UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CENTRO DE INFORMÁTICA

2016.2

Clustering dinâmico de dados intervalares baseado em distâncias quadráticas adaptativas

Proposta de Trabalho de Graduação

Discente: George Harrison Alvares de Oliveira
Orientador: Francisco de Assis Tenório de Carvalho

Recife, Setembro de 2016

Contexto

Métodos de *clustering* buscam organizar um conjunto de objetos em grupos, ou *clusters*, de forma que os itens dentro de um mesmo *cluster* possuam um alto grau de *similaridade* entre si, enquanto que itens de *clusters* diferentes possuam um alto grau de *dissimilaridade* entre si. Estes métodos podem ser divididos em hierárquicos ou particionais.

Os métodos hierárquicos buscam produzir uma sequência aninhada dos dados de entrada, i.e., uma hierarquia, e podem ser aglomerativos ou divisivos. Métodos aglomerativos produzem uma hierarquia iniciando com os n itens de entrada como grupos triviais, onde cada item está em um único grupo, e realizando sucessivas fusões destes grupos para formar grupos maiores. Já os métodos divisivos produzem uma hierarquia iniciando com os n itens em um mesmo grupo e realizando sucessivas divisões deste em grupos menores.

Os métodos particionais buscam produzir uma única partição dos dados de entrada em um número fixo de *clusters* através da otimização de algum *critério de adequação*. De forma a obter melhores resultados, estes algoritmos são executados múltiplas vezes utilizando diferentes inicializações e a melhor partição obtida é dada como saída.

Algoritmos de *clustering* particionais dinâmicos são algoritmos iterativos que envolvem a construção de novos *clusters* e de representações, ou *protótipos*, para cada *cluster* a cada iteração através da otimização de algum critério de adequação entre os clusters e seus protótipos.

Algoritmos de *clustering* dinâmicos adaptativos também otimizam um critério de adequação baseado em alguma medida envolvendo os *clusters* e seus protótipos. A diferença é que as distâncias utilizadas para comparar os *clusters* aos seus protótipos não são absolutas e mudam a cada iteração, podendo ser diferentes para dois *clusters*. A vantagem dessas distâncias adaptativas é que o algoritmo é capaz de formar *clusters* de diferentes formas e tamanhos.

Motivação

Normalmente, os itens que serão agrupados são representados por um vetor de medidas qualitativas ou quantitativas, onde cada coluna representa uma variável, ou *feature*. Cada item assume um único valor para cada variável.

Na prática, esta modelagem é muito restrita e não é capaz de representar dados complexos, em que as variáveis podem ser intervalos, conjuntos de categorias ou distribuições de frequências, por exemplo. Tais tipos de dados são o escopo de estudo da Análise de Dados Simbólicos (SDA, do inglês *Symbolic Data Analysis*). O objetivo da SDA é estender técnicas clássicas da análise de dados, e.g., *clustering*, árvores de decisão, SVMs, a estes tipos de dados complexos chamados dados

simbólicos. A Tabela 1 traz exemplos de dados simbólicos, onde cada variável assume valores de intervalos de números reais.

	Frequência cardíaca	Pressão sistólica	Pressão diastólica
1	[60, 72]	[90, 130]	[70, 90]
2	[70, 112]	[110, 142]	[80, 108]
3	[54, 72]	[90, 100]	[50, 70]
4	[70, 100]	[130, 160]	[80, 110]
5	[63, 75]	[60, 100]	[140, 150]
6	[44, 68]	[90, 100]	[50, 70]

Tabela 1 Dados cardíacos de $n = 6$ pacientes e $p = 3$ variáveis intervalares

Objetivos

Neste trabalho, vamos propor um algoritmo de *clustering* dinâmico adaptativo para conjuntos de dados em que as variáveis assumem valores intervalares.

Cronograma

	Agosto		Setembro		Outubro		Novembro	
Elaboração da proposta								
Estudo da literatura e Estado da Arte								
Implementação								
Testes e Experimentos								
Escrita da monografia								
Preparação para apresentação oral								
Apresentação								

Assinaturas

_____

George Harrison Alvares de Oliveira

Discente

_____

Francisco de Assis Tenório de Carvalho

Orientador