Applied Soft Computing 19 (2014) 134–146

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data

T. Velmurugan *

PG and Research Department of Computer Science, D.G. Vaishnav College, Chennai - 600106, India

ARTICLE INFO

Article history:
Received 27 April 2012
Received in revised form 11 January 2014
Accepted 9 February 2014
Available online 19 February 2014

Keywords:
k-Means algorithm
Fuzzy C-Means algorithm
Data clustering
Data analysis
Telecommunication data

ABSTRACT

Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. Cluster analysis is often used as one of the major data analysis technique widely applied for many practical applications in emerging areas of data mining. Two of the most delegated, partition based clustering algorithms namely k-Means and Fuzzy C-Means are analyzed in this research work. These algorithms are implemented by means of practical approach to analyze its performance, based on their computational time. The telecommunication data is the source data for this analysis. The connection oriented broad band data is used to find the performance of the chosen algorithms. The distance (Euclidian distance) between the server locations and their connections are rearranged after processing the data. The computational complexity (execution time) of each algorithm is analyzed and the results are compared with one another. By comparing the result of this practical approach, it was found that the results obtained are more accurate, easy to understand and above all the time taken to process the data was substantially high in Fuzzy C-Means algorithm than the k-Means.

1. Introduction

Data mining (DM) is the extraction of useful and non-trivial information from the large amount of data that is possible to collect in many and diverse fields of science, business and engineering. DM is part of a bigger framework, referred to as knowledge discovery in databases (KDD) that covers a complex process from data preparation to knowledge modeling. Within this process, DM techniques and algorithms are the actual tools that analysts have at their disposal to find unknown patterns and correlation in the data. Typical DM tasks are classification (assign each record of a database to one of a predefined set of classes), clustering (find groups of records that are close according to some user defined metrics) and association rules (determine implication rules for a subset of record attributes). A considerable number of algorithms have been developed to perform these and others tasks, from many fields of science, from machine learning to statistics through various computing technologies like neural and fuzzy computing. What was a hand tailored set of case specific recipes, about ten years ago, is now recognized as a proper science. It is sufficient to consider the remarkable wide spectrum of applications where DM techniques are currently being applied to understand the ever growing interest from the research community in this domain.

Data mining can be viewed as an essential step in the process of knowledge discovery. Data are normally preprocessed through data cleaning, data integration, data selection, and data transformation and prepared for the mining task. Started as little more than a dry extension of DM techniques, DM is now bringing important contributions in crucial fields of investigations and in the traditional sciences like astronomy, high energy physics, biology and medicine [8] that have always provided a rich source of applications to data miners. An important field of application for data mining techniques is also the World Wide Web. The Web provides the ability to access one of the largest data repositories, which in most cases still remains to be analyzed and understood. Recently, data mining techniques are also being applied to social sciences, home land security and counter terrorism. A DM system is therefore composed of a software environment that provides all the functionalities to compose DM applications, and a hardware back-end onto which the DM applications are executed.

This paper discusses about one of the application areas of partition based clustering algorithms k-Means and Fuzzy C-Means by

* Correspondence to: Tel.: +91 9381032070.
E-mail address: velmurugan_dgvc@yahoo.co.in

http://dx.doi.org/10.1016/j.asoc.2014.02.011
1568-4946/

means of an experimental approach choosing a real time telecommunication data. Many applications have been proposed by using different algorithms. Now, it is necessary to discuss some of the applications of related areas. This will be helpful to understand the related breakthrough in computations and engineering applications. Ling-Zhong Lin and Tsuen-Ho Hsu give a study about designing a model of FANP in brand image decision-making in their paper [21]. They discuss that both theoretical and practical efforts in band images often neglect the characteristics having interactions and mutual influence among attributes or criteria, even in the stages of different brand life cycles. This study aims to create a hierarchical framework for brand image management. The analytical network process and fuzzy sets theory have been applied to both mindshare in brand images and inherent interaction/interdependencies among diverse information resources. A real empirical application is demonstrated in the department store. Both the theoretical and practical background of this paper have shown the fuzzy analytical network process can capture expert's knowledge existing in the form of incomplete and vague information for the mutual inspiration on attribute and criteria of brand image management.

