World Wide Journal of Multidisciplinary Research and Development

World Wide Journal of Multidisciplinary Research and Development

WWJMRD 2016; 2(6): 58-60
www.wwjmrd.com
Impact Factor MJIF: 4.25
e-ISSN: 2454-6615

Sapna Mittal
Assistant Professor in
Computer Science Post
Graduate, Deptt of Computer
Science and Applications
R.S.D College, Ferozepur City,
India

Correspondence:
Sapna Mittal
Assistant Professor in
Computer Science Post
Graduate, Deptt of Computer
Science and Applications
R.S.D College, Ferozepur City,
India

A Review on Various Approaches of Clustering For Set of Instances in Data Mining

Sapna Mittal

Abstract

Data mining is the process of extraction of various types of information from different types of dataset that contains various types of attributes. is an approach that divides the whole information into different clusters. After processing of division of data values into different clusters centeroid have been computed. Cluster centeroid has been done on the basis of distance from other cluster members available in the particular clusters. The main problem in the clustering for data mining process is that text mining contains different problem for division of the text dataset into different cluster. Sometimes in the process of clustering by default empty cluster has been developed. To overcome this issue K-mean clustering has been used with hybridization of K-mediod algorithm.

Keywords: Data Mining, Clustering, K Mean Clustering, Honey Bee, Firefl, Ward

Introduction

1.1 Data mining

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is also known as Knowledge Discovery in Data (KDD)[9]. Data mining uses mathematical algorithms to part the data and evaluate the probability of future events. It automatically searches large volume of data to discover pattern and trend. Data mining software is one of a number of analytical tools for analyzing data. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is a powerful tool that can help to find patterns and relationships within our data. Data mining discovers hidden information from large databases.

1.2 Different Forms of Data Mining

- Spatiotemporal Data Mining: Spatiotemporal data are data that relate to both space and time. It refers to the process of discovering patterns and knowledge from spatiotemporal data.
- Multimedia Data Mining: It is discovery of interesting patterns from multimedia databases that store and manage large collections of multimedia objects, including image data, video data, and audio data.
- Web Mining: It is the application of data mining techniques to discover patterns, structures and knowledge from web.
- Spatial data mining: Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationship and spatial autocorrelation. Spatial data are the data related to objects that occupy space. A spatial database stores spatial objects represented by spatial data types and spatial relationship among such objects.
- Spatial Data Mining Techniques: There are different types of spatial data mining techniques. These are:
  1. Clustering
  2. Outlier Detection
  3. Association and Co-Location
  4. Classification
  5. Trend-Detection

1.3 clustering

Clustering is the process of partitioning a set of data objects into subsets such that the data elements in a cluster are similar to one another and different from the element of other cluster [9]. The set of cluster resulting from a cluster analysis can be referred to as a clustering. In this context, different clustering methods may generate different clustering's on the same data set. The partitioning is not performed by humans but by the clustering algorithm. Cluster analysis has wide range of application in business intelligence, image pattern recognition, web search, biology, and security [24]

- Spatial Clustering: Spatial clustering is a process of grouping a set of spatial objects into clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters [23]. For example, clustering is used to determine the "hot spots" in crime analysis and disease tracking. Hot spot analysis is the process of finding unusually dense event clusters across time and space. Many criminal justice agencies are exploring the benefits provided by computer technologies to identify crime hot spots in order to take preventive strategies such as deploying saturation patrols in hot spot areas. Spatial clustering can be applied to group similar spatial objects together; the implicit assumption is that patterns in space tend to be grouped rather than randomly located. However, the statistical significance of spatial clusters should be measured by testing the assumption in the data. The test is critical before proceeding with any serious clustering analyses.

1.4 Different types of Clustering Methods

These are:

- Partitioning Method: It first creates an initial set of k partitions, where parameter k is the number of partitions to construct. Typically partitioning method includes k-mean, k-medoid and CLARANS.
- K-mean Clustering:- The well-known k-means clustering is a typical example of partitioning techniques. It has become one of the most popular clustering methods because it is computationally easy to implement and it is generally accessible in most statistical software and clustering packages. The K-means clustering algorithm is a partitioning clustering method that separates data into k groups [9].Various algorithms have been developed to search for the optimal partition of the k-means clustering, which are frequently referred to as k-means algorithms since it involves the calculation of the mean (centroid) of each cluster.

