A Novel Data Classification using Fuzzy C-Means Clustering with Privacy Preserving Data Mining

B.Karthika[1] Dr.J.Suguna[2]
[1]Research scholar [2]Associate Professor
[1,2]Department of Computer Science Engineering
[1,2]Vellalar College for Women, Erode-12 Tamilnadu, India

Abstract— In data mining, clustering is a method of grouping data into different groups, so that the data in each group share similar trends and patterns. The fuzzy clustering is a promising approach that works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. The main focus of this paper is to hide certain confidential data so that they cannot be discovered through data mining techniques. In this work, the Fuzzy C-means algorithm is used for clustering the medical data set based on various attributes by preserving the privacy of the data by AES encryption and decryption method. Here, SPARCS Medical data set is used for processing and the system is developed using MATLAB. It is found that the proposed system reduces the computation time, increases the accuracy and F-measure is used to validate the quality of the cluster obtained.

Key words: Data mining, Fuzzy C-means algorithm, Clustering

I. INTRODUCTION

Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering [1]. Traditionally data mining techniques were used in various domains. However, it is introduced relatively late into the Healthcare domain. Nevertheless, as on today lot of research is found in the literature. This has led to the development of intelligent systems and decision support systems in Healthcare domain for accurate diagnosis of diseases, predicting the severity of various diseases, and remote health monitoring. Especially the data mining techniques are more useful in predicting heart diseases, lung cancer, and breast cancer and so on [9].

Data mining is the process of extraction of hidden, predictive information from large databases [1]. The overall Knowledge Discovery and Data Mining (KDD) process deals with turning low level data into high level knowledge. The process of data mining begins with the understanding of the application domain. This includes relevant prior knowledge as well as the goals of the system. First, data cleaning and pre-processing is carried out on the raw data for removal of noise and handling of missing data. Next, data reduction and projection are performed to find the minimal set of features to represent the data [9]. An appropriate data mining model is used to extract the patterns for classification. Finally, the knowledge obtained is incorporated into the performance system.

The need for privacy is sometimes essential. Privacy-preserving data mining (PPDM) considers the problem of running data mining algorithms on confidential data that is not supposed to be revealed even to the party running the algorithm. The main consideration of PPDM is twofold (Verykios et al., 2004). First, sensitive raw data like identifiers, names, addresses and so on, should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy. So, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. The former is referred as individual privacy preservation and the latter is as referred collective privacy preservation (Stanley et al., 2004). Our basic approach to preserving privacy is to let users provide a modified value for sensitive attributes. A number of techniques such as Trust Third Party, Data perturbation technique, Secure Multiparty Computation and game theoretic approach, have been suggested in recent years in order to perform privacy preserving data mining.

In the proposed research work, the modified value may be generated using AES encrypted and decrypted method.

The four important steps in data mining are preprocessing, clustering, feature extraction and classification. In the field of clustering analysis, a number of methods have been put forward and many successful applications have been reported. Clustering algorithms can be loosely categorized into the following categories: hierarchical, partition-based, density-based, grid-based and model-based clustering algorithms. [5] Among them, partition-based algorithms which partition objects with some membership matrices are most widely studied.

The SPARCS medical data set is taken for Processing. AES Encryption and Decryption Method is used for encrypting and decrypting the dataset. Fuzzy C-means clustering algorithm is used to cluster the data. The results are evaluated for the quality using some validation measures.

The first chapter describes the introduction about data mining, confidentiality issues in data mining and also gives the need for Privacy preserving data mining with its methods. The second chapter describes the literature survey related to privacy preserving data mining and various clustering methods used .The third chapter depicts the AES Encryption and Decryption method. The fourth chapter discusses the datasets used by the system, intermediate results obtained, graphs showing the comparative analysis of various algorithms used and their validation measures. The fifth chapter concludes the thesis and discusses the scope for further work enhancement.

