International Journal of Engineering Technology and Applied Science
(ISSN: 2395 3853), Vol. 3 Issue 5 May 2017

A Survey on Privacy Preservation and Utility Mining in Data Mining

1 Vikram Rajpoot, 2 Vinod Jaiswal
[1]M-Tech Scholar, [2] Professor
Department of of CSE, LNCT, Bhopal, India

rajputvikram022@gmail.com, vinodjaiswal87@gmail.com

***Abstract*—Preservation of privacy in data mining has emerged as an absolute prerequisite for exchanging confidential information in terms of data analysis, validation, and publishing. Ever-escalating internet phishing posed severe threat on widespread propagation of sensitive information over the web. An emerging topic in the field of data mining is Utility Mining. The main objective of Utility Mining is to identify the itemsets with highest utilities, by considering profit, quantity, cost or other user preferences. In this paper we introduces Privacy preservation, utility mining and Privacy preservation data mining**

Keywords— Data mining; Privacy Preservation; Utility Mining; Itemsets.

I. INTRODUCTION

During the last ten years, Data mining, also known as knowledge discovery in databases has established its position as a prominent and important research area. The goal of data mining is to extract higher-level hidden information from an abundance of raw data. Data mining has been used in various data domains. Data mining can be regarded as an algorithmic process that takes data as input and yields patterns, such as classification rules, itemsets, association rules, or summaries, as output. Data Mining tasks can be classified into two categories, Descriptive Mining and Predictive Mining.The Descriptive Mining techniques such as Clustering, Association Rule Discovery, Sequential Pattern Discovery, is used to find human-interpretable patterns that describe the data. The Predictive Mining techniques like Classification, Regression, Deviation Detection, use some variables to predict unknown or future values of other variables [1].

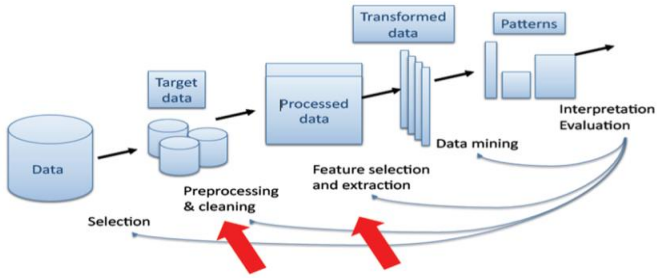

Fig. 1 Data Mining

II. PRIVACY PRESERVATION

Privacy preservation is becoming more and more a serious problem for future progress of data mining techniques with great potential access to datasets having private, sensitive, or confidential information. The major challenge for existing data mining algorithms is extracting accurate data mining results while still maintaining privacy of datasets. Due to the increasing concern on privacy, a new category of data mining called privacy preserving data mining (PPDM) has been introduced. But, the privacy-preserving data mining has turned into a major problem in recent years because of the huge amount of private data which is tracked by several business applications. In many situations, the users are reluctant to provide personal information unless the privacy of sensitive information is assured. PPDM was first introduced by Agrawal and Srikant in 2000. PPDM algorithms are developed by integrating privacy protection mechanism to conceal sensitive data before executing data mining algorithms. Then several different branches with different goals have been developed. Privacy preserving classification techniques prohibit a miner from building a classifier which is capable of forecasting the personal data.

The main consideration in privacy preserving data mining is the sensitive nature of raw data. The data miner, while mining for comprehensive statistical information about the data, should not be able to access data in its original form with all the sensitive information. This necessitates more robust techniques in privacy preserving data mining that intentionally alter the data to conceal sensitive information as well as protect the inherent statistics of the data which is vital for mining purpose. The latest trend in business collaboration is they are keen to share data or mined results to gain mutual benefit. But, it has also increased a potential threat of disclosing sensitive information when releasing the data. Data sanitization is the process, which hides the sensitive item sets present in the source database with proper modifications and discloses the modified database [2].

III. PRIVACY PRESERVING METHOD

Our basic approach to preserving privacy is to let users provide a modified value for sensitive attributes. The modified value may be generated using custom code, a browser plug-in, or extensions to products such as Microsoft's Passport (http://www.passport.com) or Novell's Digital Me

(http://www.digitalme.com). We consider two methods for modifying values:

Value-Class Membership In this method, the values for an attribute are partitioned into a set of disjoint, mutually-exclusive classes. We consider the special case of discretization in which values for an attribute are discretized into intervals. All intervals need not be of equal width. For example, salary may be discretized into 10K intervals for lower values and 50K intervals for higher values. Instead of a true attribute value, the user provides the interval in which the value lies. Discretization is the method used most often for hiding individual values.

