Data Privacy Preservation Using K-Anonymization with Clustering Technique

A. Malaisamy
Associate Professor, Department of Computer Applications,
S. S. M. College of Engineering, Komarapalayam, Tamilnadu, India

Dr. G. M. Kadhar Nawaz
Director, Department of Computer Applications,
Sona College of Technology, Salem, Tamilnadu, India

Abstract - Nowadays, data sharing in network suggests a new threats and challenges to the individual privacy and organizational confidentiality. In recent times, data privacy preservation consumed significant interests in data mining study. For individual data privacy preservation scheme, *k*-anonymity model is used for practical approach. *k*-anonymization techniques have been the center of powerful research in the recent years. An essential condition for such strategies is to make certain anonymization of data, whereas at the same time decreasing the information loss, ensuing from data modifications. In this paper we propose an efficient individual data privacy preservation approach that uses the idea of clustering to reduce information loss and thus guarantee good data quality. The main objective of the approach is that data records that are logically similar to each other should be fraction of the identical equivalence class. The individual data privacy preservation scheme considers both the sensitive and categorical attributes for a better privacy scheme. As part of the proposed DPPSCT approach we enlarge a proper metric to estimate the information loss commenced by overview of the scheme, which works for both numeric and categorical data. Experimental evaluation is conducted to estimate the performance of the proposed scheme on separate data sets to show its purity and precision when compared to other k-anonymity generalization and suppression based methods.

Keywords: Data privacy preservation, *k*-Anonymity, Clustering, Privacy Preserving Data Mining.

I. INTRODUCTION

Data mining has appeared as an input tool for a broad range of applications, sorting from national safety to market examination. Many of these applications engross mining data that comprise private and responsive information about users. For case, medical research might be conducted by relating data-mining algorithms on patient medical records to recognize disease patterns. A widespread performance is to de-identify data before discharging it and concerning a data-mining process so as to preserve the privacy of users. On the other hand, private information about users might be opened to the datasets when linking de-identified data with outer public sources. Industries, organizations, and governments must assure loads for electronic discharge of information besides demands of privacy from individuals whose individual data might be revealed by the process.

Privacy-preserving data mining (PPDM) deals with the transaction among the efficiency of the mining procedure and confidentiality of the subjects, aspiring at reducing the privacy revelation with negligible result on mining results. A dataset fulfills with k-anonymity fortification if every individual's record amassed in the unconfined dataset cannot be renowned from in any case k-1 individuals whose data also emerges in the dataset. This fortification assures that the possibility of recognizing an individual supported on the unrestricted data in the dataset does not go beyond 1/k. Generalization and inhibition are the most widespread techniques employed for de-identification of the data in k-anonymity standard algorithms.

Generalization comprises of alternate characteristic values with semantically reliable but less specific values. For instance, the month of birth can be restored by the year of birth which happens in more records, so that the identification of a specific individual is more difficult. Generalization preserves the accuracy of the data at the testimony level but fallout in less precise information that might involve the accurateness of device learning algorithms functioned on the k-anonymous dataset. Diverse schemes employ different techniques for choosing the attributes and records for simplification as well as the generalization method.

Suppression specifies to eliminating a definite attribute value and restoring incidences of the value with a unique value "?", representing that some value can be located as a substitute. Suppression can radically decrease the

worth of the data if not correctly employed. This is the motivation why most k-anonymity associated studies have paying attention on generalization.

Quasi-identifier is a place of features whose connected values might be helpful for connecting with one more dataset to re-recognize the individual that is the focus of the data. One main problem of existing generalization methods is that physically produced domain hierarchy trees are essential for each quasi-identifier quality of datasets previous to k-anonymity can be useful. A quasi-identifier is a negligible place of attributes $X1; :::;Xd$ in table T that can be connected with exterior information to re-recognize personality records.

The major purpose of the *k*-anonymity representation is therefore to change a table so that no one can create high-probability relations among records in the table and the equivalent things. To facilitate this objective, the *k*-anonymity representation needs that any verification in a table be interchangeable from no less than $(k-1)$ other records concerning the prearranged quasi-identifier. A collection of records that are impossible to tell apart to all other is frequently referred to as an equality class. By implementing the *k*-anonymity condition, it is definite that although an opponent recognizes that a *k*-anonymous table surrounds the proof of an exacting person and also recognizes a few of the quasi-identifier. Usually the data includes extremely exact clatters that are critical to association. To construct a classifier, noises need being complete into samples that are widespread by more records in the similar class.