II. LITREATURE REVIEW

Changyu Dong et al. [2014] has proposed a fast secure dot product protocol with application to privacy preserving association rule mining. For privacy concerns, various

privacy preserving data mining techniques have been proposed. An efficient secure dot product protocol and its application in privacy preserving association rule mining shows one of the most widely used data mining techniques. The protocol is a order of magnitude faster than previous protocols because it employs mostly cheap cryptographic operations. The performance has been further improved by parallelization. They implemented the protocol and tested the performance. The efficiency comes from the fact that the protocol relies mostly on cheap cryptographic operations, i.e. hashing, modular multiplication and bit operations that in addition to association rule mining, secure dot product has also been shown to be an important building block in many other PPDM algorithms such as naive Bayes classification, finding K-Nearest Neighbors (KNN) and building decision trees. Therefore, this protocol can also be used to boost the performance of those PPDM tasks and it has proved that the protocol is secured in the semi-honest model in terms of multiparty secure computation.

Stanley R. M. Oliveira et al. [2001] has proposed Earlier Background on Privacy Preserving Clustering by Data Transformation revisited a family of geometric data transformation methods (GDTMs) that distort numerical attributes by scaling, rotations, translations. This method was designed to specify privacy-preserving clustering, in context where data owners must meet privacy requirements as well as guarantee valid clustering results. Authors also provided a particularized, broad and advanced picture of methods for privacy-preserving clustering by data transformation.

Preserving the privacy of individuals when data are shared for clustering was a complex problem. The challenge was how to protect the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis [8]. The geometric data transformation methods (GDTMs) that distort confidential numerical attributes in order to meet privacy protection in clustering analysis.

End users are able to use their own tools so that the constraint for privacy has to be applied before the mining process on the data by data transformation. Data owners must not only meet privacy requirements but also guarantee valid clustering results.

Dong Chen et al., [2014] has proposed a method which is almost based on the special encryption protocol known as Secure Multiparty Computation (SMC) technology. SMC originated with Yao's Millionaires' problem. The basic problem is that two millionaires would like to know who is richer, with neither revealing their net worth. Abstractly, the problem is to simply compare two numbers, each held by one party, without either party revealing its number to the other.

Cliffton. C, Vaidya. J et al. [2004] discussed the problem of decision tree learning with the popular ID3 algorithm which is considerably more efficient than generic solutions and demands both very few rounds of communication and reasonable bandwidth. It shows how the involved data mining problem of decision tree learning can be efficiently computed, with no party learning anything other than the output itself.

III. AES ENCRYPTION AND DECRYPTION METHOD

A. Encryption and Decryption:

It is the process of converting ordinary information (called plaintext) into unintelligible text called cipher text. To encrypt more than a small amount of data, symmetric encryption is used. A symmetric key is used during both the encryption and decryption processes. Decryption is the reverse in other words, moving from the unintelligible cipher text back to plaintext. A cipher is the pair of algorithms that create the encryption and the reversing encryption [2]. The detailed operation of the cipher is controlled both by the algorithm and in each instance by a "key". This is a secret (ideally known only to the communicants), usually a short string of characters, which is needed to decrypt the cipher text.

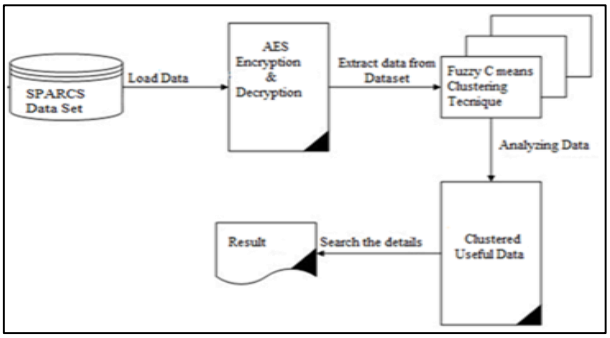

Fig. 3.1: System Architecture

B. Process of AES Encryption:

AES comprises of three block ciphers, AES-128, AES-192 and AES-256. Each cipher encrypts and decrypts data in blocks of 128 bits using cryptographic keys of 128-, 192- and 256-bits, respectively. Symmetric or secret-key ciphers use the same key for encrypting and decrypting, so both the sender and the receiver must know and use the same secret key.

