ISSN:0975-9646

Enhanced Incremental Clustering of Time-Series by Fuzzy Clustering

[1]K. Suresh Babu, [2]K.Priyanka, [3]A.Madhuri, [4]G.Chaitanya, [5]G.Gowri

[1] *Professor,* [2,3,4,5]*Student,Department of CSE,*
LENDI Institute Of Engineering & Technology, Vizianagaram, AP, India

Abstract: **Today, when we analyze the user's behavior, has gained more importance in the data mining community. Typically, the behavior of a user is defined as a time series of his or her activities. In this paper, users are clustered based upon the time series extracted from their behavior during the interaction with the given system. Although there are several different techniques to cluster time series and sequences, this paper will attack the problem by utilizing a novel incremental fuzzy clustering strategy in order to achieve the objective. In this paper we consider the customer groups. Segmentation of customers is done basing on their groups and also in time series where time series is the changes which occur in user database. In this incremental clustering of customers we are going for fuzzy clustering.**

Clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. We divide them depending on the time series i.e., the behavior of the user

Keywords: **Data mining, Incremental clustering, Fuzzy Clustering, Time-series.**

I.INTRODUCTION

Definition of Data Mining: The nontrivial extraction of implicit, previously unknown, and potentially useful information.

Data explosion problem:

Automated data collection tools and mature database technology lead to tremendous amount of data stored in database and other information repositories.

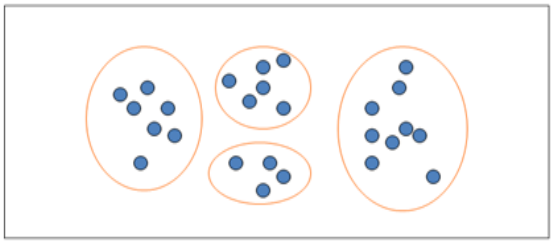
We are drowning the data, but starving for knowledge!

Solution: Data mining

Extraction of interesting knowledge (non-trivial, regularities, patterns, constraints)from data in large data base.

Now, various data mining systems are being used to examine the data within different domains. In data mining, clustering techniques are typically used in order to group data based upon their similarities.

Clustering: Each point represents the customers.



Grouping the members together who have similar characteristics

Applications:

Widely applied on fraud detection, business and finance, science Rapid computerization of businesses produce huge amount of data. The behaviour of users in interaction with various systems like finances, healthcare, and business is stored as historical data dynamically.

There are several researches involving time series clustering, and there are different techniques used in time series clustering. In these systems, extracted knowledge should be updated based upon new added objects or changes that take place in the existing objects over time.

We are using fuzzy clustering algorithm for new arrived time series (e.g. college placement application of new students, recruitments, and percentage details and others..) or changes in on hand time series (if student do not continue their activity in the college i.e. detained or some other reasons), we could justify the usage of the traditional techniques for clustering time series systems. If, however, a system accepts new time series or changes to existing time series over time, then it should be updated incrementally.

Another issue is limited space in memory to store entire time series data set in the memory for clustering. A large number of student details transactions are stored daily in databases on massive servers. In this case, the clustering of student based upon their behaviour forces of marks, recruitments in company to deal with a huge amount of multidimensional data. Based on observant projects in different sections of college, the data is usually converted to a lower dimension in a pre-processing phase for clustering purposes because of high computational costs.

The classification of customers and defining the behaviour of student are used for different purposes like aggregate or events in college with different aggregated fields in use such as student details, maximum and minimum qualifications, overall performance graph in a year.

II.PURPOSE

The main purpose of this paper is to present a detailed description of the "Enhanced Incremental Clustering of Time series by Fuzzy Clustering". It is about the clustering of the Customer Data based on their time series. We are planning to group the similar customer records together in a cluster, based on time series similarity, where the Datasets are incremented without overlapping.

III.PROBLEM DEFINITION

Databases are the storage areas for large amount of data. Databases provide flexible operations for manipulation of the data. But,when the amount of data increases, these

operations of retrieving and storing of data becomes complex. When data such as details of daily transactions of a customer or number of people using a website in a day or student details of a particular organisation are stored in a database without any differentiation, the problem arises when any organisation wants to retrieve a particular data specific to a particular time. Also this process takes a long time.

IV. MEANS OF SOLUTION

This paper is about the clustering of the Customer Data based on slowly changing dimensions. We are planning to group the similar customer records together in a cluster, based on time series similarity, where the Datasets are incremented without overlapping. After grouping the similar records into clusters, the clusters are stored in the database and finally the unwanted clusters can be eliminated. The clusters are updated incrementally and periodically.

Brief Review Of Fuzzy C-Means Algorithm:

In this paper, to group the similar data or records into clusters we use the Fuzzy C-Means (FCM) algorithm. FCM works by partitioning a collection of 'n' vectors into 'c' fuzzy groups and finds a cluster centre in each group such that the dissimilarity measure is minimized. Given c as number of classes, centre of class j for j = {1,…, c}, n as the number of time series and μij as the degree of membership of the time series *i* to cluster j for i = {1, …, n}, distance of each time series Fi from each cluster center(vj) is denoted as such:

$$dij = (Fi, vj) = DLCSS(Fi, vj).$$

Let the centers be shown by vj = {v1,…,vc} and each time series by Fi that i = {1, …,n} and dji as distances between centers and time series. Therefore, the membership valuesμij are obtained with:

$$\mu_j(x_i) = \frac{(\frac{1}{d_{ji}})^{\frac{1}{m-1}}}{\sum_{k=1}^{p}(\frac{1}{d_{ji}})^{\frac{1}{m-1}}}.$$

