Universidade Federal de Pernambuco
Centro de Informática
Graduação em Engenharia da computação

Orientador: Francisco de Assis Tenório de Carvalho
Aluno: Gibson Belarmino Nunes Barbosa

"Block Clustering L2 para variáveis simbólicas do tipo intervalo"

Análise de agrupamento é um método muito utilizado em mineração de dados e reconhecimento de padrões, que visa particionar dados em grupos, subgrupos ou categorias, de acordo com o principio que indivíduos no mesmo grupo tenham grau de similaridade mais elevado que indivíduos em grupos diferentes.

Para uma descrição mais precisa do mundo real através de informações que possam ser processadas computacionalmente deve-se levar em conta a variabilidade e/ou incerteza dos dados, para tanto os dados podem ser descritos por conjuntos de categorias, histogramas, intervalos, distribuições de pesos, etc [De Carvalho, Brito e Bock 2006]. Para este trabalho utilizamos dados do tipo intervalo que levam em consideração o intervalo possíveis de valores para determinada variável, como exemplo teríamos o clima em uma cidade qualquer que poderia variar de 15°C à 29°C.

Enquanto alguns métodos têm como objetivo gerar partições de indivíduos com o máximo de precisão e alguns outros gerarem partições de variáveis, métodos de agrupamento baseados em blocos (ou block clustering methods) consideram o grupo de indivíduos e o de variáveis simultaneamente e organizam os dados em partições homogêneas, assim é possível com essas partições gerar conjuntos de dados menores que o de inicio para serem processados com menos tempo computacional, modos de utilização podem ser visto em [Govaert e Nadif 2002] [Govaert e Nadif 2005] [Govaert e Nadif 2007].

Estudos vêm sendo desenvolvidos visando o agrupamento de dados simbólicos com o mínimo de custo computacional e taxas de acertos mais altas [Bock 2002][Chavent,De Carvalho, Lechevallier e Verde 2006]. Como a técnica de block clustering para dados simbólicos no contexto de dados do tipo intervalo não foi explorada, buscamos com esse trabalho expor sua eficiência em termos de tempo de processamento e precisão (taxa de acertos).

Cronograma

	Maio		*Junho*				*Julho*				*Agosto*				*Setembro*		
Pesquisa bibliográfica			x	x	x	x	x	x	x	x							
Definir escopo e determinar tema	x	x	x	x													
Proposição do modelo			x	x	x												
Implementação do modelo					x	x	x										
Experimentos							x	x	x	x							
Análises comparativas							x	x	x	x	x						
Escrita do relatório final					x	x	x	x	x	x	x	x	x	x	x		
Preparação para apresentação															x	x	x

Preparação da proposta de TG – 20/05/13 até 21/06/13
Pesquisa bibliográfica – 02/06/13 até 26/07/13
Definir escopo e delimitar tema – 20/05/13 até 14/06/13
Proposição do modelo – 02/06/13 até 21/06/13
Implementação do modelo – 15/06/13 até 05/07/13
Experimentos – 01/07/13 até 26/07/13
Análises comparativas – 01/06/13 até 02/08/13
Escrita do relatório final – 15/06/13 até 06/09/13
Preparação para apresentação – 01/09/13
Defesa do TG – entre 01/09/13 e 24/09/13

Referencias

Bock H.H. (2002) “Clustering algorithms and kohonen maps for symbolic data. Proc.”, Journal of the Japanese Society of Computational Statistics, vol. 15, p. 1-13

Chavent, M., De Carvalho F.A.T., Lechevallier, Y. & Verde, R. (2006) “New clustering methods for inteval data”, Computational Statistics, vol. 21, p. 211-229

Govaert, G., Nadif, M. (2002) “Clustering with block mixture models”, Pattern Recognition, vol. 27, p.463-473

Govaert ,G., Nadif, M. (2005) “An EM Algorithm for the Block Mixture Model”, IEEE transactions on pattern analysis and machine intelligence, vol. 27, p.643-647

Govaert ,G., Nadif, M. (2007) “Block clustering with Bernoulli mixture models: Comparison of different approaches”, Computational Statistics & Data Analysis, vol. 52, p.3233-3245.

De Carvalho F.A.T., Brito, P., Bock, H.H (2006) “Dynamic Clustering for Interval Data Based on L2 Distance”, Computational Statistics, vol. 21, p.231-250

_____

Francisco de Assis Tenório de Carvalho (orientador)

_____

Gibson Belarmino Nunes Barbosa (proponente)