Survey of Clustering Applications

Kamran Shaukat Dar
University of the Punjab Jhelum Campus, Jhelum. Pakistan.

Imran Javed
University of the Punjab PUCIT, Lahore. Pakistan.

Warda Amjad
University of the Punjab Jhelum Campus, Jhelum. Pakistan.

Samreena Aslam
University of the Punjab Jhelum Campus, Jhelum. Pakistan.

Aroosa Shamim
University of the Punjab Jhelum Campus, Jhelum. Pakistan.

Abstract — Data mining is the process of collecting and analyzing useful patterns from huge amount of data, it has five major functions, clustering is one of them. In clustering, we make clusters of same data. The items in one group of cluster are alike while different from items which are in some other group of cluster. In image segmentation clustering approach is used for making segments of images and identifying the disease, for this purpose the Fuzzy C-means algorithm is used. Clustering is used in social network analysis using collaborative filtering recommendation, in which memory-based and model-based algorithms are used. K-means algorithm is used in credit card fraud detection for identifying the clusters. It works along the HMM model for identifying credit card frauds. In medical world for disease analysis an improved Fuzzy C-means algorithm which is termed as 3SW-FCM algorithm is used which can analyze the characteristics of disease and provide an accurate base for doctor's diagnosis. Greedy Agglomerative technique is a clustering category used for combining objects together grounded on resemblance which is used in grouping with respect to Reference Matching of High-Dimensional records.

Index Terms - Data Mining, Clustering, Social Network Analysis, Image Segmentation, Clustering applications, survey

1. INTRODUCTION

Clustering is the process of merging a set of items in such a manner that items in the identical set are more alike than to those in other sets. It's a core job of examining DM, and a usual method for statistical records study, some of It's applications are: Plant and animal ecology, Sequence analysis, Human genetic clustering, Medical imaging, Market research, Grouping of shopping items, Social network analysis, Software evolution, Image segmentation, Evolutionary algorithms, Crime analysis, Recommender systems, Petroleum geology, and Physical geography. There are different representations of clustering according to different researches which are: Connectivity representations, Centroid representations, Distribution representations, Density representations, Subspace representations, Graph-based representations.

We are summarizing five major applications of clustering, which are: Clustering in image segmentation, Clustering for network analysis using recommendation system, Clustering in disease analysis, Clustering for huge dimensional records with reference similarity and Clustering in credit card fraud detection. In our survey we are briefly covering all the techniques and algorithms which are being used in the above mentioned papers.

2. LITERATURE REVIEW

As technology is increasing rapidly in this era, huge amount of data is producing. It has become important to control and manage data into a valuable source. Different technologies are being used to maintain data in useful ways, data mining is one of them. DM is the process in which we examine the data from different viewpoints and reduce it into useful information, information which is further used to increase profit and it cuts cost. DM tools forecast future developments and actions, which allows the businessmen to make positive judgments which are knowledge determined. DM tools save time and can response to those business queries that were too time consuming to solve. It permits users to examine the records from different bases, classify it, and precise the recognized relations. DM functions and algorithms are used in different business sectors, they predict and discover trends.

It has five major functions. Cluster Analysis, Generalization, Association and Correlation Analysis, Classification, Outlier Analysis. Clustering is one of the major and most widely using rule, it is almost using in every filed. Clustering is the

procedure in which we put those items in one group that are more alike to each other and they are dissimilar from those items that are in other groups. It is used for exploratory data mining techniques, and it is a collective procedure to examine data. Clustering is used in many grounds, including the machine knowledge, in pattern learning. in image examination, in information retrieval, and in bioinformatics. There are different types of clustering models according to different researches which are: Connectivity representations, Centroid representations, Distribution representations, Density representations, Subspace representations, Graph-based representations. Broadly used clustering algorithms are: the K-means, the Fuzzy C-means, the Hierarchical clustering and the Mixture of Gaussians.

Clustering analysis has been widely used in every filed of life and it has different applications:

Clustering approach is used in collaborative filtering using recommendation system. Collaborative filtering is the collective record of the buyers and the items. Recommendation system, recommends the items on the basis of their ratings. Recommendation system can be separated into three classes: Collaborative filtering, content based recommendation and hybrid approaches. Collaborative Filtering is a extensively used method in recommender systems. There are two types of methods in used in CF which are either memory-based or model-based.