II. LITERATURE REVIEW

Privacy Preserving Data Mining (PPDM) is a comparatively novel study locale that aspires to avoid the contravention of isolation that might effect from data-mining processes on datasets. PPDM algorithms adjust unique datasets so that seclusion is conserved even after the mining procedure is stimulated, as simply touching the mining consequences excellence. Existing study on privacy-preserving data distributing hubs on relational data: in this circumstance, the purpose is to implement privacy-preserving paradigms, for instance k-anonymity and ℓ-diversity, as reducing the information failure acquired in the anonymizing procedure (i. e. , maximize data efficacy). The author in [1] proposes two groups of new anonymization methods for spare high-dimensional data.

The author in [7] revise the anonymization of time-series as annoying to hold up compound queries, for instance range and prototype identical queries, on the available data. The conformist k-anonymity representation cannot efficiently lecture to this crisis as it might undergo harsh pattern defeat. A new anonymization representation is offered termed (k, p)-anonymity for prototype loaded time-series. To present this data possession solitude [8], the cloud's dispersed calculating resources are leveraged to realize an anonymizing course supported on or, from side to side which users present personal data and jobs. The paper [11] enlarges the meaning of k-anonymity to numerous relations and demonstrates that formerly proposed methodologies either not succeed to defend privacy or excessively decrease the usefulness of the data in a numerous relation surroundings.

In [4], the authors present a new structure for modeling, analyzing and assessing anonymity in sensor networks. The author in [5] believe two techniques to reidentify unidentified position traces, objective tracking, and house recognition, and scrutinize that branded solitude algorithms cannot attain high submission accuracy necessities or not succeed to supply privacy assurances for drivers in low-density areas. To conquer these confronts, the author obtain a novel time-to-confusion measure to set apart privacy in a locational data set and suggest a revelation organize algorithm (called uncertainty-aware path cloaking algorithm) that selectively exposes GPS samples to limit the maximum time-to-confusion for all vehicles [12].

One method to facilitate effective data mining whilst preserving privacy is to anonymize the data set that comprises confidential information concerning subjects before being unrestricted for data mining. In [2], we suggest a novel technique for attaining k-anonymity termed K-anonymity of Classification Trees Using Suppression (kACTUS). In kACTUS, proficient multidimensional inhibition is achieved, i.e., values are suppressed only on definite records based on other attribute values, devoid of the requirement for physically formed domain hierarchy trees [3]. A useful technique to contest such linking attacks, termed k-anonymization, is anonymizing the connecting characteristics so that at least k released records contest each value mixture of the linking attributes [4].

In definite applications, the positions of events accounted by a sensor network require to stay behind anonymous. That is, illegal observers must be incapable to notice the starting point of such events by examining the network traffic [6]. Termed as the source anonymity trouble [8], this crisis has appeared as an significant topic in the safety of wireless sensor networks, with range of methods supported on diverse adversarial hypothesis being proposed. In [9], we suggest a novel framework for modeling, examining and computing anonymity in sensor networks. To solve the issues, in this work, we are going to present a technique for k-anonymity problem to increase the individual data privacy preservation scheme.

III. PROPOSED DATA PRIVACY PRESERVATION USING k-ANONYMITY CLUSTERING METHODOLOGY

The main objective of the proposed individual data privacy preservation approach is that the *k*-anonymization trouble can be analyzed as a clustering problem. Clustering is the analysis of dividing a collection of objects into groups such that objects in the similar group are more analogous to each other than objects in further groups with regard to some distinct similarity principles. Naturally, a best possible solution of the *k*-anonymization problem is certainly a position of equivalence classes such that records in the similar equivalence class are very parallel to each other, thus involving a minimum generalization. The architecture diagram of the proposed data privacy preservation scheme using clustering techniques is shown in figure 3.1.

