International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 1, No 1, 2014. ISSN: 2348-6600
Reference ID: IJCS-031 PAGE NO: 166-171

K-Anonymity Multidimensional Suppression

P.PONSEKAR[#1], S.R.SARANYA*[2]
[1]Research Supervisor,
Department of Computer Science,
Kovai Kalaimagal College of Arts & Science,
Coimbatore-641 109, India.
Pon84sekar@gmail.com
[2]Research Scholar,
Kovai Kalaimagal College of Arts & Science,
Coimbatore-641 109, India.
srsaranya687@gmail.com

Abstract— **Knowledge Discovery in Databases (KDDs) is the process of identifying valid, novel, useful, and understandable patterns from large data sets. Data Mining is the core of the KDD process, involving algorithms that explore the data, develop models, and discover significant patterns. Data mining has emerged as a key tool for a wide variety of applications, ranging from national security to market analysis. Many of these applications involve mining data that include private and sensitive information about users. Many applications that employ data mining techniques involve mining data that include private and sensitive information about the subjects. One way to enable effective data mining while preserving privacy is to anonymize the data set that includes private information about subjects before being released for data mining. One way to anonymize data set is to manipulate its content so that the records adhere to k-anonymity. Two common manipulation techniques used to achieve k-anonymity of a data set are generalization and suppression. Generalization refers to replacing a value with a less specific but semantically consistent value, while suppression refers to not releasing a value at all. Generalization is more commonly applied in this domain since suppression may dramatically reduce the quality of the data mining results if not properly used.**

***Index Terms*— Privacy-preserving data mining, k-anonymity, deindentified data, decision trees.**

INTRODUCTION

Intuitively, a cluster on a set of numeric attributes identifies a "dense region" in the attribute space. That is, the region consists of a significantly larger number of tuples than expected. We generalize this intuitive notion for the class of hyper-rectangular clusters to the categorical domain.

As an illustrative example, the region [1,2]x[2,4]x[3,5] may correspond to a cluster in the 3-d space spanned by three numeric attributes. In general, the class of rectangular regions can be expressed as the cross product of intervals. Since domains of categorical attributes are not ordered, the concept of an interval does not exist. However, a straightforward generalization of the concept of an interval to the categorical domain is a set of attribute values.

Consequently, the generalization of rectangular regions in the numeric domain to categorical domain is the cross product of sets of attribute values. We call such regions interval regions. Intuitively, a cluster consists of a significantly larger number of tuples than the number expected if all attributes were independent. In addition, a cluster also extends to as large a region as possible. We now formalize this notion for categorical domains by first defining the notion of a tuple belonging to a region, and then the support of a region, which is the number of tuples in the dataset that belong to the region.

It extends kACTUS to handle unusually large attribute value domains as well as to identify clusters in subspaces. The results of a detailed evaluation of the speed and scalability of kACTUS on synthetic and real datasets and we examined whether the clusters discovered where intuitive and sensible. The compared the performance of kACTUS with the performance of STIRR. Until now, assumed that the domains of categorical attributes are such that the inter-attribute summary of any pair of attributes and the intra-attribute summary of any attribute fits in main memory. For the sake of completeness, we modify the summarization phase of kACTUS to handle arbitrarily large domain sizes. Recall that the summary information only consists of strongly connected pairs of attribute values.

For large domain sizes, the number of strongly connected attribute value pairs (either from the same or from different attributes) relative to the number of all possible attribute

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 1, No 1, 2014. ISSN: 2348-6600
Reference ID: IJCS-031 PAGE NO: 166-171

value pairs is very small. It exploit this observation to collapse sets of attribute values on each attribute into a single attribute value thus creating a new domain of smaller size. The intuition is that if a pair of attribute values in the original domain is strongly connected, then the corresponding pair of transformed attribute values are also strongly connected, provided the threshold for strong connectivity between attribute values in the transformed domain is the same as that for the original domain. An application of CACTUS to a combination of two sets of bibliographic entries. The results from the application show that CACTUS finds intuitively meaningful clusters from the dataset thus supporting our definition of a cluster.

I. BACKGROUND STUDY

Researchers have shown that the HIPAA privacy rules have affected significantly their ability to perform retrospective, chart-based research. Privacy-preserving data mining (PPDM) deals with the trade-off between the effectiveness of the mining process and privacy of the subjects, aiming at minimizing the privacy exposure with minimal effect on mining result PPDM is a relatively new research area that aims to prevent the violation of privacy that might result from data mining operations on data sets.

PPDM algorithms modify original data sets so that privacy is preserved even after the mining process is activated, while minimally affecting the mining results quality. Verykios et al. classified existing PPDM approaches based on five dimensions:

1. Data distribution
2. Data modification
3. Data mining algorithms.
4. Data or rule hiding.
5. privacy preservation. One of the PPDM techniques is k-anonymity.

The k-anonymity concept requires that the probability to identify an individual by linking databases does not exceed 1/k. Generalization is the most common method used for deidentification of the data in k-anonymity-based algorithms. Generalization consists of replacing specific data with a more general value to prevent individual identification

Privacy-preserving data mining (PPDM) deals with the trade-off between the effectiveness of the mining process and privacy of the subjects, aiming at minimizing the privacy exposure with minimal effect on mining results.

