**I J C T A**, 9(13) 2016, pp. 6341-6347


# Clustering and Association using K-Mean over Well-Formed Protected Relational Data

**Aparna[*] and J. Prathipa[**]**

**ABSTRACT**

Clustering is a technique generally applied in almost different aspects of life. It is used for grouping the different set of objects in a group in a way that objects similar to each other in any sense are placed in same group and dissimilar items in different group. Association technique in data mining is used to find items that occur frequently for e.g. It is used in data mining for predicting and analyzing behavior of the customer. Its role is important for market basket analysis, making store layout etc. Association rules are made by analyzing if/then pattern that occur frequently. In project clustering and association is performed over well-formed encrypted data. The data is stored in encrypted format so that only authorized person can perform data mining over the data. The technique of both cryptography and data mining is used in my project. Cryptography is applied to ensure the data is safe from the outside world and is available only to the authenticated person.

***Keywords:*** Encryption, Clustering, Association, Cryptography

## I. INTRODUCTION

Data mining is the process of deriving hidden knowledge and useful patterns from large data sets. It includes several complications which include side information, clustering, association and classification. The tremendous amount of data present in the database is both relevant and irrelevant. In order to get the relevant information we need to know the relative information which is difficult to find. The overall goal of the data mining process is to derive important details from a data set and transform it to the understandable format to be used in future. Recently, many new ways of gathering data have resulted in a need for applications which work effectively and efficiently with data streams. Problems of data stream include classification, clustering and association. Data mining is primarily used today by organization which has a strong consumer focus like - retail, financial, communication, and marketing organizations. It enables the organization to get the relationship between the internal factors such as price, product positioning, staff skills etc. and the external factors such as economic indicators, competition etc. Also helps the organization to check the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to perform full in-depth operation in order to get into summary information for retrieving detail of the transactional data. Classification is a data mining technique which helps to predict the membership of data instances. For Example we can use the data stored to perform classification in order to predict the weather of the particular day as sunny, rainy or cloudy. Different classification techniques are K-Nearest Neighbor. Cluster analysis or clustering is the task of assembling a set of objects in such a way that objects in the same group/ cluster are more identical in any sense to each other than those objects belonging to other set. Different algorithms are implemented by researchers for different type of cluster. There are different cluster models like connectivity models, centroid models, distribution models, density models, subspace models etc.

[*] Student, M.Tech Computer Science and Engineering, Department of Computer Science, SRM University, Kattankulathur-603203, Chennai, India, *E-mail: aparnakamal21@gmail.com*

[**] Assistant Professor (O.G), Department of Computer Science and Engineering, SRM University, Kattankulathur-603203, Chennai, India, *E-mail: Prathipa.j@ktr.srmuniv.ac.in*

Clustering algorithms give performance on the basis of measurement and inadequacy. Therefore it is highly desirable to trim the measurement of feature space. There are two more often used modes to deal with this problem: feature extraction and feature selection. Clustering applications include Economic Science, Pattern Recognition, Spatial Data Analysis, Image Processing, document classification. K-Mean algorithm is used to perform clustering because K-Mean algorithm can be applied to larger data sets and further filtering of results can be done using X-Mean algorithms. It gives better efficient result of clustering.

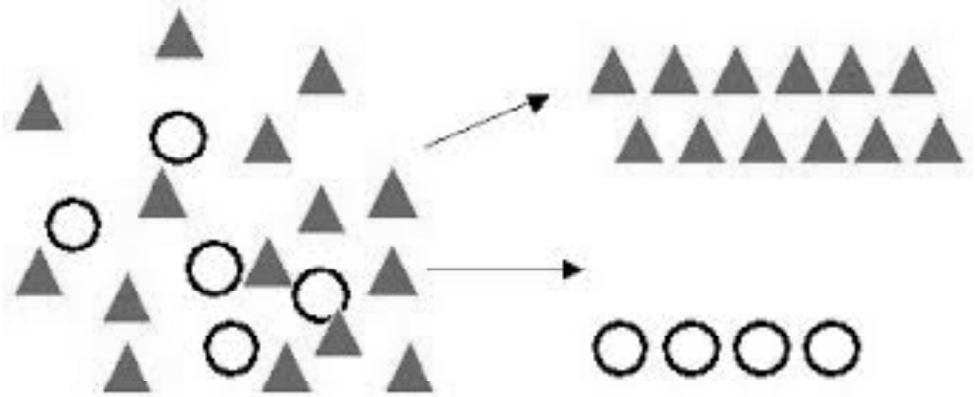

**Figure 1: Clustering**

Advancement in technology enables the organization to gather huge amount of data from their businesses. These datasets of these organizations are very valuable for the company because it helps to find the unknown knowledge by mining into the data sets. Mining association rules from the dataset is a serious topic for research these days. Consider the database of all the transaction of the retail industry. The goal of association rule mining is to find association rules from the existing data sets. In association rule mining, main computational step is to find the frequent pattern which is used to derive the association rule. The mining of association rules is one of the most popular problems of all these. The identification of set of items, products, behavior and characteristics which often occur together in the given database are the basic task of data mining. The problem of searching frequent patterns which exists in a same database is called Frequent Item Set Mining (FIM). Apriori algorithm can be implemented for searching frequent item sets. This algorithm at first identifies the frequent individual item in database and then extending them to larger item sets until the item set occur sufficiently often in database.