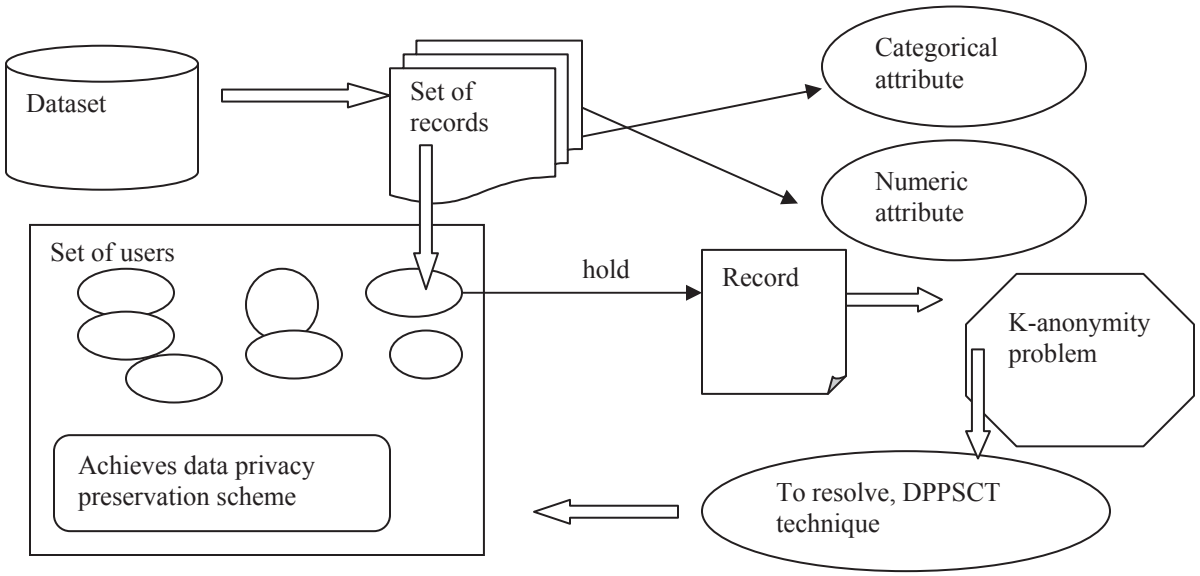

Figure 3.1 Architecture diagram of the proposed DPPSCT technique

From the figure 3.1, it is being observed that the proposed DPPSCT technique provides an individual data privacy preservation process. The k-anonymity problem on the set of data is resolved by the proposed DPPSCT technique. The proposed DPPSCT technique is done by the clustering the set of records from the dataset and the process of identifying the best record are identified reliably. Through this, the individual data privacy preservation scheme achieved.

3.1 k-anonymization as a clustering problem

Typical clustering problems need that a precise number of clusters be established in solutions. On the other hand, the *k*-anonymity crisis does not contain a limitation on the number of clusters; as an alternative, it needs that every cluster holds at least k records. Therefore, we create the *k*-anonymity as a clustering problem, termed as the k-member clustering problem.

The *k*-member clustering crisis is to discover a collection of clusters from a specified set of *n* records with each cluster has at least *k* ($k \leq n$) data points and that the average values of all intra-cluster distances is reduced. Officially, let *S* be a collection of *n* records and *k* the precise anonymization constraint. Then the best solution of the *k* clustering crisis is a collection of clusters $E = \{e1, \ldots, em\}$ which is expressed in Figure 3.2,

Step 1; For all $\forall i \neq j \in \{1,...,m\}, e_i \cap e_j = \phi$

Step 2: $U_{i=1,....,m} e_i = S,$

Step 3: $\forall e_i \in \varepsilon, |e_i| \geq kand$

Step 4: $\sum_{l=1,...,m} |e_l| . MAX_{i,j=1,...,|e_l|}$

Here $|e|$ is the size of the cluster. We think the average values of all intra-cluster distances, where an intra-cluster space of a cluster is termed as the utmost distance among any two data points in the group (i.e., the distance of the cluster).

At the heart of every clustering problem are the distance functions that determine the dissimilarities amongst data points and the cost utility which the clustering crisis tries to reduce. The distance functions are frequently processed by the type of data (i.e., numeric or categorical) being grouped, whilst the cost function is termed as the precise intention of the clustering problem. A distance function in a clustering crisis is identified with the measures how different two data points are. As the data we judge in the *k*-anonymity crisis are person-specific records that naturally comprises of both numeric and categorical attributes, we require a distance function that be able to hold both types of data at the identical time. For a numeric attribute, the disparity among two numeric and categorical values (e.g., $|x-y|$) logically explains the dissimilarity (i.e., distance) of the values. This measure is also appropriate for the *k*-anonymization crisis.

Consider D be a numeric domain. Then the normalized space among two values *vi, vj* □ *D* is defined as:

$$\delta_N(n1, n2) = |n1 - n2| / |D|$$

Where |D| is the size of the respective domain.