Clustering application is being used to categorize the companies in stock market according to their similarities in profit and loss .Grouping of companies in stock market is supportive for the investors, mangers and policy creators. A three-phase clustering presentation is used to classify companies, on the roots of resemblances in the figure of their stock market. At first, a low-determination time sequences facts are used to classify the companies. After that, during the second stage, pre-grouped companies are divided into some pure sub-groups. Then the sub-groups are combined in the third stage.

Clustering application is used in image segmentation by using Fuzzy C-means technique. Image segmentation is the actual partition of an image into a quantity of non-overlying districts, which have similar features like gray level, tone, color and texture for computer vision and understanding. Szilagyi have proposed the enhanced FCM (EnFCM) and Cai fast have generalized FCM (FGFCM) algorithm, in which there is small computational time but these algorithms are not able to straightly relate on the unique image. To overcome and solve these problems Stelios have proposed the fuzzy limited information c-means clustering algorithm (FLICM), this clustering algorithms is found on kerne techniques, which are applied to many fields of image partition.

Clustering application is used in credit card fraud detection. Due to the quick expansion in the electronic business methods, the usage of credit cards has been enlarged intensely. The most common method of payment for online purchases is credit card, in developed countries most of the money is paid through credit card for online and offline both type of transactions. As the usage of credit card has dramatically increased worldwide, opportunities for attackers to hack credit card information has increased, and they make fake transaction easily. Attackers need only little information related to the cardholder to hack the card. Clustering is used for modeling the arrangement of procedures which are used in credit card operation handling by using a Hidden Markov Model and shows how it can be used for detecting frauds in credit cards. This model (HMM) is very useful for credit card fraud recognition and it gives the instant reports, while the fraudulent transactions are on process.

Clustering is used in high-dimensional records with application to reference similarity. Many problems include clustering large datasets. A new method for clustering these giant, high-dimensional records is given. The idea holds using an inexpensive, estimated distance quantity to professionally split the records into overlying subgroups named canopies. The clustering is achieved by measuring the precise spaces only among points that arise in a common covering. Using canopies methods large clustering issues can be solved. Canopies methods are beneficial to many areas and used with a diversity of clustering methodologies, including the Greedy Agglomerative Grouping, the K-means and the Expectation-Maximization.

Clustering methods are used in medical analysis. With the development of DM methods, the clustering examination method is the active tool for medical inquiry .The clustering technique plays a significant part in examining the characteristics of diseases and offers accurate root for doctor's diagnosis. FCM algorithm is presented, which are commonly used in the examination of the disease detection. But the accuracy and the reply time have some drawbacks in the outdated FCM algorithm. Therefore, an enhanced algorithm, in which the combination of the point concentration weighted and it regulates the ideal number of groups based on the leveling methods, is suggested. Its results are better than outdated FCM algorithm.

3. APPLICATIONS

3.1. Fuzzy C-Means Grouping for Image division:

The image division is the procedure that can segment (divide an image into small parts) the images to confirm disease and to recognize disease creation portions. It divides images to definite non-overlapping sections which have same features such as the gray level, color tones and texture. Many

