Review of Literature Correlation Between Clustering and Classification of Novel Techniques in Computer Science

M Ramakrishna, Research Scholar, Department of Computer Science, SunRise University, Alwar, Rajasthan (India)
Dr. Praveen Kumar. Associate Professor, Department of Computer Science, SunRise University, Alwar, Rajasthan (India)

Abstract:

By clustering the data, people can obtain the data distribution, observe the character of each cluster, and make further study on particular clusters. In addition, cluster analysis usually acts as the preprocessing of other data mining operations. Therefore, cluster analysis has become a very active research topic in data mining. Data mining is a new technology, developing with database and artificial intelligence. It is a processing procedure of extracting credible, novel, effective and understandable patterns from database. Cluster analysis is an important data mining technique used to find data segmentation and pattern information... As the development of data mining, a number of clustering methods have been founded, The study of clustering technique from the perspective of statistics, based on the statistical theories, our paper make effort to combine statistical method with the computer algorithm technique, and introduce the existing excellent statistical methods, including factor analysis, correspondence analysis, and functional data analysis, into data mining. The present study is undertaken to develop a Data Mining workflow using clustering and classification of data, solving clustering problem as well as extracting potentially interesting association rules. Use the appropriate proximity measure, and to select the optimal clustering model to solve clustering problems. Develop a Data Mining workflow to extract potentially interesting association rules.

Keywords: Review Of Literature, Correlation, Clustering, Classification of Novel Techniques

Introduction:

Data clustering, by definition, is an exploratory and descriptive data analysis technique, which has gained a lot of attention, e.g., in statistics, data mining, pattern recognition etc. It is an explorative way to investigate multivariate data sets that contain possibly many different data types. These data sets differ from each other's in size with respect to a number of objects and dimensions, or they contain different data types etc. Undoubtedly, the data clustering belongs to the core methods of data mining, in which one focuses on large data sets with unknown underlying structure. The intention of this report is to be an introduction into specific parts of this methodology called cluster analysis. So-called partitioning-based clustering methods are flexible methods based on iterative relocation of data points between clusters. The quality of the solutions is measured by a clustering criterion. At each iteration, the iterative relocation algorithms reduce the value of the criterion function until convergence. By changing the clustering criterion, it is possible to construct robust clustering methods that are more insensitive to erroneous and missing

This greedy algorithm starts by randomly selecting a number of examples from the dataset as the initial set of representatives. Clusters are then created by assigning examples to the cluster of their closest representative. Starting from this randomly generated set of representatives, the algorithm tries to improve the quality of the clustering by adding a single non-representative example to the set of representatives as well as by removing a single representative from the set of representatives. The algorithm terminates if the solution quality (measured by q(X)) does not show any improvement. Moreover, we assume that the algorithm is run r (input parameter) times starting from a randomly generated initial set of representatives each time, reporting the best of the r solutions as its final result. The pseudo-code of algorithm SRIDHCR that was used for the evaluation of supervised clustering is given in Figure. It should be noted that the number of clusters k is not fixed for SRIDHCR; the algorithm searches for "good" values of k.

Clustering is the grouping together of similar data items into clusters. Clustering analysis is one of the main analytical methods in data mining; the method of clustering algorithm will influence the clustering results directly. This paper discusses the various types of algorithms like k-means clustering algorithms, etc…. and analyzes the advantages and shortcomings of the various algorithms. In each type, we can calculate the distance between each data object

and not all cluster centers in each iteration, which makes the efficiency of clustering is high. This paper provides a broad survey of the most basic techniques and identifies .This paper deals with the issues of clustering algorithm such as time complexity and accuracy to provide the better results based on various environments. The results are discussed on huge datasets. Mythili S1, Madhiya E2 2014 The World is overflowing with various kind of data like - scientific data, environmental data, financial data, and mathematical data. Manually analyzing, classifying, and pruning of the data is a tedious task for human, because the data is growing at a faster speed in this age of network and information sharing. Clustering is important in data analysis and data mining applications. A clustering method groups the data set into several data set based on the concept of maximizing the intra- class similarity and minimizing the inter- class similarity. This paper analyze or give an overview or review about the various clustering methods: Partitioning method, hierarchical method, Density based method, Grid based method, and Model based method in data mining. Kavita Nagar 2015