For categorical attributes, nevertheless, the disparity is no longer appropriate as most of the categorical areas cannot be specified in any specific order. The most uncomplicated resolution is to presume that every value in such a field is regularly diverse to each other; e.g., the distance of two values is 0 if they are the equal, and 1 if unlike. On the other hand, some areas might have some semantic associations between the values. In such areas, it is enviable to classify the distance functions supported on the existing associations. Such associations can be simply detained in a taxonomy tree.

Let C = $\{c_1, \ldots, c_k\}$ be a group of classes (i.e., equivalence class) where the quasi-identifier comprises of numeric attributes $NA_1, \ldots, NA_m$ and categorical attributes $CA_1, \ldots, CA_n$. Let TC_i be the taxonomy tree termed as the area of categorical attribute. Assume MINI_N and MAX_N be the min and max values in *e* with respect to attribute *Ni*, and let U_{CA} be the combination of values in C respecting attribute *CAi*.. Then the quantity of information loss happened by simplifying *e*, specified by *IL* (C), is defined as:

$$IL(C) = |C| .\left(\sum_{i=1,\ldots,m} \frac{MAX_N - MIN_N}{|N_i|}\right) + \sum_{j=1,\ldots,n} \frac{H(\Lambda(U_{CAj}))}{H(T_{Cj})}$$

Where |C| is the number of records, |*N*| represents the size of numeric domain *N*, $\Lambda(\cup Cj)$ is the subtree entrenched at the lowest widespread antecedent of every value in $\cup Cj$, and $H(T)$ is the height of taxonomy tree T.

3.2 Anonymization algorithm

Armed with the distance and cost functions, we are now prepared to confer the *k*-member clustering algorithm. As in most clustering problems, an comprehensive search for an best result of the *k*-member clustering is potentially exponential. With the intention of precisely distinguish the computational difficulty of the problem; we classify the *k*-member clustering trouble as a result problem as follows.

For a specified number of records, a clustering method is termed as $\varepsilon = \{c_1, \ldots, c_l\}$ such that,

Step 1: $|c_i| \geq k, 1 < k \leq n,$ the size of each cluster is greater than or equal to a positive integer *k*, and

Step 2: $\sum_{i=1,\ldots,l} IL(c_i) < C$ the clustering scheme is less than a positive constant *c*.

In most *k*-anonymity process, the main objective is greatly positioned on the quasi-identifier, and consequently other attributes are frequently disregarded. On the other hand, these attributes earn more vigilant consideration. Actually, we desire to diminish the deformation of quasi-identifier not only as the quasi-identifier itself is significant information, but also as a more precise quasi-identifier will direct to good prophetic models on the distorted table. In fact, the association among the quasiidentifier and other attributes can be considerably damaged or disturbed owing to the uncertainty commenced by the simplification of the quasi-identifier. Thus, it is serious that the generalization procedure does defend the discrimination of classes employing quasi-identifier. The algorithm below describes the process of clustering with the set of records.

Input: Set of records S, threshold value t

Output: set of clusters which contains at least t records
Step 1: If S < k
Step 2: Return set of records
Step 3: End if
Step 4: Pick some set of records r randomly
Step 5: While (|S|>k)
Step 6: S=S-r
Step 7: c= {r}
Step 8: While (|c|<k)
Step 9: S=S-r
Step 10: c=cU{r}
Step 11: End while
Step 12: With the set of records from S
Step 13: Identify the best cluster
Step 14; End
Step 15: End
In precise, the above 15 steps is now required to decide records with the similar class label for a cluster, and the amount of enforcement is proscribed by the weight of sentence. With this minor variation, our proposed DPPSCT algorithm can efficiently decrease the cost of categorization metric without raising much information loss.

IV. EXPERIMENTAL EVALAUTION

The main objective of the evaluation of experiments was to examine the performance of the proposed DPPSCT approach in terms of running time, anonymity level, and scalability. To accurately evaluate the proposed DPPSCT approach, we also compared our implementation with the existing works like kACTUS and ACD (Anonymizing classification data). The experiments were performed on a 2.66 GHz Intel *IV* processor machine with 1 GB of RAM. The operating system used in the machine was Microsoft Windows XP Professional Edition, and the implementation was built and run in Java 2 Platform, Standard Edition 5.0. For our experiments, we used both the synthetic and real dataset from the UC Irvine Machine Learning Repository, which is considered as a benchmark for evaluating the performance of the proposed DPPSCT *k*-anonymity algorithms. We detached records with mislaid values and maintained only the original attributes. For *k*-anonymization, we measured *{age*, *education*, *occupation*, *gender}* as the quasi-identifier. Among these sets of quasi-identifiers, *age* and *education* were treated as numeric attributes while the other six attributes were treated as categorical attributes. Besides to that, we also preserved the respective attribute to appraise the classification metric. The performance of the proposed data privacy preservation using k anonymization with clustering techniques is measured in terms of

