ISSN 2321-807X
Volume 12 Number12
Journal of Advances in chemistry

A Fuzzy Logic based Privacy Preservation Clustering method for achieving K-Anonymity using EMD in dLink Model

Dr.D.Palanikkumar, Ms.S.Priya
Professor, Department of Computer Science and Engineering, Nehru Institute of Technology, Coimbatore
palanikkumard@gmail.com
Assistant Professor,Dept of CSE&IT,Coimbatore Institute of Technology, Coimbatore
priyapalanikkumard@gmail.com

ABSTRACT

Privacy preservation is the data mining technique which is to be applied on the databases without violating the privacy of individuals. The sensitive attribute can be selected from the numerical data and it can be modified by any data modification technique. After modification, the modified data can be released to any agency. If they can apply data mining techniques such as clustering, classification etc for data analysis, the modified data does not affect the result. In privacy preservation technique, the sensitive data is converted into modified data using S-shaped fuzzy membership function. K-means clustering is applied for both original and modified data to get the clusters. t-closeness requires that the distribution of sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. Earth Mover Distance (EMD) is used to measure the distance between the two distributions should be no more than a threshold t. Hence privacy is preserved and accuracy of the data is maintained.

Indexing terms/Keywords

S-shaped fuzzy function, t-closeness, Earth Mover Distance

Academic Discipline And Sub-Disciplines

Computer Science – Information and Communication Engineering

SUBJECT CLASSIFICATION

Computer Science – Data Modelling – Algorithmic Approach

TYPE (METHOD/APPROACH)

Classification and Clustering method, k - anonymity.

1. INTRODUCTION

1.1 Overview of data mining

Data mining is a recently emerging field, connecting the three worlds of Databases, Artificial Intelligence and Statistics (Sairam et al.2011). The information age has enabled many organizations to gather large volumes of data (Y.Li et al.2009). However, the usefulness of this data is negligible if “meaningful information” or “knowledge” cannot be extracted from it. Data mining, otherwise known as knowledge discovery, attempts to answer this need (Agarwal et al.2002). In contrast to standard statistical methods, data mining techniques search for interesting information without demanding a priori hypotheses. As a field, it has introduced new concepts and algorithms such as association rule learning (Fienberg et al.2005). It has also applied known machine-learning algorithms such as inductive-rule learning (e.g., by decision trees) to the setting where very large databases are involved. Data mining techniques are used in business and research and are becoming more and more popular with time.

1.1.1 Confidentiality issues in data mining

A key problem that arises in any en masse collection of data is that of confidentiality. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests (Muralidhar et al.2006).However, there are situations where the sharing of data can lead to mutual gain. A key utility of large databases today is research, whether it be scientific, or economic and market oriented. Thus, for example, the medical field has much to gain by pooling data for research; as can even competing businesses with mutual interests. Despite the potential gain, this is often not possible due to the confidentiality issues which arise.

1.2 Applications of Privacy-Preserving Data Mining

The problem of privacy-preserving data mining has numerous applications in homeland security, medical database mining, and customer transaction analysis. Some of these applications such as those involving

bio-terrorism and medical database mining may intersect in scope (Agarwal et al.2002). In this section, we will discuss a number of different applications of privacy-preserving data mining methods.

a. Medical Databases
b. Bioterrorism Applications
c. Homeland Security Applications

1.3 Data Mining Methods

The main reason for applying data mining methods to text document collections is to structure them. A structure can significantly simplify the access to a document collection for a user. Well known access structures are library catalogues or book indexes (Muralidhar et al.2006). However, the problem of manual designed indexes is the time required to maintain them. Therefore, they are very often not up-to-date and thus not usable for recent publications or frequently changing information sources like the World Wide Web. The existing methods for structuring collections either try to assign keywords to documents based on a given keyword set (classification or categorization methods) or automatically structure document collections to find groups of similar documents (clustering methods). In the following we first describe both of these approaches. The proposed methods to automatically extract useful information patterns from text document collections.

Clustering

Clustering method can be used in order to find groups of documents with similar content. The result of clustering is typically a partition (also called) clustering P, a set of clusters P. Each cluster consists of a number of documents d. Objects in the case documents of a cluster should be similar and dissimilar to documents of other clusters. Usually the quality of clustering's considered better if the contents of the documents within one cluster are more similar and between the clusters more dissimilar (R.K.Ahuja et al.1993). Clustering methods group the documents only by considering their distribution in document space (for example, a n-dimensional space if we use the vector space model for text documents).

Clustering algorithms compute the clusters based on the attributes of the data and measures of (dis)similarity. However, the idea of what an ideal clustering result should look like varies between applications and might be even different between users. One can exert influence on the results of a clustering algorithm by using only subsets of attributes or by adapting the used similarity measures and thus control the clustering process. To which extent the result of the cluster algorithm coincides with the ideas of the user can be assessed by evaluation measures.