The FCM objective function (standard loss) that is attempted to be minimized takes the form:

$$J = \sum_{j=1}^{c} J_j = \sum_{j=1}^{c}\sum_{i=1}^{n}[\mu_{ij}]^m d_{ij}$$

where μij is a numerical value between [0; 1]; dij is the Euclidian distance between the jth prototype and the ith time series; and m is the exponential weight which influences the degree of fuzziness of the membership matrix. In order to update a new cluster center value in iterations, the following formula is employed:

$$v_j = \frac{\sum_{i=1}^{n}(\mu_{ij})^m (F_i)}{\sum_{i=1}^{n}(\mu_{ij})^m}, \forall j \in \{1, \ldots, c\}.$$

System Model:

Whenever a user enters the information and perform any transaction using the valid id and password, these details are stored in the database in the form of clusters using fuzzy clustering algorithm. The process that happens is, the similar data i.e data which was entered within a time period is separated into clusters.

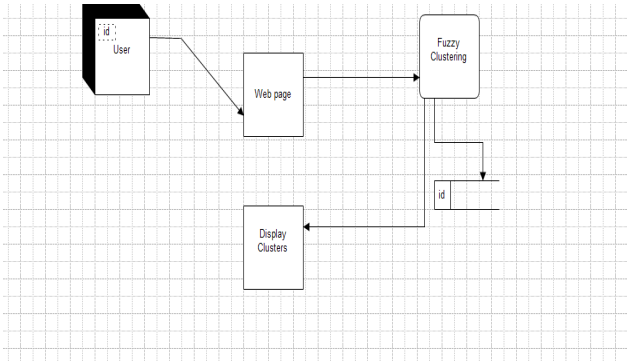
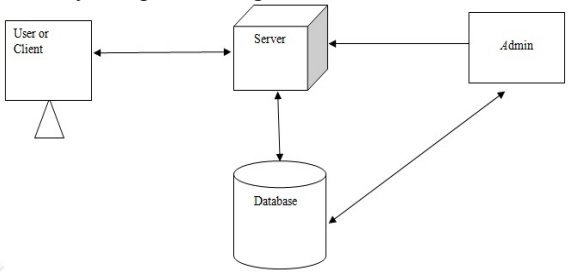


Fig: Dataflow Design

In the above fig., the user who wants to retrieve the data, access details by signing in with a valid user id and password. Then the request is processed by the server. The server retrieves the required data and then finally displays the information to the user or admin**.**

Architectural Design:

The architectural design is the design of the entire software system; it gives a high-level overview of the software system, such that the reader can more easily follow the more detailed descriptions in the later sections. It provides information on the decomposition of the system into modules (classes), dependencies between modules, hierarchy and partitioning of the software modules.

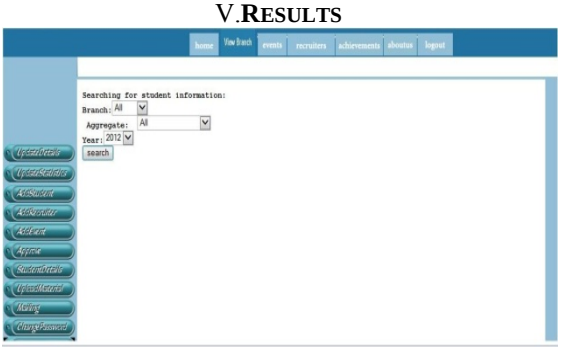


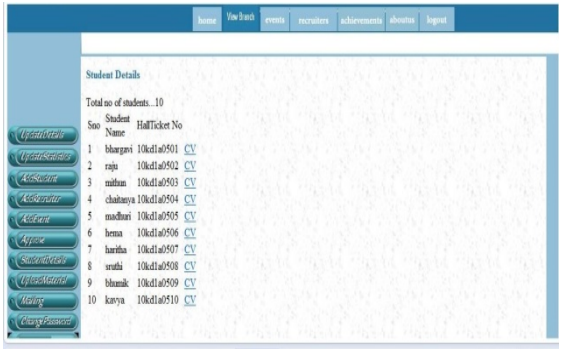
The User or Client logins into his account with the valid username and password and his details will be maintained in the database. And admin searches for the similar records and group them into clusters. He can also access that data which is stored in the database. The user can access the database only through the server but he don't have the direct access with the database. But, admin can access the details in the database directly or through the server.

The user first sends the request to the server, then the server processes that request and sends back the response to the user or client.

And the project we implement is a 3-tier Architecture as the user or client and the admin can access the database through the server.

V.RESULTS





VI.CONCLUSION

This paper provides a solution to the complex data retrieval problem by organising the data in database into clusters. Fuzzy mechanism is used to arrange only the similar data into clusters. So the user can easily access the data in small amount of time without much complexity. The time complexity can also be reduced by performing the search basing upon the keyword.

REFERENCES

1. M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On clustering validation techniques",*Journel of Intelligent Information Systems*, Vol. 17, 2001, pp. 107-145.

2. P. Cotofrei and K. Stoffel, Classification rules + time=temporal rules," in *Proceedings of International Conference on Computational Science*, Part 1, 2002, pp. 572-581.

3. T. Fu, F. Chung, V. Ng and R. Luk, "Pattern discovery from stock time series using self- organizing maps", in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Temporal Data Mining*, 2001, pp. 27-37.

4. V.Kavitha and M. Punithavalli, "Clustering time-series data stream-A literature survey," *International Journal of Computer Science and Information Security*, Vol. 8, 2010, pp. 289-294.

5. E. Keogh and C. Ratanamahatana, "Exact indexing of dynamic time warping**,**" *Knowledge and Information Systems*, Vol. 7, 2005, pp. 358-386