Value Distortion: Return a value xi + r instead of xi where r is a random value drawn from some distribution. We consider two random distributions:

- Uniform: The random variable has a uniform distribution, between [-α; +α]. The mean of the random variable is 0.
- Gaussian: The random variable has a normal distribution, with mean $\mu = 0$ and standard deviation.

We fix the perturbation of an entity. Thus, it is not possible for snoopers to improve the estimates of the value of a field in a record by repeating queries.

A. Quantifying Privacy

For quantifying privacy provided by a method, we use a measure based on how closely the original values of a modified attribute can be estimated. If it can be estimated with c% confidence that a value x lies in the interval [x1; x2], then the interval width (x2 □ x1) defines the amount of privacy at c% confidence level. Table 1 shows the privacy offered by the different methods using this metric. We have assumed that the intervals are of equal width W in Discretization. Clearly, for 2 = W, Uniform and Discretization provide the same amount of privacy. As increases, privacy also increases. To keep up with Uniform, Discretization will have to increase the interval width, and hence reduce the number of intervals. Note that we are interested in very high privacy. (We use 25%, 50%, 100% and 200% of range of values of an attribute in our experiments.) Hence Discretization will lead to poor model accuracy compared to Uniform since all the values in a interval are modified to the same value. Gaussian provides significantly more privacy at higher confidence levels compared to the other two methods. We, therefore, focus on the two value distortion methods in the rest of the paper [3].

Table.1 Privacy Matrix

	Confidence		
	50%	**95%**	**99.9%**
Discretization	**0.5 W**	**0.95**	**0.99**
Uniform	**0.5 2**	**0.95**	**0.99**
Gaussian	**1.34**	**3.92**	**6.8**

IV. PRIVACY PRESERVING IN DATA MINING

Recently, the relevance of privacy-preserving data mining techniques is thoroughly analyzed and discussed. Utilization of specific methods revealed their ability to preventing the discriminatory use of data mining. Some methods suggested that any stigmatized group must not be targeted more on generalization of data than the general population. the technique called 'Privacy- Preserving Record Linkage' (PPRL), which allowed the linkage of databases to organizations by protecting the privacy. Thus, a PPRL methods based taxonomy is proposed to analyze them in 15 dimensions. Overviewed several available techniques of data mining for the privacy protection depending on data distribution, distortion, mining algorithms, and data or rules hiding. Regarding data distribution, only few algorithms are currently used for privacy protection data mining on centralized and distributed data. acknowledged the need to add or to multiply the protocol based homomorphic encryption along with the existing concept of digital envelope technique in obtaining collaborative data mining while keeping the private data intact among the mutual parties. The proposed technique exhibited considerable influence on different applications.

Analyzed the current privacy preserving solutions for cloud services, where the solution is outlined based on advanced cryptographic components. The solution offered the anonymous access, the unlink ability and the retention of confidentiality of transmitted data. Finally, this solution is implemented, the experimental results are obtained and the performance is compared. compared a set of fuzzy-based mapping methods in the context of privacy-preserving characteristics and the capability to maintain the same connection with other fields. This comparison is subjected to: (1) the four front modification of the fuzzy function definition, (2) the introduction of the seven ways to join different functional values of a particular data item to a single value, (3) the utilization of several similarity metrics for the comparison of the original data and mapped data, and (4) the evaluation of the influence of mapping on the derived association rule [4].

V. PRIVACY PRESERVING TECHNIQUE

We present here four efficient methods for privacy-preserving computations that can be used to support data mining. Not all are truly secure multiparty computations in some, information other than the results is revealed { but all do have provable bounds on the information released. In addition, they are efficient: the communication and computation cost is not significantly increased through addition of the privacy preserving component. This is by no means an exhaustive list of efficient secure multiparty computations. Some other examples can be found in.