algorithms which based on clustering procedures have been suggested for image division. Among of all these procedures, the most popular procedure is Fuzzy grouping, which is more viable and retain more information about an image than the hard clustering. The most famous and broadly used clustering technique is the Fuzzy C-means technique, firstly suggested by Dunn [4] and future enhanced by Bezdek [5]. Conventional FCM method is used for this reason and it works fine on the noise-free pictures, but fails to divide the pictures ruined by sound and outliers. To overcome this problem an improved FCM_S method was given, but it has one shortcoming which is the spatial neighborhood method have to be calculated in every iteration phase, so this is an extremely time taking method. To cut the computational stages and cost, Chen Zhang [6] have given two variants, which are (FCM-S1 and FCM-S2). Then the Szilagyi [7] offered the improved FCM (EnFCM), to speed up the image division procedure, hence the computational time of the EnFCM method is very small than other grouping methods. Then the rapid prodeuced FCM (FGFCM) technique offered by Cai [12] is a better method of FCM, its calculation time is very small. On the other hand, these methods cannot directly work on the unique image; they need some instructions which is not at all an easy job. In the result of parameter selection difficultythe Setlios [1] presented a local information C-means method FLICM, which maintain the image division performance by describing Fuzzy factor. Then kernel distance measure method is introduced to expand the performance of FLICM that is getting more consideration in machine learning community. So, the technique which is founded on kernel scheme has been useful to enormous fields of image division. Zhan [12] introduced a new kernel-method brought distance measure into the main function of FCM(KCM) to advance the predictable measure. Chen [6] recommended two variants of KCM, KFCM-S1 and KFCM-S2 to reshape the computational cost of image division. The FLICM use the Euclidean distance which direct to unsatisfactory results on division of image. Therefore some researchers implement another method so-called robust distance measures to decrease the result of outliers on grouping outcomes. Motivated by all these methods of clustering an improved FLICM term as KWFLICM algorithm is proposed which is the combination of kernel method and weighted fuzzy aspect. This proposed KWFLICM algorithm is tested and compared with the other clustering algorithms which are NNcut method, FLICM, RFLICM, WFLICM. The results indicate that the suggested technique as compare to other four techniques can eliminate the noise whereas maintaining the image facts and attain the brilliant presentation of image division.

3.2. Clustering approach for collaborative filtering reference:

Over the past few years huge amount of data is being uploaded on world-wide web and all over the social networks. It is very necessary to manage the data and store it in useful patterns. Clustering method using collaborative filtering with the help of recommender system is used to search the items based on their previous records. Recommender system, recommends the items on the basis of their previous rating to users. The items which are highly rated are more recommended than low rating items. Collaborating filtering is used in recommendation system. There are two kinds of collective filtering: 1) memory-based 2) model-based. In memory-based system works on the entire matrix of consumers rating and generates the recommendations by identifying the items. Model-based techniques construct a model based on the previous ratings and the model is used for recommendations. In the memory based collaborative filtering, previous ratings of users are used to make entire matrix of an item, because there's the Possibility that the new user like to buy the item on the basis of previous suggestion and ratings. In collaborative filtering algorithms similarity computation is the most important step. The main indication of comparison calculation is co-rating of item I and j. If it's a consumer centered CF, relationship between two consumers is calculated using the items which have been ranked by both customers and for item-based CF, comparison between two substances is calculated by working on the consumers who have ranked both of these items. There are many techniques to analyze similarity, but the most widely used are, Pearson-correlation and vector-Cosine. After discovering the resemblance calculation, CF techniques have to invent out the most alike customers for the vigorous customer. The most vital phase because the references are made using the scores of neighbors and therefore neighborhood has an effect on the approval of products high scoring. The neighborhood choice strategy is selected by focusing on the similarity measures and the application fields. Then the references are made extremely rated item are extremely recommended and low rating are less recommended.

3.3. Clustering approach in credit card fraud Detection:

Ghosh and Reilly [8] invented the (credit card fault) CCF sensing machine that is skilled (checked) on the huge model of label credit card. Cho and Park [14] recommended a Hidden Markov Model built interference finding scheme that advances the displaying period and performance. Ourston et al [15] given the application of Hidden Markov Model which detects the multi stage of network attacks. Hoang et al [16] has also given new way to method the arrangement of system using HMM. HMM works with two grading levels. A HMM is a set of states that is related with a possibility distribution. A likely result or statement can be linked with symbol of observation of possibility distribution. Therefore states are hidden from the outside it is known as Hidden Markov Model. CCFD system is created on HMM, it do not required fake signs but is skilled of detecting fraud just by behavior in notice of a cardholder and particular acquired items in only

transaction that are unknown to any Credit card user. FDS that is consecutively in banks disputes recognition card to the cardholders. Each of operation is submitted to FDS and then it is used for confirmation purpose. The FDS accepts card information such as credit card number. Hidden Markov Model detects fake transaction thorough credit card, creates the group of training set then classify the outline of cardholder. It stores data of different amount of transactions in the form of cluster which can be in low, medium and high value ranges. For the security reasons, the SI is store in database, security form has different queries like account number, date of birth, mother name and the user have to give the correct answer to verify the operation unit. The information is only known by the card holder. The method works with two stages. 1) Training stage. 2) Detection stage.
Training Stage: This is vital part of the FDS. In this stage Hidden Markov Model is being skilled. For the teaching Hidden Markov Model, changes transaction of cardholder into statement symbols and form structures. At the end of teaching, we converted HMM consistent to each cardholder. This step is achieved offline, it do not disturb the credit card performance, which requests online answer. Although different type of clustering methods can be recycled, we use K-means clustering process to define the clusters. The basic steps of K-Means Grouping techniques are: We define the number of groups current and adopt it to be K and also adopt the middle of these clusters.