i) running time
ii) anonymity level and
iii) scalability

V. RESULTS AND DISCUSSION

In section v, we are going to compare the proposed data privacy preservation using k anonymization with clustering techniques with the existing kACTUS and Anonymizing classification data. The table and graph below describes the performance of the proposed scheme.

No. of records	Running time (sec)		
	Proposed DPPSCT	Existing kACTUS	Existing ACD
200	52	95	100
400	75	110	125
600	89	120	150
800	97	135	165
1000	110	150	180

Table 5.1: No. of records vs. running time

The above table (table 5.1) shows the running time required to process the privacy preservation scheme and compared the results with the existing kACTUS and anonymizing classification data.

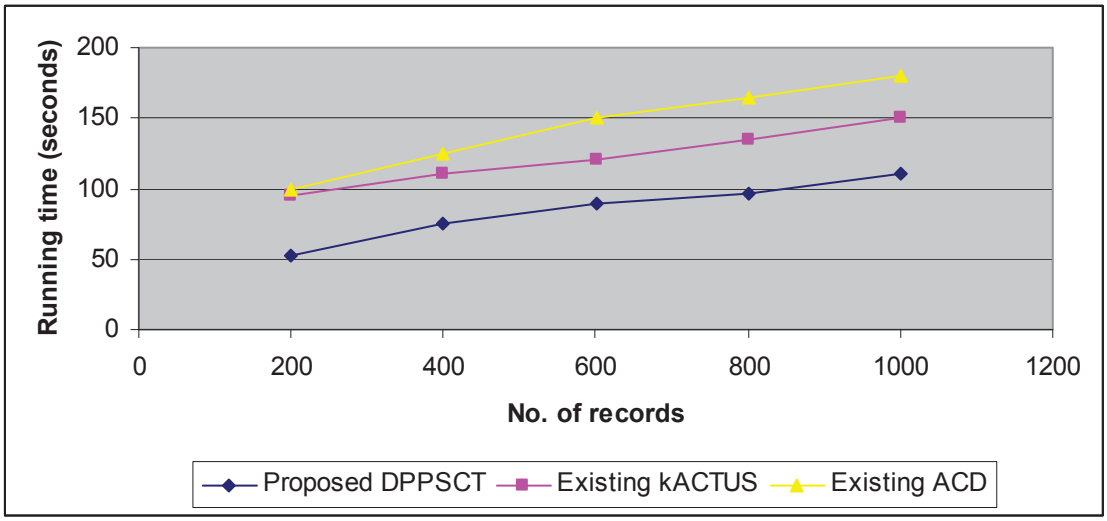

Figure 5.1: No. of records vs. running time

Figure 5.1 shows the process of running time required to process the individual data privacy preservation scheme with the corresponding clustering techniques. Running time is measured in terms of seconds (secs). Running time is in lesser ratio in the proposed DPPSCT scheme compared with the existing kACTUS and ACD. In the proposed DPPSCT scheme, while records in the dataset increases, execution time decreases dramatically. Compared to the existing works like kACTUS and ACD, the proposed DPPSCT consumes less running time to cluster the set of records with the respective threshold value and the variance is 50-60% less in the proposed DPPSCT.

No. of records	Anonymity level (%)		
	Proposed DPPSCT	Existing kACTUS	Existing ACD
200	75	65	50
400	80	70	55
600	83	72	65
800	87	75	70
1000	95	73	73

Table 5.2: No. of records vs. Anonymity level

The above table (table 5.2) represents the level of anonymity required for privacy preservation scheme and compared the results with the existing kACTUS and anonymizing classification data.