Review of Literature

Preparation of meaningful and more robust algorithms for effective clustering and classification of data, which in turn can be used for more effective data mining using computing methodologies. The list of techniques, which can be considered under such a definition, ranges from link analysis/associations, sequential patterns, analysis of time series, and classification by decision trees or neural networks, cluster analysis to scoring models. Hence, old and well-known statistical and mathematical as well as neural network methods get their new or resurrected role in Data Mining or more general in Business Intelligence. Since the 90s of the 20th century, with the information technology and the rapid development of Database technology, people can Very easy to access and store large amounts of data. The face of large-scale mass data, traditional data analysis work With only some surface treatment, but cannot get the inherent relationship between the data and the underlying information, from Fall into the "data rich, knowledge poor" dilemma [1]. To escape this dilemma, people urgently need a species can intelligently and automatically transform the data into useful information and knowledge of techniques and tools, which are on the strong force the urgent needs of data analysis tools make data mining (Data Mining) technology emerged [2]. Data mining in recent years with the database and artificial intelligence developed a new technology that the big amount of raw data to discover the hidden, useful information and knowledge to help policy makers to find the potential between the data Associated factors found to be ignored. Data mining because of its huge business prospect, are now becoming an international data library and information policy-making in the field of cutting-edge research, and caused extensive academic and industry relations note [3]. At present, data mining has been in business management, production control, electronic commerce, market analysis and scientific science and many other fields to explore a wide range of applications [4]. The face of huge amounts of data, the first task is to sort them out, cluster analysis is to classify the raw data as a reasonable way. The so-called clustering is a group of physical or abstract objects, according to the degree of similarity between them, divided into several groups, and makes the same data objects within a group of high similarity, and different groups of data objects are not similar [5][6]. As an important function of data mining, clustering analysis can serve as a stand-alone tool to get data on the distribution of observed characteristics of each class, focus on a specific class to do some further analysis Data mining aims to discover hidden from the database, meaningful knowledge, mainly into the following categories function [12]: (1). Concept description Concept description is called as summary description, which aims to concentrate the data, given its comprehensive descriptions, or will compare it with other objects. By summing up the data, you can achieve an overall grasp of the data. Description of the most simplest concept is the use of statistics in the traditional method to calculate the various data items in the database total, mean, variance, etc., or use OL "(On Line Processing, online analytical processing) achieve multi-dimensional query and calculation of data. (2). Correlation Analysis Correlation analysis found that large amounts of data items from the set of interesting association or correlation between the contacts. With the large number of

continuously collect and store data and many people in the industry from their database for mining association rules increasingly the more interesting. Records from a large number of business services found interesting correlation could help many business decisions making. (3). Classification and Prediction Classification and prediction are two forms of data analysis can be used to extract models describing important data classes or pre-future trends measured data. Classification and Prediction of a wide range of applications, for example, you can create a classification model. On the bank's loan customers to classify, to reduce the risk of the loans; also through the establishment of the classification model the functioning of the factory machines to classify, to predict the occurrence of machine failure. (4). Cluster Analysis Category according to maximum similarity and minimize between-cluster similarity principle, makes the same class of objects with high similarity with other classes of objects is very similar. each cluster formation. Class can be seen as an object class, which it can export rules. Clustering is also easy to observe the contents of the organization into hierarchical structure to organize similar events together. (5). Outlier Analysis Database may contain data objects, their general behavior with the data or the model inconsistent. This data objects are outliers. Many data mining algorithms attempt to minimize the impact of outliers, or row. In addition to them, however, in some applications may be an isolated point of a very important message. For example, in fraud detection, isolated points may indicate fraud. (6). Time Series Analysis In time series analysis, the data attribute value is changing over time. These data generally equal time intervals to obtain, but cannot get equal time intervals. Through the time, series map can be time-series data visualization. There are three basic functions in time series analysis: First, dig mode excavation, that is, by analyzing the time series of historical patterns to study the behavioral characteristics of affairs. Second, trend analysis that is, using historical data for time series forecasting the future value. Third, similarity search, which uses distance measures to ensure, given the similarity of different time series.