Detection Phase: Training stage is normally done offline, while detection is online procedure. When the Hidden Markov Model restrictions are erudite, we use the signs from the cardholder then design an initial arrangement of signals. The threshold value can be used empirically and the new arrangement is used as vile arrangement and use for defining the authority of following operation. The goal for counting non-malicious signs is for detecting the varying performance of cardholder.

3.4. Clustering technique for Diseases Analysis:

The clustering study can observe the characteristics of disease and give a accurate base for doctor's analysis. The outdated FCM algorithm used for this purpose but it is a limited research and not works fine for huge number of clusters. To reduce this problem the scientists required to convert the outdated algorithm to the Fuzzy c-means algorithm, which is grounded on smooth knowledge. In smooth knowledge, a smooth function is measured for precisely processing the facts and resulting global ideal solution. An improved FCM algorithm combines the two features; point density and the process of determine the best cluster to get a good result. The clustering algorithms are based on clustering method. The clustering methods are density based method, grid based method and model method, partitioning method and hierarchal method. The partitioning method in clustering algorithm is the method that splits the data into K partitions and every partition must belong to only one set of data. The typical algorithms that use this method are K-MEDIODS algorithm, CALARNS algorithm and K-means algorithm. Hierarchical method is a method which sets the hierarchy of given dataset, the algorithms which based on this method are CHAMELEON algorithm, BIRCH algorithm and CURE algorithm. And the algorithms which based on density method are DENCLUE algorithm, DBSCAN algorithm and OPTICS algorithm. The grid based method is a fast processing method, and the algorithms which based on this method are WAVE-CLUSTER, CLIQUE and STING algorithms. The algorithms which based on model, for that algorithms we consider a model, that contain a dataset, processing data set, features extraction, the selection and design of models and algorithms, cluster analysis, result analysis and new knowledge. Many researchers are working on clustering-algorithm to analysis the medical information and Alizadeh [9][10][11] used the hierarchical based algorithm in observing the tumor data[13]. Motivated by all the mentioned algorithms the new algorithms 3SW-FCM algorithm is proposed which based on smooth technology. This new proposed algorithm combines the point density as weighting factor, the method of determining the optima-number of cluster and Fuzzy c-means clustering algorithm. The FCM is a clustering algorithm which first proposed by Dumm [2] and future enhanced by Bezdek [3]. Through experiment it is proved that this proposed 3SW-FCM algorithm is more effective algorithm than the traditional FCM algorithm which has some disadvantages in accuracy and the response time.

3.5. In Grouping of High-Dimensional records with Reference Similarity:

Clustering methods is used in many important areas. By grouping patient histories, health care can be revealed. By making groups of addresses, clustering can be used for making new classes. By grouping documents, ranked organizations are derived. Canopy technique is used to make groups of items. The main use of the canopy is that it can decrease the space required for grouping the records into overlying groups. Canopies are used by different source of information. Space metrics for text used by search engines and are based on the inverted index. Inverted index, is a matrix demonstration we can directly get into the list of forms. The use of inverted index can be applied to high-dimensional real-valued data [17]. Each data is effectively converted into documents. Greedy Agglomerative method, is a grouping type used for positioning objects together due to similarity. Canopies also used by Expectation Maximization clustering. This is another type of clustering used by canopies. This method is normally not used by canopies because they are unable to specify how many clusters to use. We can

formally use computational savings for the canopies technique. This method has two mechanisms: A phase where the canopies are designed followed by a low grouping procedure. If we create canopies by means of the inverted directory, we do not need to accomplish whole pair-wise space relations. If we generate groups by using the Greedy Agglomerative method, we have to provide the distance metric of bibliographic citations [18]. Many related work is perform for clustering the large data records and canopies method is used for grouping the similar data sets. Canopy clustering makes some problems but it reduces the repetition of data records.