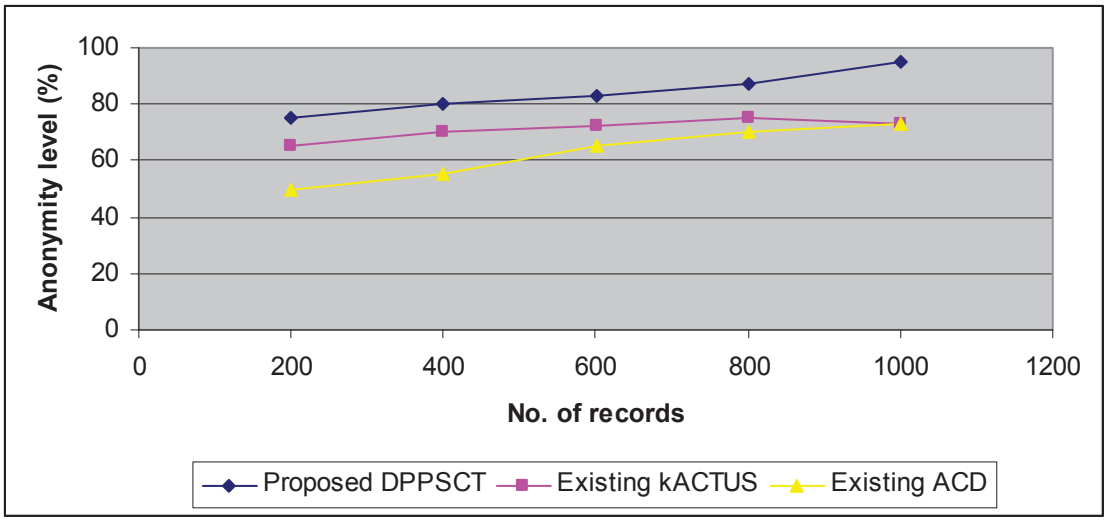

Figure 5.2: No. of records vs. Anonymity level

Figure 5.2 shows the process of process of number of records and anonymity level of individual data with the corresponding clustering techniques. Anonymity level is measured in terms of rate (%). Anonymity level is in higher ratio in the proposed DPPSCT scheme compared with the existing kACTUS and ACD. In the proposed DPPSCT scheme, while records in the dataset increases, level of anonymity increases dramatically. Compared to the existing works like kACTUS and ACD, the proposed DPPSCT provides high level of anonymity to protect the set of individual records with the respective threshold value and the variance is 55-65% less in the proposed DPPSCT. The anonymity level of the proposed DPPSCT scheme on set of records in size from 200 to 1200 is high.

No. of records	Scalability (%)		
	Proposed DPPSCT	Existing kACTUS	Existing ACD
200	70	55	45
400	75	60	50
600	80	63	55
800	84	68	58
1000	90	72	62

Table 5.3: No. of records vs. Scalability

The above table (table 5.3) shows the level of scalability required for individual data privacy preservation scheme and compared the results with the existing kACTUS and anonymizing classification data.