Anonymization Approach

Data anonymization is a type of information sanitization whose intent is to ensure privacy protection. It is the process of either encrypting or removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous.

K-anonymity

A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appear in the release.

L-diversity

This model is an extension of the K – Anonymity. It is a form of group based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation. An equivalence class is said to have l-diversity if there are at least l "well-represented" values for the sensitive attribute. Distinct l-diversity, Entropy l-diversity, Recursive (c-l) - diversity are the different types of L-diversity models.

t-closeness

It is a further refinement of l-diversity group based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation. An equivalence class is said to have *t*-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold *t.* t-closeness anonymization is more effective than many other privacy-preserving data mining methods.

2. RELATED WORKS

In recent year's lot of research work has been carried out to preserve data privacy before releasing the data for various research purposes which adopts various techniques like Data Auditing, Data Modification, Cryptographic methods and k-anonymity (Ren et al.2012).

In Modification-Based Techniques a number of techniques have been developed for a quantity of data mining techniques like classification, association rule discovery and clustering (Weng et al.2015). Based on the hypothesis that discerning data modification or sanitization is an NP-Hard problem, and for this basis, alteration can be used to address the complexity issues like swapping values between records, replacing the original

database by a sample from the same distribution, adding noise to the values in the database, adding noise to the results of query, sampling the results of a query (Z.Qin et al.2013).

In Cryptographic methods, data is encrypted using protocols like secured multiparty computation (SMC) (Bayardo et al.2005)..It is a study of mathematical techniques, related to aspects of information security such as confidentiality, data integrity, entity authentication and data origin authentication is shaping the way that information is safely and securely transmitted over the internet. Sensitive information is quite large, such as Credit card information, Security numbers, Private correspondence, Military statement, Bank account information.

Refurbishing-based techniques are techniques where the original circulation of the data is reconstructed from the randomized data.

3. PRIVACY PROTECTED DATA PUBLISHING TECHNIQUES

In this section, this analyzes how rule based slicing can provide membership disclosure protection.

Bucketization

This block examines how an adversary can infer membership information from bucketization. Because bucketization releases each tuple's combination of QI values in their original form and most individuals can be uniquely identified using the QI values, the adversary can determine thhjje membership of an individual in the original data by examining whether the individual's combination of QI values occurs in the released data (Duncan et al.2001).

Rule Based Slicing

Slicing offers protection against membership disclosure because QI attributes are partitioned into different columns and correlations among different columns within each bucket are broken (Lambert et al.1986).

The proposed two quantitative measures for the degree of membership protection offered by rule based slicing which identifies the background knowledge about the data.

The first is the fake-original ratio (FOR), which is defined as the number of fake tuples divided by the number of original tuples. Intuitively, the larger the FOR, the more membership protection is provided.

Generalization

By generalizing attribute values into "less-specific but semantically consistent values," generalization offers some protection against membership disclosure.

It was shown in that generalization alone (e.g., used with k-anonymity) may leak membership information if the target individual is the only possible match for a generalized record (Givens et al,1984). The intuition is similar to our rationale of fake tuple. If a generalized tuple does not introduce fake tuples (i.e., none of the other combinations of values are reasonable), there will be only one original tuple that matches with the generalized tuple and the membership information can still be inferred

Also, the protection against membership disclosure depends on the choice of the background table. Therefore, with careful anonymization, generalization can offer some level of membership disclosure protection.

4. BASIC ALGORITHMS

4.1 K-anonymity

K-anonymity is a popular measure of privacy for data publishing: It measures the risk of identity-disclosure of individuals whose personal information is released in the form of published data for statistical analysis and data mining purposes (e.g. census data) (Iyengar et al.2002). Higher values of k denote higher level of privacy (smaller risk of disclosure).

In many applications, the data records are made available by simply removing key identifiers such as the name and social-security numbers from personal records. However, other kinds of attributes (known as pseudo-identifiers) can be used in order to accurately identify the records. For example, attributes such as age, zip-code and sex are available in public records such as census rolls. When these attributes are also available in a given data set, they can be used to infer the identity of the corresponding individual. A combination of these attributes can be very powerful, since they can be used to narrow down the possibilities to a small number of individuals.