Chien-wen Shen et al. presents a fuzzy AHP-based fault diagnosis for semiconductor lithography process [22]. This study proposes the approach of fuzzy analytic hierarchy process (FAHP) for the ambiguous fault evaluations of lithography process. The application of FAHP has several advantages over conventional approaches because it is able to enumerate the managerial causes of lithography faults and to homogenize the differences among the subjective judgments of on-site engineers. Together with the fuzzy set theory, this study provides a systematic mechanism to construct a hierarchy of FAHP model and a FAHP diagnosis map for the lithography process. A paper titled as "Application of quality function deployment to improve the quality of Internet shopping website interface design," is presented by Hui-Ming Kuo and Cheng-Wu Che [23]. They analyzed that the popularization and rapid development of the Internet has fostered the growth of online shopping leading it to become an important new channel for consumers to make purchases. The Internet users' rate of satisfaction has declined since online shopping has become an important consumer option. In order to improve customer satisfaction and to enhance the shopping experience, it is very important to understand the customer quality needs particular to the Internet shopping website, then to meet these needs through suitable website interface design. A B2C shopping website is used as an example in this study. Quality function deployment (QFD) is utilized to attain an understanding of customer quality needs, quality elements, and the relationship between them. Suggestions for improving the quality of website design are proposed based on the case study and the major performance indices discussed. Conclusions can be used as reference for online shopping website operators wishing to enhance the keenness of their websites in the highly competitive online shopping market.

"Fuzzy control of interconnected structural systems using the fuzzy Lyapunov method", a study carriedout by C.W. Chen [24]. In this study, a closed-form, easy-to-use fuzzy control method for interconnected structural systems is developed. First, the structural systems are reviewed. Interconnected schemes are employed to represent the structural systems and their subsystems. The representation of the interconnected systems consists of J interconnected subsystems. Stability is ensured by the criteria derived from the fuzzy Lyapunov functions. The fuzzy Lyapunov function is defined as the fuzzy blending of quadratic Lyapunov functions. Common solutions can be obtained by solving a set of linear matrix inequalities that are numerically feasible. To verify the effectiveness of the interconnected structural system, explanatory examples are presented for a practical tuned mass damper mounted on a building structure. Finally, the proposed control method is demonstrated using an example in which the interconnected technique is utilized to represent large-scale structural systems.

A paper presented by Cheng-Wu Chen titled as "Stability conditions of fuzzy systems and its application to structural and mechanical systems" [25]. This paper provided the stability conditions and controller design for a class of structural and mechanical systems represented by Takagi–Sugeno (T–S) fuzzy models. In this design procedure of controller, parallel-distributed compensation (PDC) scheme was utilized to construct a global fuzzy logic controller by blending all local state feedback controllers. A stability analysis was carried out not only for the fuzzy model but also for a real mechanical system. Furthermore, this control problem can be reduced to linear matrix inequalities (LMI) problems by the Schur complements and efficient interior-point algorithms are now available in Matlab toolbox to solve this problem. A simulation example was given to show the feasibility of the proposed fuzzy controller design method.

Data mining is to be performed on various types of databases and information repositories, but the kind of patterns to be found are specified by various data mining functionalities like class/concept description, association, correlation analysis, classification, prediction, cluster analysis etc. Among these, Cluster analysis is one of the major data analysis method widely used for many practical applications in emerging areas [8,11]. The quality of a clustering result depends on both the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns. There is a number of clustering techniques that have been proposed over the years [3]. Different clustering approaches may yield different results. The partitioning based algorithms are frequently used by many researchers for various applications in different domains. This research work deals two of the partitioning based clustering techniques as stated. With these discussions, about data mining and clustering methods, the next section discusses some of the application areas of the two algorithms.

The rest of the paper is structured as follows. Section 2 presents some of the application areas of chosen algorithms and their approaches. In Section 3, the basics of k-Means and fuzzy clustering methods are described. Section 4 discusses about the experimental setup of the proposed method is discussed. Section 5 covers the experimental studies, results, and a brief discussion. Finally, Section 6 presents the conclusions of the research work.

2. Applications of cluster analysis

Clustering means creating groups of objects based on their features in such a way that the objects belonging to the same groups are similar and those belonging in different groups are dissimilar. Clustering is one of the standard workhorse techniques in the field of data mining. Its intention is to systematize a dataset into a set of groups, or clusters, which contain similar data items, as measured by some distance function. The major applications of clustering include document categorization, scientific data analysis, customer/market segmentation and www. The other areas include pattern recognition, artificial intelligence, information technology, image processing, biology, psychology, and marketing. Some of the areas specifically used clustering concepts nowadays are:

- *Marketing*: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- *Land use*: Identification of areas of similar land use in an earth observation database.