## Concept of association mining

**Item:** Transaction database field.

Transaction: It refers to a record of the database. Transaction is marked usually with small letter t and item is marked with small letter i. ti={i1,i2,…,ip}. TID is the unique identifier of each transaction. The database D consists of all the transactions. D={t1,t2,…,tn}

**Support:** The association rule is given as X->Y for the transaction database. This support is referred to as a ratio**.** The ratio is the ratios between total numbers of item set which contain X and Y and the count of the entire item set.

**Confidence:** It is defined as the ratio of total count of transaction having X and Y to the total count of X. It is marked as confidence(X->Y).

**Frequent Item set:** The item set which occur frequently and the support of which is always higher than the minimum support.

**Strong rule and Weak rule:** If support(X -> Y)>=Min Support and Confidence (X->Y)>=Min Confidence, then association rule X-> Y is marked as strong rule, otherwise it is marked as weak rule.

## II. RELATED WORK

Data mining is a technique which can be used to extract useful and important information. There are different data mining methods like association, classification and clustering. Tools of data mining are used to forecast the future trends and behaviors which permit the businesses to make dedicated, knowledge-driven decisions. Due to the growing economy demand for securing data is increasing in order to protect from unauthorized access. Cryptography is the art or science of secret writing and storing the information for shorter or longer period of time in certain form which is only revealed to the person who is authorized and hidden from others. Different application areas of data mining include banking, research, government agencies etc. Clustering and association is one of the commonly used tasks in data mining applications. For past decade many practical and theoretical solution have been proposed for the clustering and association problem. The recent popularity of cloud computing, users have opportunity to outsource their data, in encrypted form, as well as the data mining task can be performed on cloud. In this paper, we focus on solving the clustering and association problem over encrypted data. In particular, we propose a k-mean and x-mean algorithm over encrypted data in the cloud for the clustering and Apriori algorithm for association. Also, analysis of efûciency is done for our proposed protocol using a real-world dataset under different parameter settings. [1] Discussion is made about the problem of secure distributed classification. It assumes a privacy-preserving data mining scenario where collaboration of data sources is made to develop global model, but their data is not disclosed to others. These days enormous amount of data is present and different organizations use it as it is splitted among them. These data can be used by organizations to make predictive models which are accurate without revealing their own databases. These days we have enormous amount of data and in many situations, data is split between multiple organizations. Naïve Bayes classification was used for distributed data and it was considered as the baseline classifier. It was used to predict the class membership probability like checking the particular tuple for the class membership. Bayesian classification is based on Bayes' theorem and has high accuracy for larger databases. The main objective of privacy preserving data classification is to hide the private information and create accurate classifier. [2] Discussion was made about the problem of privacy preserving mining of association rules. Increase in the digital data concern is the preservation of privacy of user. Framework is presented for mining association rules that contains categorical items were randomized data is used for privacy preservation of the individual transaction. They analyze the nature of privacy breaks and propose a class of randomization operators that are much more effective than uniform randomization in limiting the breaks. Then the formulae is derived for an unbiased support estimator and its variance, which allow us to find item set supports from random datasets, and show how to utilize these formulae into mining algorithms. Finally, an experimental result is presented which validate the algorithm by applying it to the real data sets. [3] discussion was made about the Secure Nearest Neighbor (SNN) problem in which a query is issued by the client which contains query point E(q) to a cloud service provider and asks for them to give the encrypted data point in the encrypted database E(D) which is closest to the query point, without permitting the server to gain knowledge of the plaintexts of the data or the query and the result of the query. The relationship is established between the SNN problem and the order-preserving encryption (OPE) problem from the cryptography field and it is shown that the SNN is at least as hard as OPE. One cannot expect to find the exact encrypted nearest neighbor based on E(q) and E(D) because it is difficult to construct a secure OPE scheme in standard security models.