In k-anonymity techniques, we reduce the granularity of representation of these pseudo-identifiers with the use of techniques such as generalization and suppression(Koudas et al.2007).In the method of

generalization, the attribute values are generalized to a range in order to reduce the granularity of representation. For example, the date of birth could be generalized to a range such as year of birth, so as to reduce the risk of identification. In the method of suppression, the value of the attribute is removed completely (Dwork et al.2011). It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data.In order to reduce the risk of identification, the k-anonymity approach requires that every tuple in the table be Indistinguish ability related to no fewer than k respondents. This can be formalized as follows:

Anonymizing Data: k-anonymity

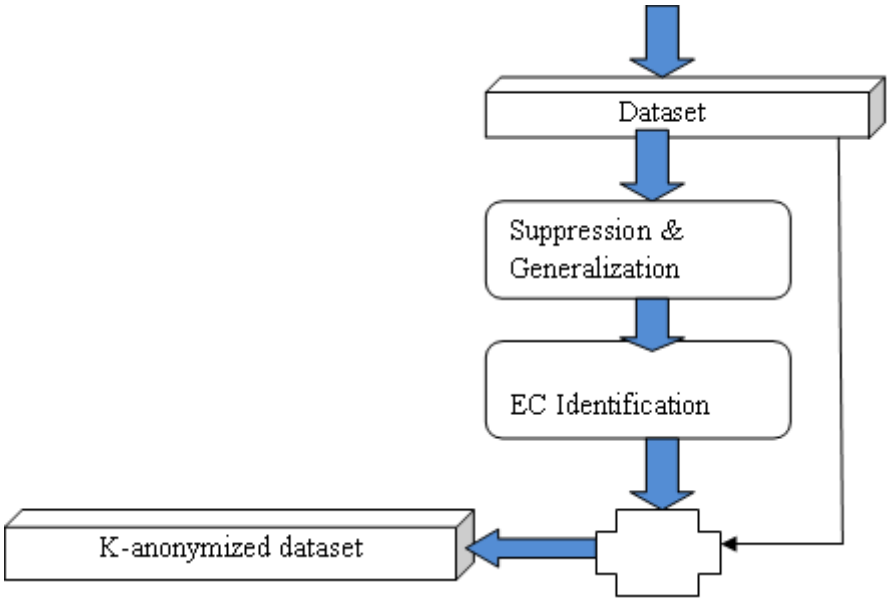

Figure 1: Data Anonymizing Process Flow

There are four basic methods for anonymizing data:

- Replacement - substitute identifying numbers
- Suppression - omit from the released data
- Generalization - for example, replace birth date with something less specific, like year of birth
- Perturbation - make random changes to the data

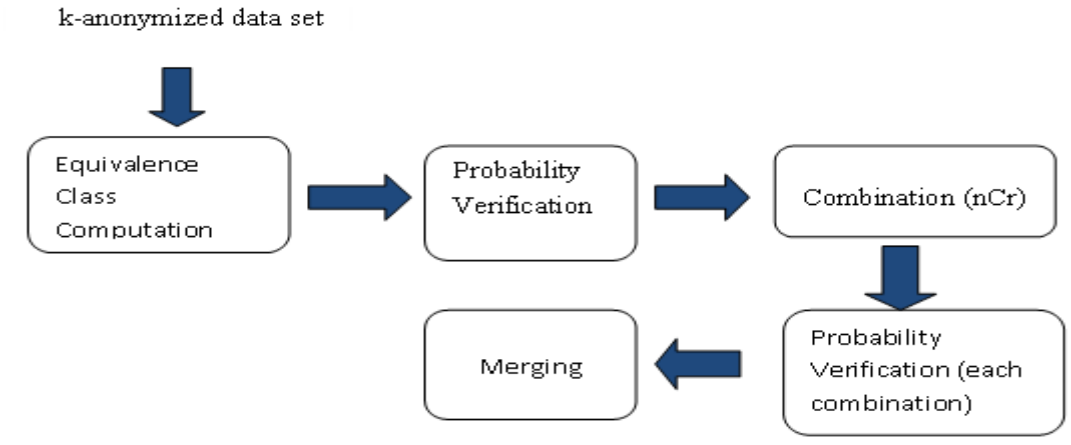

Figure 2: K anonymity Process Flow

Table 1: Sample Dataset

ISSN 2321-807X
Volume 12 Number12
Journal of Advances in chemistry

Name	Race	Birth	Gender	Zip	Problem
Sean	Black	12-3-1993	M	02141	Short Breath
Daniel	Black	13-4-1991	M	02141	Chest Pain
Kate	Black	15-5-1989	F	02138	Hypertension
Marion	Black	17-6-1987	F	02138	Hypertension
Helen	Black	18-8-1995	F	02138	Obesity
Reese	Black	12-6-1997	F	02138	Chest Pain
Forest	White	13-7-1988	M	02138	Chest Pain