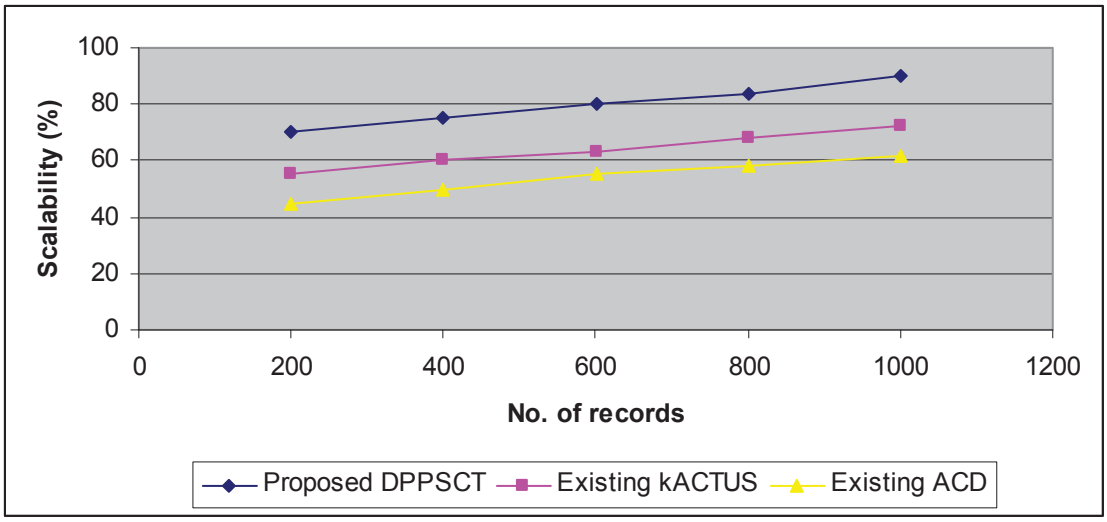

Figure 5.3: No. of records vs. Scalability

Figure 5.3 shows the process of number of records and scalability level of individual data with the corresponding clustering techniques. Scalability level is analyzed based on the analysis of both the data anonymity and privacy level of the anonymized data. Scalability level of the data anonymity and privacy over set of records in the database are measured. Scalability level is in higher ratio in the proposed DPPSCT scheme compared with the existing kACTUS and ACD, because the proposed DPPSCT improves the preservation of sensitive and categorical attributes of the data objects. In the proposed DPPSCT scheme, while records in the dataset increases, level of scalability increases dramatically. Compared to the existing works like kACTUS and ACD, the proposed DPPSCT provides high level of scalability to protect the set of individual records with the respective threshold value and the variance is 65-75% less in the proposed DPPSCT. The anonymity level of the proposed DPPSCT scheme on set of records in size from 200 to 1200 is high.

VI. CONCLUSION

In this paper, we proposed an efficient *k*-anonymization DPPSCT algorithm by transforming the *k*-anonymity problem to the *k*-member clustering problem. Also proposed two important elements of clustering, that is, distance and cost functions, which are specifically tailored for the *k*-anonymization problem. We emphasize that our proposed DPPSCT scheme naturally captures the data distortion introduced by the generalization process and is general enough to be used as a data quality metric for any *k*-anonymized dataset. Performance evaluations are conducted to show the effectiveness of the proposed scheme with the existing works like kACTUS and ACD. Experimental results showed that the proposed DPPSCT scheme provides an efficient k-anonymization for individual data privacy preservation process. Compared to the other works like kACTUS and ACD, the proposed DPPSCT scheme provides 60-70% efficiency in k-anonymization problem.

REFERENCES

[1] Gabriel Ghinita, Panos Kalnis and Yufei Tao, "Anonymous Publication of Sensitive Transactional Data", IEEE Transactions on Knowledge and Data Engineering, 2011.
[2] Slava Kisilevich, Lior Rokach, Yuval Elovici and Bracha Shapira, "Efficient Multi-Dimensional Suppression for K-Anonymity", IEEE transactions on knowledge and data engineering, 2010.
[3] Benjamin C. M. Fung, Ke Wang and Philip S. Yu, "Anonymizing Classification Data for Privacy Preservation", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 2007.
[4] Basel Alomair, Andrew Clark, Jorge Cuellary and Radha Poovendran, "Towards a Statistical Framework for Source Anonymity in Sensor Networks", IEEE Transactions on Mobile Computing, 2011.
[5] Baik Hoh and Marco Gruteserand Hui Xiong, "Achieving Guaranteed Anonymity in GPS Traces via Uncertainty-Aware Path Cloaking", : IEEE Transactions on Mobile Computing, 2010.
[6] Roberto J. Bayardo and Rakesh Agrawal, " **Data Privacy Through Optimal k-Anonymization**", Proceedings.21st International Conference on Data Engineering, 2005.
[7] Xuan Shang, Ke Chen, Lidan Shou, Gang Chen and Tianlei Hu, "Supporting Pattern-Preserving Anonymization For Time-Series Data", IEEE Transactions on Knowledge and Data Engineering, 2011.

[8] Safwan Mahmud Khan and Kevin W. Hamlen, "AnonymousCloud: A Data Ownership Privacy Provider Framework in Cloud Computing", IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2012.
[9] Yang Du, Tian Xia, Yufei Tao, "On Multidimensional k-Anonymity with Local Recoding Generalization", IEEE 23rd International Conference on Data Engineering, 2007. ICDE 2007.
[10] Kristen LeFevre, David J. DeWitt and Raghu Ramakrishnan, "Mondrian Multidimensional K-Anonymity", :Proceedings of the 22nd International Conference on Data Engineering, 2006. ICDE '06..
[11] Nergiz N. Ercan, Clifton Chris and Nergiz A. Erhan, "Multirelational k-Anonymity", IEEE Transactions on Knowledge and Data Engineering, 2009.
[12] Zhi-yuan Li, Liang-min Wang and Si-guang Chen, "Network Coding-Based Mutual Anonymity Communication Protocol for Mobile P2P Networks", IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2012.