The AES Decryption operation is based on 4 control signals and takes a total of 24 clock cycles. The first 12 clock cycles are used to load in the decryption key. On the first clock cycle, the 128-bit key is sent as input, and the 'key load' control signal is pulsed for only one clock cycle. 12 clock cycles later, the 'key done' control signal is asserted by the core as acknowledgement. After this step, the load control signal may be pulsed for one (and only one) clock cycle. On that cycle, the core reads in the 128-bit encrypted data block from text in and begins processing. After another 12 clock cycles, the done control signal is asserted and the system can read back the decrypted data from text out.

For correcting decryption, the key that is used for decryption must match the key used for encryption.

C. Fuzzy C-Means Clustering:

The Fuzzy C-means (FCM) algorithm is using the weights that minimize the total weighted mean-square error:

$$J(w_{qk}, z^{(k)}) = \Sigma_{(k=1,K)} \Sigma_{(k=1,K)} (w_{qk}) \| x^{(q)} - z^{(k)} \|^2 \quad (3)$$

$$\Sigma_{(k=1,K)} (w_{qk}) = 1 \text{ for each } q$$

$$w_{qk} = (1/(D_{qk})^2)^{1/(p-1)} / \Sigma_{(k=1,K)} (1/(D_{qk})^2)^{1/(p-1)} , \; p > 1 \quad (4)$$

The FCM allows each feature vector to belong to every cluster with a fuzzy truth value (between 0 and 1), which is computed using Equation (4). The algorithm assigns a feature vector to a cluster according to the maximum weight of the feature vector over all clusters [3].

D. Eliminating Empty Clusters:

After the fuzzy clustering loop it adds an Equation $\kappa = D_{min}^2 / \{\sum_{(k=1,K)} \sigma_k^2\}$ to eliminate the empty clusters. This step is put outside the fuzzy clustering loop and before calculation of modified XB validity [4]. Without the elimination, the minimum distance of prototype pair used in Equation $\kappa = D_{min}^2 / \{\sum_{(k=1,K)} \sigma_k^2\}$ may be the distance of empty cluster pair. It calls the method of eliminating small clusters by passing 0 to the process so it will only eliminate the empty clusters.

After the Fuzzy C-means iteration, for the purpose of comparison and to pick the optimal result, it add Equation (9) to calculate the cluster centers and the modified Xie-Beni clustering validity κ.

The Xie-Beni validity is a product of compactness and separation measures. The compactness-to-separation ratio ν is defined by Equation (6).

$$\nu = \{(1/K)\sum_{(k=1,K)} \sigma_k^2\}/D_{min}^2 \quad (6)$$

$$\sigma_k^2 = \sum_{(q=1,Q)} w_{qk} \| x^{(q)} - c^{(k)} \|^2 \quad (7)$$

D_{min} is the minimum distance between the cluster centers.
The Modified Xie-Beni validity κ is defined as

$$\kappa = D_{min}^2 / \{\sum_{(k=1,K)} \sigma_k^2\} \quad (8)$$

The variance of each cluster is calculated by summing over only the members of each cluster rather than over all Q for each cluster, which contrasts with the original Xie-Beni validity measure.

$$\sigma_k^2 = \sum_{\{q:\ q \text{ is in cluster } k\}} w_{qk} \| x^{(q)} - c^{(k)} \|^2 \quad (9)$$

E. Sensitivity:

The sensitivity of a test is its ability to determine the patient cases correctly. It should calculate the proportion of true positive in patient cases. This can be stated as

$$Sensitivity = \frac{No.of\ TruePositive}{TruePositive + TrueNegative}$$

F. Accuracy:

The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, it should calculate the proportion of true positive and true negative in all evaluated cases. This can be stated as

$$Accuracy = \frac{\sum Truepositive + \sum Truenegative}{\sum Totalpopulation}$$

A negative result in a test with high sensitivity is useful for ruling out disease. A high sensitivity test is reliable when its result is negative, since it rarely misdiagnoses those who have the disease. A test with 100% sensitivity will recognize all patients with the disease by testing positive. A negative test result would definitively rule out presence of the disease in a patient.