Take a table 1 for example, with rows and attributes. Each attributes is either part of a quasi-identifier (like a name or address), or is sensitive information (like the fact you had an operation on a particular afternoon). A quasi-identifier is a set of attributes that, perhaps in combination, can uniquely identify individuals (Kullback et al.2014). Sensitive information includes the attributes that we want to keep private (Lambert,1993). The driving license number is an identifier; a driving record is sensitive information. The table satisfies k-anonymity if each sequence of values in any quasi-identifier appears with at least k occurrences. Bigger k is better. If the user removes all the attributes except for the problem we have a much anonymized data set (k=11)

Table 2: Anonymized dataset

Name	Race	Birth	Gender	Zip	Problem
Sean	Black	12-3-1993	M	02141	Short Breath
Daniel	Black	13-4-1991	M	02141	Chest Pain
Kate	Black	15-5-1989	F	02138	Hypertension
Marion	Black	17-6-1987	F	02138	Hypertension
Helen	Black	18-8-1995	F	02138	Obesity
Reese	Black	12-6-1997	F	02138	Chest Pain
Forest	White	13-7-1988	M	02138	Chest Pain

On the other hand, if user just removes the name and generalize the zip code and date of birth we have a less anonymized set. Exercise: convince yourself that k=2 for this set.

Table 3: Another Kind of anonymized dataset

Name	Race	Birth	Gender	Zip	Problem
Sean	Black	1965	M	0214*	Short Breath
Daniel	Black	1965	M	0214*	Chest Pain
Kate	Black	1965	F	0213*	Hypertension
Marion	Black	1965	F	0213*	Hypertension
Helen	Black	1964	F	0213*	Obesity
Reese	Black	1964	F	0213*	Chest Pain

Forest	White	1964	M	0213*	Chest Pain

Of course, the issue is utility. There is a tradeoff between keeping the data useful for research and maintaining privacy. Researchers and attackers are doing the same thing after all: looking for useful patterns in the data (Xiaokui et al.2015). With the k=2 data set you can ask questions about correlation of problems with gender, or with geography to some extent (although not very specific geographical factors, like toxic leaks).

4.2 The l-diversity Method

The k-anonymity is an attractive technique because of the simplicity of the definition and the numerous algorithms available to perform the anonymization. Nevertheless the technique is susceptible to many kinds of attacks especially when background knowledge is available to the attacker. Some kinds of such attacks are as follows:

Homogeneity Attack: In this attack, all the values for a sensitive attribute within a group of k records are the same. Therefore, even though the data is k-anonymized, the value of the sensitive attribute for that group of k records can be predicted exactly.

Background Knowledge Attack: In this attack, the adversary can use an association between one or more quasi-identifier attributes with the sensitive attribute in order to narrow down possible values of the sensitive field further (Xiaokui et al.2015). An example given in the following is one in which background knowledge of low incidence of heart attacks among Japanese could be used to narrow down information for the sensitive field of what disease a patient might have. A detailed discussion of the effects of background knowledge on privacy may be found the existing approaches. Clearly, while K-anonymity is effective in preventing identification of a record, it may not always be ef fective in preventing inference of the sensitive values of the attributes of that record. Therefore, the technique of l-diversity was proposed which not only maintains the minimum group size of k, but also focuses on maintaining the diversity of the sensitive attributes.

4.3 D-Link

Organizations share data about individuals to drive business and comply with law and regulation. However, an adversary may expose confidential information using quasi-identifying attributes (e.g., age, geocode and gender) across disparate data publications (Xiaokui et al.2015). Privacy protection models (e.g., k-anonymity and its extensions) fail to protect an individual's privacy against this ''composition attack". The objective is to enhance the dLink model by providing privacy preservation using t- closeness for publish data set. It includes Generalization and Suppression.