A. Secure Sum

Secure sum is often given as a simple example of secure multiparty computation. We include it here because of its

applicability to data mining (see Sections 3.1 and 3.3), and because it demonstrates the difficulty and subtlety involved in making and proving a protocol secure. Distributed data mining algorithms frequently calculate the sum of values from individual sites. Assuming three or more parties and no collusion, the following method securely computes such a sum. Assume that the value v = Ps l=1 vl to be computed is known to lie in the range [0::n]. One site is designated the master site, numbered 1. The remaining sites are numbered 2::s. Site 1 generates a random number R, uniformly chosen from [0::n]. Site 1 adds this to its local value v1, and sends the sum R + v1 mod n to site 2. Since the value R is chosen uniformly from [1::n], the number R + v1 mod n is also distributed uniformly across this region, so site 2 learns nothing about the actual value of v1. For the remaining sites l = 2::s □ 1, the algorithm is as follows. Site l receives

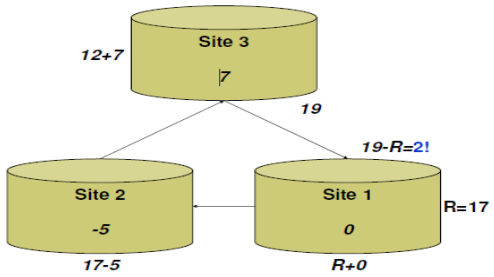

Fig. 2 secure computation ofa sum

$$V = R + \sum_{j=1}^{l-1} u_{j \bmod n}$$

Since this value is uniformly distributed across [1::n], i learns nothing. Site i then computes and passes it to site l + 1.

$$R + \sum_{j=1}^{l-1} u_{j \bmod n} = (u_j + V) \bmod n$$

Site s performs the above step, and sends the result to site 1. Site 1, knowing R, can subtract R to get the actual result. Note that site 1 can also determine by subtracting v1. This is possible from the global result regardless of how it is computed, so site 1 has not learned anything from the computation. Figure 1 depicts how this method operates. This method faces an obvious problem if sites collude. Sites l-1 and l + 1 can compare the values they send/receive to determine the exact value for vl. The method can be extended to work for an honest majority. Each site divides vl into shares. The sum for each share is computed individually. However, the path used is permuted for each share, such that no site has the same neighbor twice. To compute vl, the neighbors of l from each iteration would have to collude. Varying the number of shares varies the number of dishonest (colluding) parties required to violate security.

B. Secure Set Union

Secure union methods are useful in data mining where each party needs to give rules, frequent itemsets, etc., without revealing the owner. The union of items can be evaluated using SMC methods if the domain of the items is small.

Each party creates a binary vector where 1 in the ith entry represents that the party has the ith item. After this point, a simple circuit that or's the corresponding vectors can be built and it can be securely evaluated using general secure multi-party circuit evaluation protocols. However, in data mining the domain of the items is usually large. To overcome this problem a simple approach based on commutative encryption is used. An encryption algorithm is commutative if given encryption keys K1; : : : ;Kn 2 K, for any m in do- main M, and for any permutation i; j, the following two equations hold:

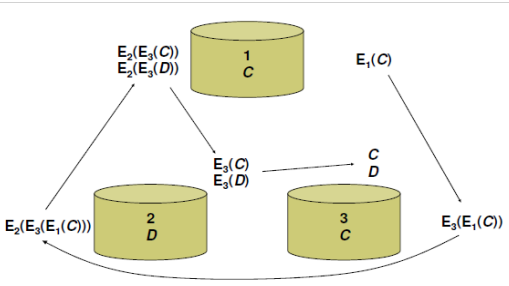

Fig. 3 determining the union of a set of items

$$E_{ki_1}(\ldots E_{ki_n}(M)\ldots) = E_{kj_1}(\ldots E_{kj_n}(M)\ldots) \quad (1)$$

$$\forall M_1 M_2 \in M \text{ such that } M_1 \neq M_2 \text{ and for given } k, \in < \frac{1}{2^k}$$
$$Pr\left(E_{ki_1}(\ldots E_{ki_n}(M1)\ldots) = E_{kj_1}(\ldots E_{kj_n}(M)\ldots)\right) < \in \quad (2)$$

With shared p the Pohlig-Hellman encryption scheme[15] satisfies the above equations, but any other commutative encryption scheme can be used.