IV. Results and Discussion

Experimental analysis is indented to be of use to researchers from all fields who want to study algorithms experimentally. The Medical dataset obtained from the SPARCS is used to test the performance of the proposed algorithm. At the same time, its properties are also empirically studied. The experimental results are summarized and discussed in the following section.

A. Dataset Description:

Medical database named Statewide Planning and Research Cooperative System (SPARCS) includes about 10, 485, 76 medical data from various countries. SPARCS is a comprehensive all payer data reporting system established in 1979 as a result of cooperation between the health care industry and government. The data set contains 10, 48,576 records with 38 attributes. Among the available data randomly 20,000 records with 8 attributes has been selected and used for the proposed work. The chosen records belong to six countries namely Africa, India, Russia, America, Australia, and China with the attributes like name of the patient, age, country, etc. The sensitive information like patient's name, Address, Disease, Date of Visit, etc are preserved by applying AES encryption and decryption methods.

Dataset	No. Of Records	No. Of Attributes
Sparcs	20,000	8

Table 1: Dataset information

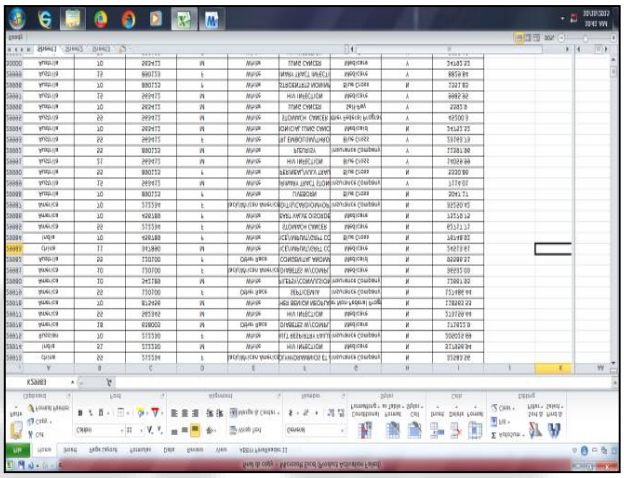

Fig. 4.1: SPARCS Medical Data set – Sample

B. Experimental Results:

Applied Fuzzy C-means Clustering algorithm for SPARCS medical dataset and evaluated the performance and found that Fuzzy C-means clustering yields better results.

The clusters are formed based on the attributes like country, age, and zip-code, gender etc. The Clustering using fuzzy C-means is shown in Figure 4.2.