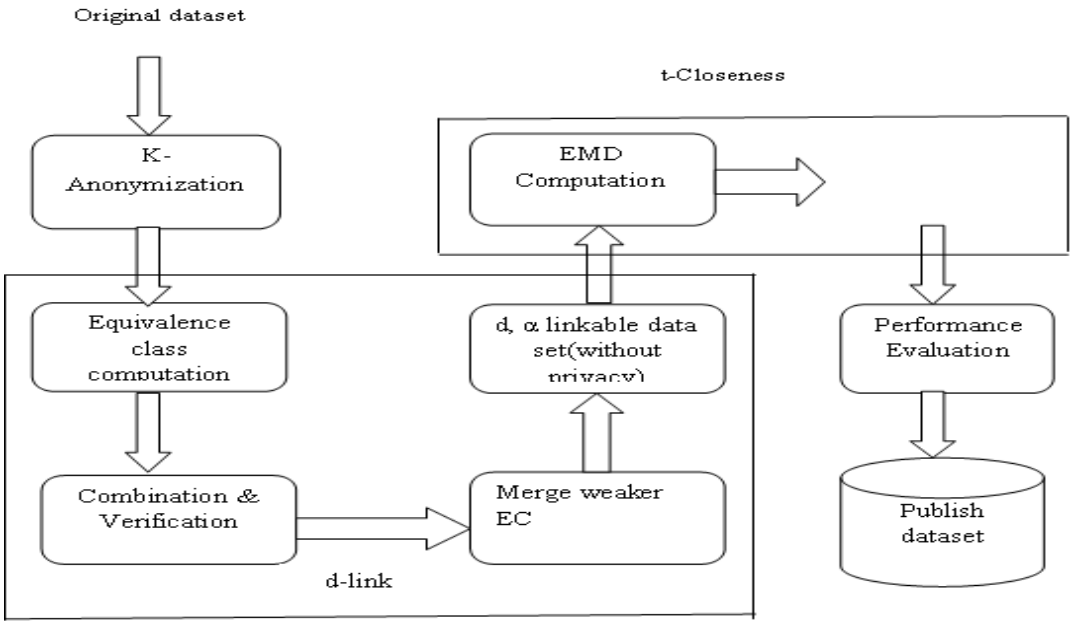

Figure 3: Privacy Preservation using t- Closeness

4.4 EMD computation

The Earth Mover's Distance (EMD) is a method to evaluate dissimilarity between two multi-dimensional distributions in some feature space where a distance measure between single features, which we call the ground distance is given. The EMD ``lifts" this distance from individual features to full distributions (Liu et al.2015).Intuitively, given two distributions, one can be seen as a mass of earth properly spread in space, the other as a collection of holes in that same space. Then, the EMD measures the least amount of work needed to fill the space(Geravand et al.2013).

A distribution can be represented by a set of clusters where each cluster is represented by its mean (or mode), and by the fraction of the distribution that belongs to that cluster(Shu et al.2012). The step by step procedure for calculating Earth Mover Distance (EMD) is given below,

Steps:

1. Order the Numerical attribute: Numerical attribute values are ordered. Let the attribute domain be {*v*1*, v*2*...*vm}, where *vi* is the ith smallest value.

   3 4 5 6 7 8 9 10 11

2. Split the dataset based on equal class.

   3 4 5 6 7 8 9 10 11

3. Find correlation between quasi-identifier attributes and sensitive attributes.
4. Find probabilities for each sensitive attribute
5. Find distance between probabilities
6. For every class find the score by applying the following formula.

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} .$$

The following section discusses about the implementation details, experimental bed set-up for further evaluation.

5. EVALUATION METHOD

The Earth Mover"s Distance (EMD) was introduced in laptop vision as associate degree improved distance live between 2 distributions. The Most frequent use of EMD is often recorded in multimedia database systems (Duncan et al. 2001). The EMD is predicated on the stripped-down quantity of work required to rework one distribution to a different by moving distribution mass between one another. Using results from running common machine learning algorithms (such as k-means clustering and logistic regression on a dataset) that EMD does not significantly affect the accuracy of data analysis (Swapnil et al. 2016). Further, we show that the method not only relieves the analysts from the burden of distributing a privacy budget between data transformation operations, it also manages to provide superior output accuracy. Evaluation criteria for privacy & utility are the most thematic consideration and it is shown through benchmarks of minimizing the composition attack using the t-closeness with dLink model.