C. Secure Size of Set Intersection

Consider several parties having their own sets of items from a common domain. The problem is to securely compute the cardinality/size of the intersection of these local sets. Formally, given k parties P1…Pk having local sets S1….. Sk, we wish to securely compute . We can do this is using a parametric commutative one way hash function. One way of getting such a hash function is to use commutative public key encryption, such as Pohlig Hellman, and

Algorithm 1 Finding secure union of items
Require: N is number of sites and Union_set = initially
{Encryption of all the rules by all sites }
for each site i do
for each X do
M = newarray[N] ;
Xp = encrypt(X; ei) ;

```
M[i] = 1 ;
Union set S (Xp,M);
end for
end for { Site i encrypts its items and adds them to the
global set. Each site then encrypts the items it has not
encrypted before }
for each site i do
for each tuple (r,M)    Union set do
if M[i] == 0 then
rp=encrypt(r,ei);
M[i]=1;
Mp= M ;
Union_set=(Union_set-{(r,M)})        p,Mp)};
end if
end for
end for
for (r,M)    Union set and (rp,Mp )    Union_set do
{check for duplicates}
if r==rp then
Union_set= Union_set-{(r,M)} {Eliminate duplicate
items before decrypting};
end if
end for
for each site i do {Each site decrypts every item to get
the union of items}
for all (r,M)    Union_set do
rd = decrypt(r,di) ;
Union_set=(Union set-{(r,M)})    {(rd)};
end for
permute elements in the Union_set
end for
return Union_set
```

Throw away the decryption keys. Commutative encryption has already been described in detail in Section 2.2. All k parties locally generate their public key-pair (Ei;Di) for a commutative encryption scheme. (They can throw away their decryption keys since these will never be used.) Each party encrypts its items with its key and passes it along to the other parties. On receiving a set of (encrypted) items, a party encrypts each item and permutes the order before sending it to the next party. This is repeated until every item has been encrypted by every party. Since encryption is commutative, the resulting values from two different sets will be equal if and only if the original values were the same (i.e., the item was present in both sets). Thus, we need only count the number of values that are present in all of the encrypted itemsets. This can be done by any party. None of the parties is able to know which of the items are present in the intersection set because of the encryption [5].

D. Scalar Product

Scalar product is a powerful component technique. Many data mining problems can essentially be reduced to computing the scalar product. One example of this, reducing association rule mining to scalar product computation, will be discussed in Section 3.2. The problem can be formally defined as follows: Assume 2 parties P1 and P2 each have a vector of cardinality n; i.e. P1 has ~X = (x1 : : : xn) and P2 has ~Y = (y1 : : : yn). The problem is to securely compute the scalar product of the two vectors, i.e., Pn i=1 xi _ yi. Recently, there has been a lot of research into this problem, which has given rise to many different solutions with varying degrees of accuracy, communication cost and security. Note that all of these techniques are limited to the 2-party version of the problem and cannot easily be extended to the general case. In the problem is modeled as Secure Multiparty Computation and the present a solution using cryptographic techniques (oblivious transfer). This, however, is not very efficient. The key insight in is to use linear combinations of random numbers to disguise vector elements and then do some computations to remove the effect of these randoms from the result. The solution is briery explained in algorithm 3. Though this method does reveal more information than just the input and the result, it is efficient and suited for large data sizes, thus being useful for data mining [5].

VI. UTILITY MINING

The limitations of frequent or rare itemset mining motivated researchers to conceive a utility based mining approach, which allows a user to conveniently express his or her perspectives concerning the usefulness of itemsets as utility values and then find itemsets with high utility values higher than a threshold .In utility based mining the term utility refers to the quantitative representation of user preference i.e. the utility value of an itemset is the measurement of the importance of that itemset in the users perspective. For e.g. if a sales analyst involved in some retail research needs to find out which itemsets in the stores earn the maximum sales revenue for the stores he or she will define the utility of any itemset as the monetary profit that the store earns by selling each unit of that itemset.

Here note that the sales analyst is not interested in the number of transactions that contain the itemset but he or she is only concerned about the revenue generated collectively by all the transactions containing the itemset. In practice the utility value of an itemset can be profit, popularity, page-rank, measure of some aesthetic aspect such as beauty or design or some other measures of user's preference.

Formally an itemset S is useful to a user if it satisfies a utility constraint i.e. any constraint in the form u(S) >= minutil, where u(S) is the utility value of the itemset an minutil is a utility threshold defined by the user [32]. In our example if we take utility of an itemset as the unit profit associated with the sale of that itemset then with utility threshold minutil = 500 then the itemset ABC has a utility value of 555 which means that this itemset is of interest to the user even though its support value is just 20%.Since while considering the total utility of an itemset S we multiply the utility values of the individual items consisting the itemset S with the corresponding frequencies of the individual items of S in the transactions that contain S, so the utility based mining approach can be said to be measuring the significance of an

itemset from two dimensions. The first dimension being the support value of the itemset i.e the frequency of the itemset and the second dimension is the semantic significance of the itemset as measured by the user. the importance of constraint based itemset mining in which the user has the privilege to specify his or her preferences by defining constraints that capture the semantic significance of the itemset in the intended application domain.