2. Review of literature

Asmaa (2013) et al. in the paper "Text Clustering Using Statistical and Semantic Data" proposed a semantic process to improve the frequency mechanism with the semantic relations of the text documents. The proposed algorithm selects iteratively relevant features and performs clustering until convergence. To evaluate its performance, experiments on two corpora have been conducted. The obtained results show that the performance of their algorithm is superior to that obtained by the existing algorithms. Harpreet (2013) et al. in the paper "A Survey on Comparison between Biogeography Based Optimization and Other Optimization Method" survey comparison of BBO and other optimization. Optimization strategies have gained wide importance in solving complex problem. Biogeography Based Optimization (BBO) is a population based evolutionary algorithm that is based on the mathematics of biogeography. Many optimization algorithm such as genetic algorithm, Ant colony optimization etc are used in various field such as image processing. BBO algorithm works on the basis of two concept-Migration and Mutation. Xin-She (2013) et al. in the paper "Firefly Algorithm: Recent Advances and Applications" briefly review the fundamentals of firefly algorithm together with a selection of recent publications. Then, we discuss the optimality associated with balancing exploration and exploitation, which is essential for all metaheuristic algorithms. By comparing with intermittent search strategy, they conclude that metaheuristics such as firefly algorithm are better than the optimal intermittent search strategy. They also analyze algorithms and their implications for higher-dimensional optimization problems. Durga (2012) et al. in the paper "A Survey on Concept Based Mining Model using Various Clustering Techniques" introduced a new concept-based mining model that analyses terms on the sentence, document, and corpus levels. The concept-based mining model can effectively discriminate between non-important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. The proposed mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. The term which contributes to the sentence semantics is analysed on the sentence, document, and corpus levels rather than the traditional analysis of the document only. The proposed model can efficiently find significant matching concepts between documents, according to the semantics of their sentences. Sayantani (2012) et al. in the paper "A tutorial review on Text Mining Algorithms" described different existing Text Mining Algorithms i.e Classification Algorithm, Association Algorithm, Clustering Algorithm, stating the merits / demerits of the algorithms. In addition some alternate implementation of the algorithms is proposed. Finally the logic of these algorithms are, merged to generate an algorithm which will perform the task of Classification of a data set into some predefined classes, establish relationship between the classified date and finally cluster the data based on the association between them into groups. Li-Yeh (2012) et al. in the paper "An Improved Particle Swarm Optimization for Data Clustering" proposed an improved particle swarm optimization based on Gauss chaotic map for clustering. Gauss chaotic map adopts a random sequence with a random starting point as a parameter, and relies on this parameter to update the positions and velocities of the particles. It provides the significant chaos distribution to balance the exploration and exploitation capability for search process. This easy and fast function generates a random seed processes, and further improve the performance of PSO due to their unpredictability. In the experiments, the eight different clustering algorithms were extensively compared on six test data. The results indicate that the performance of their proposed method is significantly better than the performance of other algorithms for data clustering problem.

3. Approaches used

Firefly algorithm

The firefly algorithm (FA) is a metaheuristic algorithm, inspired by the flashing behavior of fireflies. The primary purpose for a firefly's flash is to act as a signal system to attract other fireflies. Xin-She Yang formulated this firefly algorithm by assuming:

1. All fireflies are unisexual, so that one firefly will be attracted to all other fireflies;
2. Attractiveness is proportional to their brightness, and for any two fireflies, the less bright one will be attracted by (and thus move to) the brighter one; however, the brightness can decrease as their distance increases;
3. If there are no fireflies brighter than a given firefly, it will move randomly.