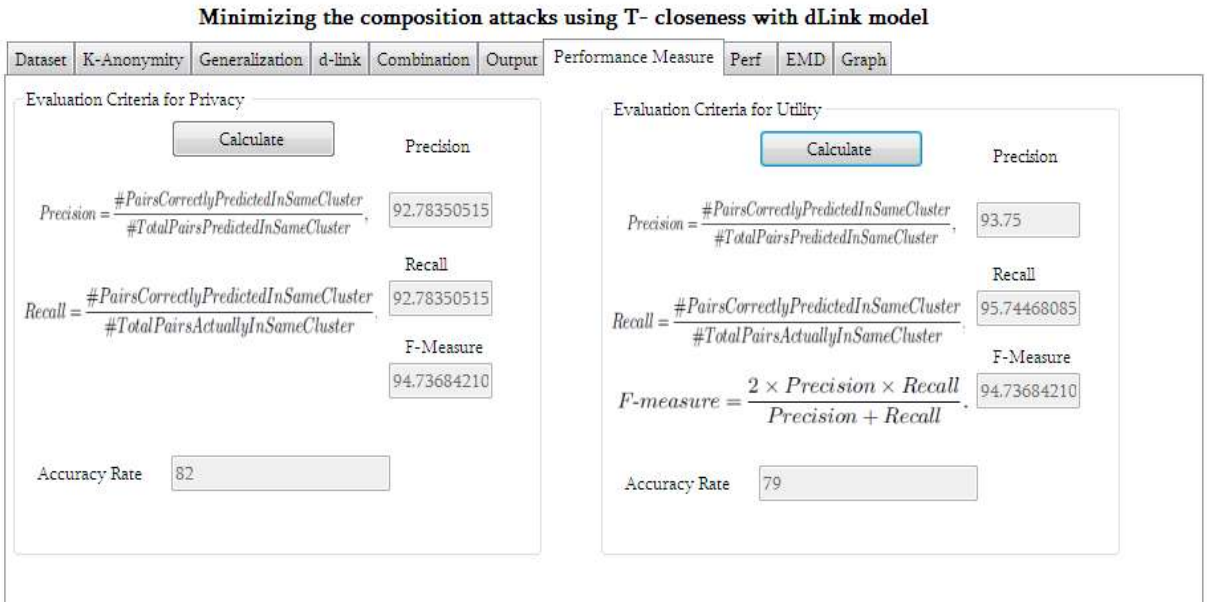

Figure 4: Evaluation criteria for privacy and utility

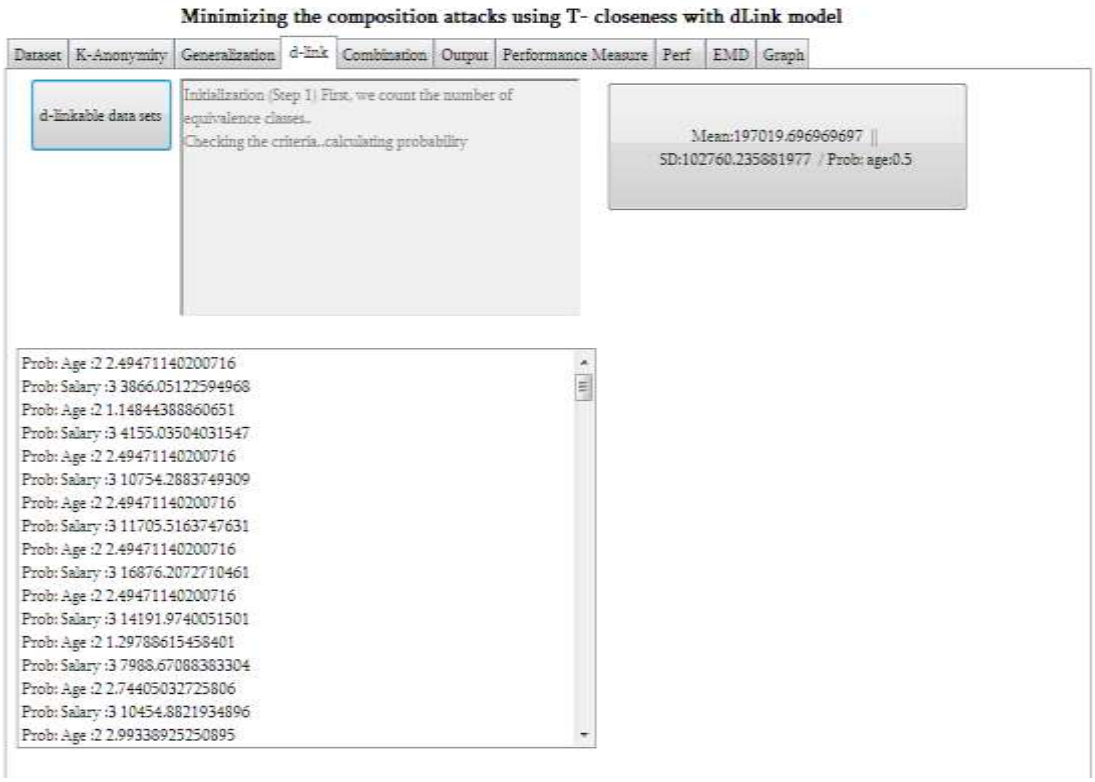

Figure 5: dLink- Partitioning Equivalence Classes

ISSN 2321-807X
Volume 12 Number12
Journal of Advances in chemistry

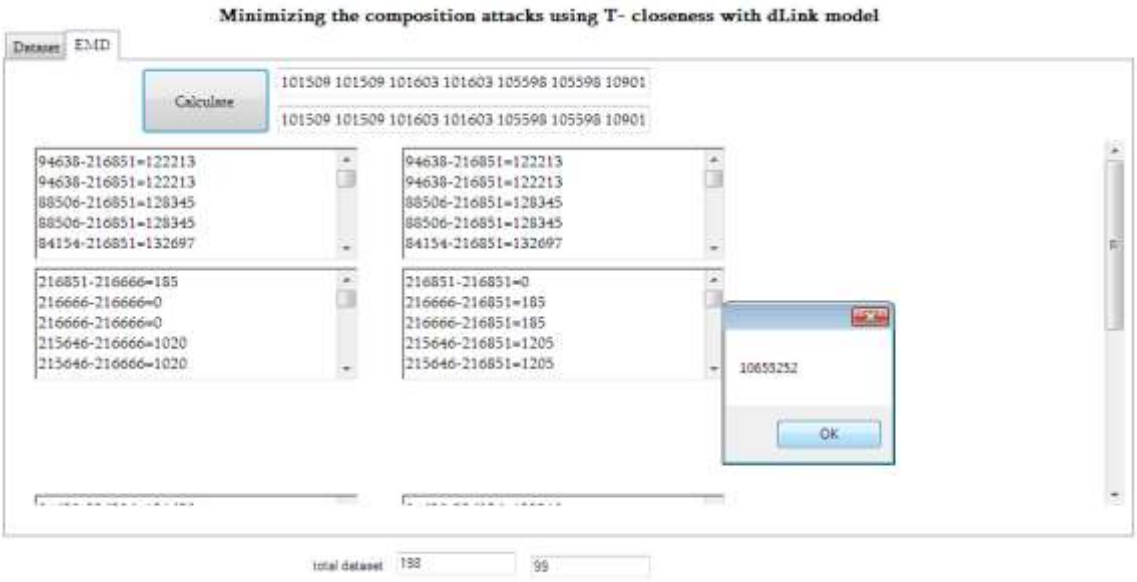

Figure 6: Probability Calculation

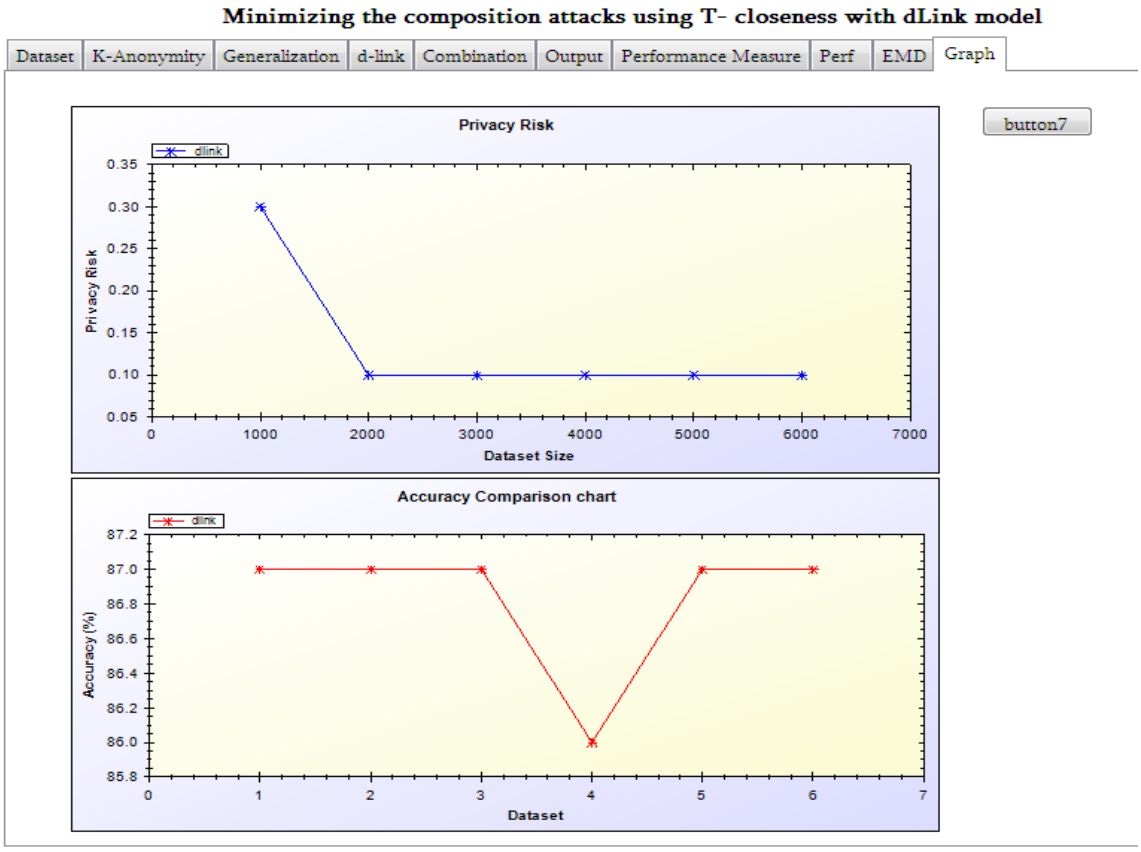

Figure 7: Minimized composition attack - EMD using T-Closeness with dLink Model

6. CONCLUSION

While *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. The notion of _-diversity attempts to solve this problem by requiring that each equivalence class has at least _ well-represented values for each sensitive attribute. We have shown that _-diversity has a number of limitations and have proposed a novel privacy notion called *t*-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold *t*). As part of future work, Data Perturbation will help to preserve data and hence sensitivity is maintained .In future, we want to propose a hybrid approach of these techniques (Valake et al, 2014)