Defines two types of utility measures for any itemset, transaction utility and external utility. The Transaction utility of an item in a transaction is defined according to the information stored in the transaction. For e.g. the quantity of an item sold in the super market transaction database. The external utility of an itemset is based on the information provided by the user and is not available in the transactions. For e.g. in case of sales database the external utility may be the profit associated with the sale of itemsets [6].

VII. LITERATURE SURVEY

Jerry Chun-Wei Lin (2016) in this paper present that, two novel algorithms, namely Maximum Sensitive Utility-Maximum item Utility (MSU-MAU) and Maximum Sensitive Utility-Minimum item Utility (MSU-MIU),are respectively proposed to minimize the side effects of the sanitization process for hiding SHUIs. The proposed algorithms are designed to efficiently delete SHUIs or decrease their utilities using the concepts of maximum and minimum utility .A projection mechanism is also adopted in the two designed algorithms to speed up the sanitization process. Besides, since the evaluation criteria proposed for PPDM are insufficient and inappropriate for evaluating the sanitization performed by PPUM algorithms, this paper introduces three similarity measures to respectively assess the data base structure, data base utility and item utility of a sanitized database. These criteria are proposed as a new evaluation standard for PPUM [7].

Alpa Shah(2016) in this paper present that, Extensive research has been carried out for preserving the privacy of identifiers in dataset during Data Mining. Various dimensions based on Cryptographic principles, Perturbation and Secure Sum Computation have been studied to achieve privacy. Effective techniques to maximize privacy and minimize information loss have always been intriguing. The work in this paper presents a comparison based on experimental study of three fundamental perturbation techniques viz. - Additive, Multiplicative and Geometric Data Perturbation [GDP] for Privacy Preserving Data Mining [PPDM]. These techniques form the basis of many advanced Perturbation techniques as described later. The literature doesn't embark a clear cut comparison amongst the three techniques based on suitable metrics. We have identified various statistical metrics that must be considered for evaluating Perturbation techniques. The facet of research is independent in this context, and this paper will try to confer the applicability of perturbation techniques by descriptive statistics through experiments under one roof. A comparison amongst the perturbation based techniques is conferred at the end to exemplify the importance of this research [8].

Vadlana Baby (2016) in this paper present that, an efficient distributed threshold privacy-preserving kmeans clustering algorithm that use the code based threshold secret sharing as a privacy-preserving mechanism. Construction involves code based approach which allows the data to be divided into multiple shares and processed separately at different servers. Our protocol takes less number of iterations compare with existing protocols and it do not require any trust among the servers or users. We also provide experiment results with comparison and security analysis of the proposed scheme [9].

Prajakta R. Padhye(2016) in this paper present that, a system which uses HUPID-Tree structure to maintain the information about the database and patterns and it is updated only with the incremented data. It reduces the time overhead of rescanning the database from the beginning. High utility itemsets (HUIs) i.e. the desirable patterns mined from the HUPID-Tree will be used for generating rules. Cross selling profit of each rule will be estimated with the help of an objective function i.e. the rule utility function. Cross selling is the practice of selling among the established customers. It uses items in the consequent part of a rule for recommendation and provides future profit information with the application of a rule. Managers can use this cross-selling profit information to maximize the profit and the itemsets which will be sold in the future will also be the high utility itemsets [10].

Zakaria Gheid (2016) in this paper present that, propose a novel privacy-preserving k-means algorithm based on a simple yet secure and efficient multiparty additive scheme that is cryptography-free. We designed our solution for horizontally partitioned data. Moreover, we demonstrate that our scheme resists against adversaries passive model [11].

Junqiang Liu (2015) in this paper present that, A novel algorithm that finds high utility patterns in a single phase without generating candidates. The novelties lie in a high utility pattern growth approach, a look ahead strategy, and a linear data structure. Concretely, our pattern growth approach is to search a reverse set enumeration tree and to prune search space by utility upper bounding. We also look ahead to identify high utility patterns without enumeration by a closure property and a singleton property. Our linear data structure enables us to compute a tight bound for powerful pruning and to directly identify high utility patterns in an efficient and scalable way, which targets the root cause with prior algorithms. Extensive experiments on sparse and dense, synthetic and real world data suggest that our algorithm is up to 1 to 3 orders of magnitude more efficient and is more scalable than the state-of-the-art algorithms [12].