Customer analysis is crucial phase for companies in order to create new campaign for their existing customers. If a company can understand customer features and make efforts to fulfill their wants and provide friendly service then the customer will be supportive to the enterprise. The aim of this study was to develop a methodology to identify the characteristics of customers. It involved identification of the demographic characteristics of customers based on the analysis of categorical data using data mining clustering methods. The extracted knowledge can help companies identify valuable customers, and enable companies to make efficient knowledge-driven decisions. Customer analysis is crucial phase for companies in order to create new campaign for their existing customers. Companies are able to group or cluster certain customers, which have similar features. This may assist companies to make better marketing strategies over certain customer groups. Companies recognize that their existing customer database is their most important asset (Athanassopoulos, 2000). It is also important that how to effectively process and use customer data. Thus, this new techniques to assist analyze, comprehend or even visualize the massive amounts of stored data obtained from business and scientific applications (Liao et al, 2004). Data mining is the process of discovering and extracting considerable customer knowledge, such as rules, patterns, associations, clusters, and significant structures from large amounts of data stored in databases (Liao et al., 2008; Coussement et al., 2010). According to a research conducted by Reinartz et al., it is more beneficial to keep and satisfy existing customers than to constantly attract new customers who are characterized by a high attrition rate (Reinartz et al., 2003). Thus, if a company can understand customer features and make efforts to fulfill their wants and provide friendly service then the customer will be supportive to the enterprise. For instance, specific measures and motivation may be proposed to the most risky customer groups, i.e. the most disposed to leave the company, they may remain constant (Burez et al., 2007). Classification and patterns extraction from customer data is very important for business support and decision-making. Timely identification of newly emerging trends is very important in business process. Large companies are having huge volume of data but starving for knowledge. To overcome the organization current issue, the new breed of technique is required that has intelligence and capability to solve the knowledge scarcity and

the technique is called Data mining. The objectives of this paper are to identify the high-profit, high-value and low-risk customers by one of the data mining technique – customer clustering. In the first phase, cleansing the data and developed the patterns via demographic clustering algorithm using IBM IMiner. In the second phase, profiling the data, develop the clusters and identify the high-value low-risk customers. This cluster typically represents the 10-20 percent of customers, which yields 80% of the revenue.

References:

1. A perusal of big data classification and Hadoop technology, International Transaction of electrical and computer engineers system, 2017, Vol 4, No.1, 26-38
2. https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwinsqrmvpjWAhVHtY8KHYdPAxIQFgglMAA&url=http%3A%2F%2Fwww.dtic.mil%2Fget-tr-doc%2Fpdf%3FAD%3DAD0699616&usg=AFQjCNEFrzqvpb8-qG_iAhvSzFRKpkEog
3. http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf
4. http://www.ijarcsms.com/docs/paper/volume2/issue12/V2I12-0095.pdf
5. M. and Heckerman, D., "An experimental comparison of several clustering and initialization method", Technical Report MSRTR-98-06, Microsoft
6. Research, Redmond, WA, February 1998.
7. Mrs. Bharati M. Ramageri, "Data Mining Techniques and Applications," Indian Journal of Computer Science and Engineering, Vol. 1 No. 4, pp.301-305, 2010.
8. Karimella Vikram and Niraj Upadhayaya, "Data Mining Tools and Techniques: a review," Computer Engineering and Intelligent Systems, Vol 2, No.8, Pp.31-39, 2011.
9. Usama Fayyad, G. Paitetsky-Shapiro, and Padhrais Smith, "knowledge discovery and data mining: Towards a unifying framework", proceedings of the International Conference on Knowledge Discovery and Data Mining, pp. 82-22, 1996
10. A Novel Method for Text and Non-Text Segmentation in Document Images International conference on Communication and Signal Processing, April 3-5, 2013, India,
11. Novel Techniques on feature clustering algorithms for text classification IJCST Vol 4, Issue Spl-4 Oct-Dec, 2013. ISSN: 2229-4333 (Print)
12. A Novel Method for text and Non-Text segmentation in document images. International Conference on Communication and signal processing, April 3-5 , 2013 978-1-4673-4866-9/13 2013 IEEE