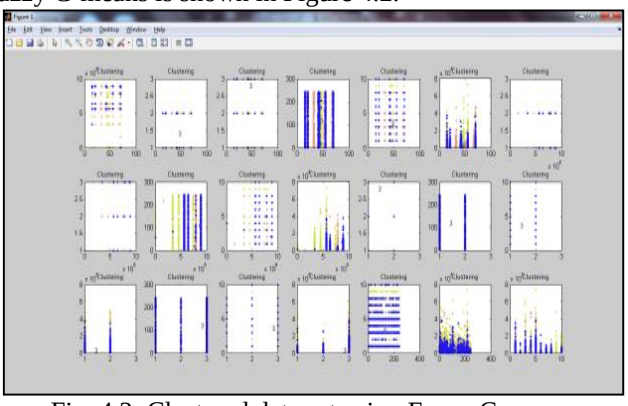

Fig. 4.2: Clustered data set using Fuzzy C-means

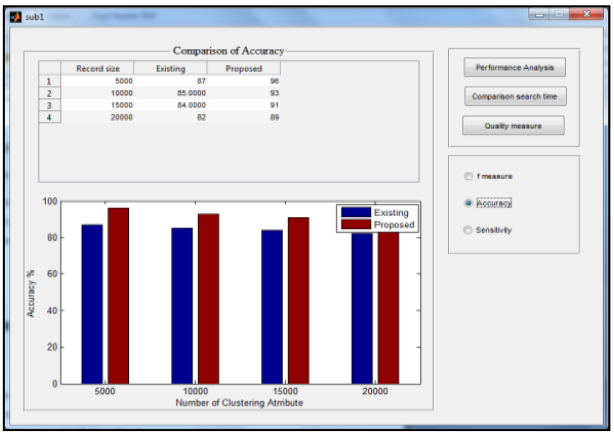

Fig. 4.3: Accuracy Comparison Existing vs. Proposed

Figure 4.3 shows the comparison of accuracy and it is found that the proposed method yields better accuracy than the existing method.

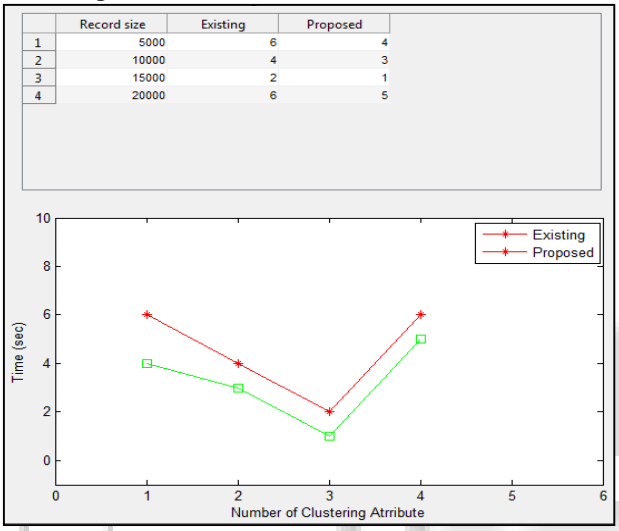

Fig. 4.4: Time Comparison Existing vs. Proposed

The proposed work reduces the time needed for processing and it is shown in the figure 4.4.

V. Conclusion

The fuzzy C-means Clustering algorithm is used for clustering the medical data set SPARCS based on various attributes by preserving the privacy of the data by AES encryption and decryption method. Privacy-preserving data mining (PPDM) considers the problem of running data mining algorithms on confidential data that is not supposed to be revealed even to the party running the algorithm. The experimental results indicate high performance in clustering the medical data set and centralized dataset using appropriate validation measures.

References

[1] Arun K. Pujari, "Data mining Techniques" university Press, First Edition 2001.
[2] Adam. R and Wortman. J.C, "Security-Control Methods for Statistical Databases: A Comparative Study",ACM Computing Surveys, vol.21, no.4, pp. 515-556, 1989.
[3] Bezdek. J.C, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenu Press, New York 1981.
[4] Bezdek. J.C, "Theory for Fuzzy Convergence C-means: Counter examples and Repairs", IEEE Trans. Syst., September/October 1987.
[5] Cliffton. C, Vaidya. J, "Privacy Preserving Naive Bayes Classifier for Vertically Partioned Data", In Proceedings of the 2004 SIAM International Conference on Data Mining, pp.522-526 , 2004.
[6] Dong Chen. C, "A Fast Secure dot product protocol with application to privacy preserving association rule mining", Pacific-Asia Conference on Knowledge Discovery.2014.
[7] Jiawei Han and Micheline Kamber, "Data Mining: Concept and techniques" Kaufman publishers, 2001.
[8] Kinoshenko. D, Mashtalir. V, and Yegorova. E, "Clustering method for fast content-based image retrieval" Computer Vision and Graphics,32,2006.
[9] Lindell. Y and Pinkas. B, Privacy Preserving Data Mining, Advances in Cryptology CRYPTO'00.Lecture Notes in Computer Science, Vol.1880, Springer-Verlag, 2000, pp.353.Earlier version of this paper.