7. REFERENCES

1. Agrawal D, Aggarwal C.C, "On the Design and Quantification of Privacy Preserving Data mining algorithms", ACM PODS Conference,2002
2. Agrawal D, Aggarwal C.C, "On the Design and Quantification of Privacy Preserving Data mining algorithms", ACM PODS Conference,2002
3. Dwork, "A firm foundation for private data analysis", Commun. ACM Janvol, vol. 54, no. 1, pp. 86-95, 2011.
4. C. R. Givens and R. M. Shortt. A class of Wasserstein met-rics for probability distributions. Michigan Math J., 31:231–240, 1984.
5. Lambert. Measures of disclosure risk and harm. J. Offi-cial Stat., 9:313, 1993.
6. Liu, X. Shu, D. Yao, and A. R. Butt, "Privacy-preserving scanning of big content for sensitive data exposure with MapReduce," in Proc.ACM CODASPY, 2015.
7. Fienberg S.E. and Mcintyre J. "Data Swapping:Variations on a theme by Dalenius and Reiss." In Journal of Official Statistics,21:309-323,2005.
8. K. Ren, C. Wang and Q. Wang, "Security challenges for the public cloud", *IEEE Internet Computing*, no. 1, pp. 69-73, 2012.
9. L. Weng, L. Amsaleg, A. Morton and S. Marchand-Maillet, "A privacy-preserving framework for large-scale content-based information retrieval", *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 1, pp. 152-167, 2015.
10. T. Duncan and D. Lambert, "Disclosure-limited data dissemination", Journal of the American statistical association, vol. 81, no. 393, (1986), pp. 10-18.
11. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman, and S. F. Roehrig. Disclosure limitation methods and informa- tion loss for tabular data.pages 135–166. Elsevier, 2001.
12. Muralidhar K. and Sarathy R. " Data Shuffling a new masking approach for numerical data", Management Science, forthcom- ing, 2006.
13. N. Koudas, D. Srivastava, T. Yu, and Q. Zhang. Aggregate query answering on anonymized tables. In Proc. 23rd Intnl. Conf. Data Engg. ICDE, 2007.
14. P. Valake Tejashri and S. Patil Sachin A Enabling Multilevel Trust Privacy Preserving Data Mining using Random Rotation Based Data Perturbation © Elsevier Publications ERCICA-2014.
15. R. J. Bayardo and R. Agrawal. Data privacy through opti- mal k-anonymization. In Proc. 21st Intnl. Conf. Data Engg. (ICDE), pages 217–228, Washington, DC, USA, 2005. IEEE Computer Society.
16. R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. Network flows: theory, algorithms, and applications. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.
17. Sairam et al "Performance Analysis of Clustering Algorithms in Detecting outliers",International Journal of Computer Science a nd Information Technologies, Vol. 2 (I) , Jan-Feb 2011, 486-488.
18. S. Geravand and M. Ahmadi, "Bloom filter applications in network security: A state-of-the-art survey," Comput. Netw., vol. 57, no. 18,pp. 4047–4064, Dec. 2013.
19. S. L. Kullback and R. A. Leibler. On information and sufficiency. Ann. Math. Stat.22:79–86, 2014
20. Swapnil Kadam, Prof. Navnath Pokale, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, January 2016
21. V. S. Iyengar. Transforming data to satisfy privacy con-straints. In Proc. 8th ACM KDD, pages 279–288, 2002.
22. Xiaokui Shu,Danfeng Yao,"Privacy-preserving detection of sensitive data exposure" in vol.,10,No.5,May 2015.
23. X. Shu and D. Yao, "Data leak detection as a service," in Proc. 8th Int.Conf. Secur. Privacy Commun. Netw., 2012, pp. 222–240.
24. Y.Li,S.Zhu,L.Wang, and S.Jajodia " A privacy enhanced micro- aggregation method", In Poe. Of 2nd International Symposium on Foundations of Information and Knowledge Systems, pp148-55, 2008.
25. Z. Qin, J. Yan, K. Ren, C. W. Chen and C. Wang, "Towards efficient privacy-preserving image feature extraction in cloud computing", pp. 497-506, 2013.