Majid Bashir Malik (2015) in this paper present that, A large number of tools and techniques have been developed for the purpose. Soft Computing methods have also emerged as a powerful tool for data mining as soft computing is tolerant to uncertainty, partial truth and imprecision. It helps in achieving solutions that are low cost, robust and tractable. Neural Networks are being extensively used for analysis purposes in every field of life from business to health sectors. In the current scenario where privacy of an individual is an important issue, people are reluctant to share their confidential information. Thereby privacy preserving in data mining (PPDM) has emerged as an indistinguishable component of data mining. The aim of this paper is to propose a model that preserves the privacy of individuals without affecting the final results of the Neural Networks [13].

Manish Shanna (2014) in this paper present that, Privacy preserving data mining techniques allow publishing data for the mining purpose while at the same time preserve the private information of the individuals. Many techniques have been proposed for privacy preservation but they suffer from various types of attacks and information loss. In this paper we proposed an efficient approach for privacy preservation in data mining. Our technique protects the sensitive data with less information loss which increase data usability and also prevent the sensitive data for various types of attack. Data can also be reconstructed using our proposed technique [14].

VIII. CONCLUSION

In many organizations large amount of data are collected. These data are sometimes used by the organizations for data mining tasks. However, the data collected may contain private or sensitive information which should be protected. Privacy protection is an important issue if we release data for the mining or sharing purpose. Privacy preserving data mining techniques allow publishing data for the mining purpose while at the same time preserve the private information of the individuals

References

[1] Jyothi Pillai, O.P.Vyas, "Overview of Itemset Utility Mining and its Applications", International Journal of Computer Applications (0975 – 8887) Volume 5– No.11, August 2010.

[2] C.SARAVANABHAVAN, R.M.S.PARVATHI, "PRIVACY PRESERVING SENSITIVE UTILITY PATTERN MINING", ISSN: 1992-8645/ 20th March 2013. Vol. 49 No.2 © 2005 - 2013 JATIT & LLS

[3] Rakesh Agrawal Ramakrishnan Srikant, "Privacy-Preserving Data Mining".

[4] Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh and Mohammad Abdur Razzaque, "A comprehensive review on privacy preserving data mining", Aldeen *et al. SpringerPlus (2015) 4:694* DOI 10.1186/s40064-015-1481-x

[5] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, Michael Y. Zhu, "Tools for Privacy Preserving Distributed Data Mining", Volume 4, Issue 2 - page 1

[6] Sudip Bhattacharya, Deepty Dubey, "High Utility Itemset Mining", ISSN 2250-2459, Volume 2, Issue 8, August 2012

[7] Jerry Chun-Wei Lin, Tsu-YangWu , PhilippeFournier-Viger, GuoLin , JustinZhan , MiroslavVoznak, "Fast algorithms for hiding sensitive high-utility itemsets in privacy-preserving utility mining", 55(2016)269–284 /Accepted 14 July 2016

[8] Alpa Shah, Ravi Gulati, "Evaluating Applicability Of Perturbation Techniques For Privacy Preserving Data Mining By Descriptive Statistics", 978-1-5090-2029-4/16/$31.00 @2016 IEEE

[9] Vadlana Baby, Dr. N. Subhash Chandra, "Distributed threshold k-means clustering for privacy preserving data mining", 978-1-5090-2029-4/16/$31.00 @2016 IEEE.

[10] Prajakta R. Padhye , R. J. Deshmukh, "A marketing solution for cross-selling by high utility itemset mining with dynamic transactional databases", 978-1-5090-0082-1/16/$31.00 ©2016 IEEE

[11] Zakaria Gheid, Yacine Challal, "Efficient and Privacy-Preserving k-means clustering For Big Data Mining", 2324-9013/16 $31.00 © 2016 IEEE.

[12] Junqiang Liu, Member, Benjamin C.M. Fung, "Mining High Utility Patterns in One Phase without Generating Candidates", 10.1109/TKDE.2015.2510012/1041-4347 _ 2015 IEEE.

[13] Majid Bashir Malik, M. Asger, Rashid Ali, Abid Sarvar, "A model for Privacy Preserving in Data Mining using Soft Computing Techniques", 978-9-3805-4416-8/15/$31.00 c 2015 IEEE.

[14] Manish Shannal ,Atul Chaudhar, Manish Mathuria,Shalini Chaudhar,Santosh Kumar, "An Efficient Approach for Privacy Preserving in Data Mining", 978-1-4799-3140-8/14/$31.00 ©2014 IEEE.