4. CONCLUSION

In this survey, we have presented different clustering applications. In social network study the modified CF algorithm is used to produce the recommendation founded on user's marks and to find the groups of similar users a complex network grouping method is applied on records. HMM application is used in credit card fault detection, with different ranges of transaction amount the validity of HMM is checked. Clustering large records is a universal task which can be performed by making canopies. Canopy technique is extensively applicable. The clustering methods can be greedy agglomerative, K-nearest neighbor or K-means these methods are used for the creation of canopies that are used to measure the large data sets. Canopies have the ability in which all the objects in group fall in the identical canopy and there is no precision on those objects. In image segmentation an enhanced algorithm KWFLICM of traditional FCM algorithm is used to divide an image into small parts to provide a correct base for doctor's diagnosis. The improved form of outdated FCM algorithm term as 3SW-FCM algorithms, based on smooth technology is used for medical analysis.

REFERENCES

[1] S. Krinidis and V. Chatzis, "A robust fuzzy local information C-means clustering algorithm,"IEEE Trans. Image Process., vol. 19, no. 5, pp.1328–1337, May 2010.

[2] Sun Yang. "Fuzzy clustering in the research and application of intelligent medical diagnosis system". MS thesis. Zhejiang University, 2006.

[3] Yang Cuiqiong, Jiang Hong, and Yu Xiaolei. "An improved FCM clustering research". Computer and Digital Engineering. vol. 38.pp.1-3, 2010

[4] J. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," J. Cybern., vol. 3, no. 3,pp. 32–57, 1974.

[5] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981.

[6] S. Chen and D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure,"IEEE Trans. Syst., Man, Cybern., B, Cybern., vol. 34, no. 4, pp. 1907–1916,Aug. 2004.

[7] L. Szilagyi, Z. Benyo, S. Szilagyii, and H. Adam, "MR brain image segmentation using an enhanced fuzzy C-means algorithm," in Proc. 25th Annu. Int. Conf. IEEE EMBS, Nov. 2003, pp. 17–21

[8] Ghosh, S., and Reilly, D.L., 1994. Credit Card Fraud Detection with a Neural-Network, 27th Hawaii International Conference on Information Systems, vol. 3 (2003), pp. 621- 630.

[9] Chae, Young Moon, et al. "Data mining approach to policy analysis in a health insurance domain." International journal of medical informatics. vol. 62, no. 2, pp. 103-111, 2001.

[10] Shi Yifang et al. "Data mining and knowledge discovery technology in patient flow analysis". Journal of Preventive Medicine.Vol. 33, no. 2, pp. 237-238, 2006 .

[11] Liu Mingxia, Ren Shiquan. "Self-organizing data mining in total health expenditure forecast". Health Economics Research. Vol.12, pp. 10-12. 2003.

[12] W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy C-means clustering algorithms incorporating local information for image segmentation,"Pattern Recognit., vol. 40, no. 3, pp. 825–838, Mar.2007

[13] Bouchachia, Abdelhamid, and Witold Pedrycz. "Data clustering with partial supervision." Data Mining and Knowledge Discoveryvol. 12, no. 1, pp. 47-48, 2006.

[14] S.B. Cho and H.J. Park, "Efficient Anomaly Detection by Modeling Privilege Flows Using Hidden Markov Model," Computer and Security, vol. 22, no. 1, pp. 45-55, 2003.

[15] D. Ourston, S. Matzner, W. Stump, and B. Hopkins, "Applications of Hidden Markov Models to Detecting MultiStage Network Attacks," Proc. 36th Ann. Hawaii Int'l Conf. System Sciences, vol. 9, pp. 334-344, 2003.

[16] X.D. Hoang, J. Hu, and P. Bertok, "A Multi-Layer Model for Anomaly Intrusion Detection Using Program Sequences of System Calls," Proc. 11th IEEE Int'l Conf. Networks, pp. 531-536, 2003.

[17] H. Hirsh. Integrating mulitple sources of informationin text classification using whril. In Snowbird Learning Conference, April 2000.

[18] D. Sankoff and J. B. Kruskal. Macromolecules: TheTheory and Practice of Sequence Comparison. Addison-Wesley, 1983