

A survey for the essential Data Mining technique: Clustering

Isabel B. Amaro¹, Mirella M. Moro², Clodoveu Davis³

¹Department of Computer Science – Federal University of Minas Gerais (UFMG)
Belo Horizonte, Brazil

{isabel.amaro,mirella,clodoveu}@dcc.ufmg.br

Abstract. *Clustering is a Data Mining technique capable of group data with some non trivial similarity. Because of the advance of technology and data generation, clustering algorithms had to be developed and improved to process and extract useful information from the current large amount of data. Therefore, many areas that use Data Mining are constantly publishing articles with new clustering techniques and algorithms. This survey will study some state-of-the-art clustering algorithms, summarize, organize and present the temporal analysis in order to support the studies of its readers.*

1. Introduction

Um dos pontos de interesse principais do avanço da tecnologia é o oferecimento de serviços personalizados baseados nas informações adquiridas, podendo essas serem específicas de um usuário, de um texto ou de um banco de dados, por exemplo. Devido a isso, uma grande quantidade de dados vem sendo gerada nos últimos anos para esse consumo, despertando uma área de interesse da computação: mineração de dados.

A mineração de dados busca identificar padrões relevantes não triviais em um banco de dados raw. Para isso, a mineração de dados utiliza a técnica de clusterização, a qual consiste no agrupamento de dados semelhantes.

Por causa disso e de outras coisas, a mineração de dados vem sendo trabalhada cada vez mais em pesquisas. Novos algoritmos para clustering vem sendo desenvolvidos e publicados em artigos a cada ano, dificultando o entendimento do verdadeiro estado-da-arte, devido ao fato da computação ser uma área muito dinâmica. Nosso objetivo é analisar artigos de áreas diferentes que utilizaram técnicas de clustering diferentes para atingir seus objetivos de forma a auxiliar o leitor na busca de informações pelo estado da arte de algoritmos de clustering.

2. Different clustering case of use

A boa e rápida detecção de genes de doença é fundamental para prevenção e tratamento de casos na medicina atual. Esta foi uma das motivações de (Diseases, Text Mining), que utilizou a mineração de dados para extrair associações genéticas em um banco de dados médicos por meio de identificação de entidades de forma a economizar tempo e facilitar a detecção precoce de doenças como o câncer.

3. Clustering

4. Data Mining

5. Algorithms

5.1. K-Means

5.2. Hierarchical clustering

5.3. DBSCAN

6. Temporal analysis

7. Conclusion