The popularity of cloud due to its flexibility and scalability motivates the service providers to provide access to the cloud databases like Amazon Relational Database Service (Amazon RDS), Microsoft SQL Azure. Data is stored over the cloud posted by the data owners for the storage management, and query

processing of databases. This framework provides great flexibility and scalability to data owners and their clients, and it is especially useful for users with strict local resources. However, security concerns are still there due to remote placement of the data , a data owner prefer the server not to gain access to the database D or query contents to D while still it require the server to provide the functionality of the database D in the cloud. For this purpose data owners encrypt the data with the encryption scheme E denoted as E(D) to be loaded to the cloud The clients also encrypt their queries q and send only the encrypted queries E(q) to the server. The server then identify the cipher text in E(D) which corresponds to the answer of q on D, using only E(q) and E(D).

## III. PROPOSED SYSTEM

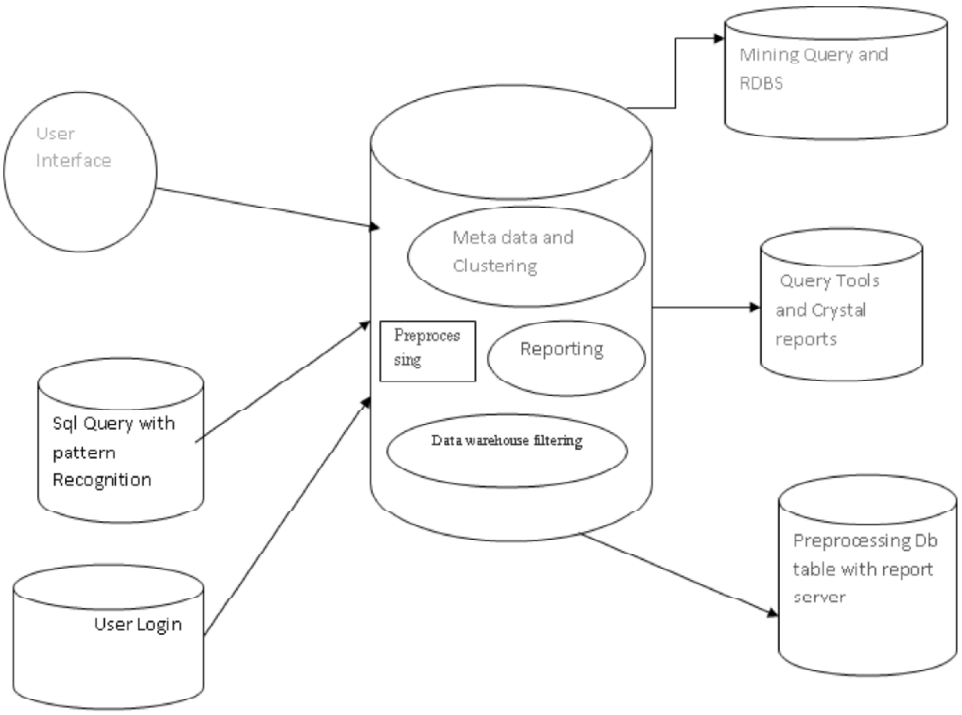


**Figure 2: Architecture diagram**

It proposes novel methods to effectively solve the DMED (Data Mining over Encrypted Data) problem assuming that the encrypted data are outsourced to a cloud. It includes the clustering and association which is considered as one of the most common data mining tasks. Because each technique has their own advantage and the cloud computing environment. Additionally, we are implementing other data mining techniques over cloud like clustering and association. User can group similar data items together in a secure environment. We apply k-mean algorithm for the task of clustering datasets and x-mean to get a filtered result. Apriori algorithm is applied for performing association. First task is to create the administrator who can perform all admin jobs like granting and revoking privilages from the user and also can perform task related to viewing details, deleting etc. New user can be added to the system by performing registration on the registration page. Our task is to get the employee details based on technology, qualification and experience so that we can make cluster according to our need also we can filter the cluster based on some other criteria. This filtering the result based on new criteria is the process called association where we associate already existing cluster to our requirement. For example we want to get the details of the employee who is working in some particular technology and have more than two years of experience then getting the user based on technology is clustering and then filtering it according to our need is association. The result can only be viewed by the

administrator of the system by posting query to the database. Then preprocessing of data is performed on the back end database and the result can be viewed in the form of crystal report. One employee can't check other employee details.

## IV. IMPLEMENTATION AND EXPECTED RESULTS

Data is stored over the cloud and the data is in encrypted form. This data access is only provided to the authorized user of the cloud. This makes the data secure and also querying over the cloud is faster. User and Administrator are present where a new user can be added by creating a new user by providing all the personal details. Details about the user are taken mainly for the technology, experience and qualification. Administrator is present who can view the details of the employee, can perform search operation and can delete the employee. Administrator can perform data mining activities to get the expected result. For Example- Administrator can search an employee suitable for the job based on qualification, technology and experience and can get the person name that is more appropriate for the job.