Honey bee optimization

The Bees Algorithm is a new population-based search algorithm, first developed in 2005 by Pham DT and Karaboga. D independently. The bee's algorithm is a population-based search algorithm The Bees Algorithm is an optimisation algorithm inspired by the natural foraging behaviour of honey bees to find the optimal solution [17]. A honey-bee colony consists of queen(s) (best solution), drones (incumbent solutions), worker(s) (heuristic), and broods (trial solutions). A colony of honey bees can extend itself over long distances in multiple directions (more than 10 km) .Flower patches with plentiful amounts of nectar or pollen that can be collected with less effort should be visited by more bees, whereas patches with less nectar or pollen should receive fewer bees .The bees who return to the hive, evaluate the different patches depending on certain quality threshold (measured as a combination of some elements, such as sugar content). They deposit their nectar or pollen go to the "dance floor" to perform a "waggle dance".

Ward's algorithm

Ward's clustering method is implemented by reducing the number of clusters one at a time starting from one cluster per compound and ending which one cluster comprises all the compounds. At each cluster reduction, the method merges the two clusters and this will gives the result of the smallest increase in the total sum of squares of the distances of each point to its cluster centroid.Thus, the Ward's algorithm forms clusters by selecting a clusters that minimizes the within cluster sum of squares or the error sum of the squares

Conclusion

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. In the processing of data mining various approaches like classification, clustering has been used to divide information into different blocks for extraction of meaning full information. In the process of data mining huge datasets have been used for extraction of knowledge based information. After processing of division of data values into different clusters centeroid have been computed. Cluster centeroid has been done on the basis of distance from other cluster members available in the particular clusters. The main problem in the clustering for data mining process is that text mining contains different problem for division of the text dataset into different cluster. Sometimes in the process of clustering by default empty cluster has been developed. We removed this issue by using K-mean clustering with hybridization of K-mediod algorithm.

References

1. Asmaa Benghabrit, Brahim Ouhbi, Hicham Behja, Bouchra Frikh, "Text Clustering Using Statistical and Semantic Data" World Congress on Computer and Information Technology (WCCIT), pp. 1-6, 22-24 June 2013.
2. Harpreet Kaur, Gaganpreet Kaur, "A Survey on Comparison between Biogeography Based Optimization and Other Optimization Method" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 2, February 2013.
3. J.Durga, D.Sunitha, S.P.Narasimha, B.Tejeswini Sunand, "A Survey on Concept Based Mining Model using Various Clustering Techniques" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012.
4. Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay, "A tutorial review on Text Mining Algorithms" International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 4, June 2012.
5. Li-Yeh Chuang, Yu-Da Lin, and Cheng-Hong Yang, "An Improved Particle Swarm Optimization for Data Clustering" proceedings of the International multiconference on engineers and Computer Scientists 2012, Vol I, March 14-16, 2012.
6. M. Thangamani and P. Thangaraj, "Integrated Clustering and Feature Selection Scheme for Text Documents" Journal of Computer Science 6 (5): 536-541, 2010.
7. O.A. Mohamed Jafar and R. Sivakumar, "Ant-based Clustering Algorithms: A Brief Survey" International Journal of Computer Theory and Engineering, Vol. 2, No. 5, October, 2010.
8. Kilian Sto_el and Abdelkader Belkoniene, "Parallel k/h-Means Clustering for Large Data Sets" *gbif.ch/files/content/sites/imi/files/shared/documents/.../Parallel.pdf*
9. Marcelo N. Ribeiro, Manoel J. R. Neto, and Ricardo B. C. Prudˆencio, "Local feature selection in text clustering"*www.cin.ufpe.br/~rbcp/papers/ICONIP08.pdf.*
10. Xin-She Yang, Xingshi He, "Firefly Algorithm: Recent Advances and Applications" Int. J. of Swarm Intelligence, 2013 Vol.1, No.1, pp.36 – 50.
11. A. Abraham, He Guo and Hongbo Liu, "Swarm Intelligence: Foundations, Perspectives and Applications", Swarm Intelligence in Data Mining, A. Abraham, C. Crosan, V. Ramos (Eds.), Studies in Computational Intelligence (series), Springer, Germany, 2006.
12. A. Abraham and V. Ramos, "Web usage mining using artificial ant colony clustering and linear genetic programming", Proc. Congress on Evolutionary Computation (IEEE Press), Australia, 2003, pp.1384-1391.