelligence and statistics V**, v. 112, p. 121-130, 1996.

ETEMADPOUR, Ronak et al. Choosing Visualization Techniques for Multidimensional Data Projection Tasks: A Guideline with Examples. In: **International Joint Conference on Computer Vision, Imaging and Computer Graphics**. Springer International Publishing, 2015. p. 166-186.

FAYAD, Jamil Abdalla et al. Crescimento e produção do tomateiro cultivado sob condições de campo e de ambiente protegido. **Horticultura brasileira**, v. 19, n. 3, p. 232-237, 2001.

HARTLEY, Stephen J. **Concurrent programming: the Java programming language**. Oxford University Press, Inc., 1998.

KONTKANEN, Petri et al. Using Bayesian networks for visualizing highdimensional data. **Intelligent Data Analysis**, 2000.

RUSSELL, S.; NORVIG, P; INTELLIGENCE, Artificial. A modern approach. **Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs**, v. 25, p. 27, 1995.

SHEPARD, R. N. ROMNEY, A. K AND NERLOVE, S. (1972), Individual Differences and Multidimensional Scaling, in **Multidimensional Scaling**: Theory and Applications in the Behavioral Sciences, N.Y. Science Press, v, I, pp. 105-155.

SQUILLACOTE, Amy Henderson et al. **The paraview guide**. Kitware, 2006.

WARD, MATTHEW O., GEORGES GRINSTEIN, AND DANIEL KEIM. **Interactive data visualization: foundations, techniques, and applications**. CRC Press, 2010

WANG, Y., ZHANG, N. L., CHEN, T., & POON, L. K. (2011, June). **Latent tree classifier**. In ECSQARU (pp. 410-421) Springer, Berlin Heidelberg.