**X-MEAN ALGORITHM**

**Require :** Data set DS, Maximum number of clusters MAXC
**Require :** Function 2Means(DS) returning two clusters
Clustering ← 2Means(DS)
BestScore ← -∞
**While** |Clustering| <MAXC **do**
NewClustering = {}
**for all** Pl ∈ Clustering **do**
Pl2 ← 2Means(Pl)
**if** Measure({Pl}) > Measure(Pl2) **then**
NewClustering ← NewClustering ∪ {Pl}
else
NewClustering ← NewClustering ∪ Pl2
endif
end for
Clustering ← NewClustering
**if** Measure(Clustering) >BestScore **then**
BestScore ← Measure(Clustering)
BestClustering ← Clustering
endif
**end** while
**return** BestClustering

Clustering of entire data is done using 2-Means inorder to obtain two clusters. It efficiently searches the number of clusters. It makes local decision about which subset of the current centroid should split themselves in order to better fit the data. Splitting decision is done by computing BIC. X-Mean produces better clustering on real life data and synthetic data and runs faster. Meanwhile there are some disadvantages of x-mean that it can't be applied to medium or large datasets, it scales poorly and also search is prone to local minima.

**K-MEAN ALGORITHM**

Let $Y = \{y_1, y_2, y_3, \ldots\ldots, y_n\}$ be the set of data points and $P = \{p_1, p_2, \ldots\ldots, p_c\}$ be the set of centers.

1) Randomly select '*n*' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4) Recalculate the new cluster center using: where, '$n_i$' represents the number of data points in $i^{th}$ cluster.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3.

K-Mean algorithm is implemented for larger data sets, it is easier to understand and is a robust algorithm. It gives good performance when data set are distinct or separated from each other.

Distance between data point and cluster center is the Euclidean distance between them which can be calculated by the formula –

$$P_i = (1/n_i)\sum_{j=1}^{n_i} x_i, \text{ where } n_i \text{ represents the data pints in } {}_{\text{i}}^{\text{th}} \text{cluster}$$

**APRIORI ALGORITHM**

Apriori (TD, $\epsilon$)

$M_1 \leftarrow$ { large1 –Itemsets}

$l \leftarrow 2$

While $M_{l-1} \neq \phi$

$C_k \leftarrow \{d \cup \{e\} \mid d \in M_{k-1} \Lambda\ e \notin d\} - \{c \mid \{f \mid f \subseteq c\ \Lambda\ |f|=l-1\} \not\subseteq M_{l-1}$

for transactions $t \in TD$

$C_t \leftarrow \{c \mid c \in C_k \Lambda\ c \subseteq t\}$

for candicates $c \subseteq C_t$

count[c] $\leftarrow$ count [c]+1

$M_k \leftarrow \{ c \mid c \subseteq C_k\ \Lambda$ count[c] $\geq \epsilon$

$l \leftarrow l+1$

return $\cup M_l$

Apriori algorithm is used for implementing association of data sets in the cloud environment. This algorithm groups the data set based on frequently occurrence of the data. For e.g. Market basket analysis can be taken as a reference where the data is collected for customers buying behavior and a rule is derived out from the collected data about the items which is purchased the most and the item which is most likely to be purchased with the same product again and again.

## V. FUTURE WORK

The growing economy requires the data to be secure and these needs are fulfilled by the enviornment called cloud. Many cloud service provider are present which provide reliable, cost effective, resource sharing and

faster service to the client. The technique clustering and association can be used to retrieve the knowledge in an secure enviornment. K-mean and x-mean algorithm is used for clustering and Apriori algorithm is used for association. In future we can try performing the techniques clustering and association using other algorithms for these two tasks over the cloud computing enviornment.

## VI. CONCLUSION

Data mining technologies over the cloud computing is very important for the business to predict the future trends and make proactive decision. In the methodology "Clustering and Association using K-Mean over Well-Formed Protected Relational Data" we perform clustering and association in an authorized environment. Here the data is stored over the cloud in encrypted format.

## REFERENCES

[1] P. Zhang, Y. Tong, S. Tang, and D. Yang (2005), "Privacy preserving Naive Bayes classiûcation," *in Proc. 1st Int. Conf. Adv. Data Mining Appl.*

[2] A. Evûmievski, R. Srikant, R. Agrawal, and J. Gehrke (2004), "Privacy preserving mining of association rules" *Inf. Syst.*, Vol. 29, No. 4.

[3] X. Xiao, F. Li, and B. Yao (2013), Secure nearest neighbor revisited," *IEEE ICDE.*

[4] X. Xiao, F. Li, and B. Yao, "Secure nearest neighbor revisited," IEEE ICDE.

[5] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, "k-nearest neighbor classiûcation over semantically secure encrypted relational data," *eprint arXiv:1403.5001*.