The main goal of this study is to introduce a new k-anonymity algorithm which is capable of transforming a nonanonymous data set into a k-anonymity data set. The transformation is aimed to achieve a predictive performance of a classifier trained on the transformed data set as similar as possible to the performance of a classifier trained on the original data set. Our approach wraps an existing classification tree induction algorithm and is referred to as K-Anonymity of Classification Trees Using Suppression (kACTUS). The classification tree inducer is used to induce a classification tree from the original data set which indicates how the attribute values affect the target class. The classification tree can be easily interpreted by a machine in order to perform the k-anonymity process. Each path from the root to a leaf can be treated as a classification rule. A subset of the data set is ascribed with each leaf. If this subset size is greater than k, then this subset of instances can be easily anonymized by suppressing all the quasi-identifier attributes that are not referenced in one of the nodes along the path from the root. Assuming that all attributes are categorical, then all quasi-identifiers attributes that are referenced have the same value for all the instances of the subset, and thus, there is no need to suppress or generalize their values). In our terminology, a leaf node complies with the k-anonymity if the number of instances that ascribed to this leaf is higher or equal to k. If all the leaves in the tree comply with the k-anonymity, the data set can be k-anonymized by suppression as described above. For leaves that do not comply with the k-anonymity, by adequately pruning them, one can obtain a new leaf which may comply with the k-anonymity. We utilize the fact that the order in which the attributes are selected for the decision tree usually implies their importance for predicting the class. Thus, by pruning the rules in a bottom-up manner, we suppress the least significant quasi-attributes.

METHODOLOGY

A. kCACTUS Algorithm

Clustering is an important data mining problem. Most of the earlier work on clustering focused on numeric attributes which have a natural ordering on their attribute values. Recently, clustering data with categorical attributes, whose attribute values do not have a natural ordering, has received some attention. However, previous algorithms do not give a formal description of the clusters they discover and some of them assume that the user post-processes the output of the algorithm to identify the final clusters.

It introduces a novel formalization of a cluster for categorical attributes by generalizing a definition of a cluster for numerical attributes. Then describe a very fast

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 1, No 1, 2014. ISSN: 2348-6600
Reference ID: IJCS-031 PAGE NO: 166-171

summarization based algorithm called kACTUS that discovers exactly such clusters in the data. kACTUS has two important characteristics. First, the algorithm requires only two scans of the dataset, and hence is very fast and scalable. kACTUS can find clusters in subsets of all attributes and can thus perform a subspace clustering of the data. This feature is important if clusters do not span all attributes, a likely scenario if the number of attributes is very large. In a thorough experimental evaluation, we study the performance of kACTUS on real and synthetic datasets.

The goal of clustering, in general, is to discover dense and sparse regions in a dataset. Most previous work in clustering focused on numerical data whose inherent geometric properties can be exploited to naturally define distance functions between points. However, many datasets also consist of categorical attributes on which distance functions are not naturally defined. Recently, the problem of clustering categorical data started receiving interest [GKR98, GRS99].

As an example, consider the MUSHROOM dataset in the popular UCI Machine Learning repository [CBM98]. Each tuple in the dataset describes a sample of gilled mushrooms using twenty two categorical attributes. For instance, the cap color attribute can take values from the domain {brown, buff, cinnamon, gray, green, pink, purple, red, white, yellow}. It is hard to reason that one color is "like" or "unlike" another color in a way similar to real numbers.

An important characteristic of categorical domains is that they typically have a small number of attribute values. For example, the largest domain for a categorical attribute of any dataset in the UCI Machine Learning repository consists of 100 attribute values (for an attribute of the Pendigits dataset). Categorical attributes with large domain sizes typically do not contain information that may be useful for grouping tuples into classes. For instance, the CustomerId attribute in the TPC-D database benchmark [Cou95] may consist of millions of values; given that a record (or a set of records) takes a certain CustomerId value (or a set of values), it cannot infer any information that is useful for classifying the records. Therefore, it is different from the age or geographical location attributes which can be used to group customers based on their age or location or both. Typically, relations contain 10 to 50 attributes; hence, even though the size of each categorical domain is small, the cross product of all their domains and hence the relation itself can be very large.

It introduces a fast summarization-based algorithm called kACTUS[2] for clustering categorical data. kACTUS exploits the small domain sizes of categorical attributes. The central idea in kACTUS is that summary information constructed from the dataset is sufficient for discovering well-defined clusters. The properties that the summary information typically fits into main memory and that it can be constructed efficiently typically in a single scan of the dataset result in significant performance.

- It formalizes the concept of a cluster over categorical attributes.
- It introduces a fast summarization-based algorithm kACTUS for clustering categorical data.
- Then extend kACTUS to discover clusters in subspaces, especially useful when the data consists of a large number of attributes.
- In an extensive experimental study, we evaluate kACTUS and compare it with earlier work on synthetic and real datasets.

B. kACTUS Steps

Steps involved in kACTUS algorithm;

1. The inputs taken in kACTUS algorithm are original dataset, Quasi Identifier, k (threshold value).
2. The take a copy of the original dataset and work in the copy.
3. Take node in classification tree where height=1
4. K value is the user defined value.
5. Based on K value, if the count of distinct combination is greater than k, those data from table will be eliminated and thus obtain table compression.
6. Finally apply suppression for compressed table.

The output will be the anonymous dataset

C. Modules

- Preprocessing of Data Set.
- Deriving the Classification Tree.
- Implementing K-Anonymity Process.
- Applying Suppression for Numeric Attributes.
- Applying Suppression
  a) Single Dimension Database.
  b) Multi Dimension Database.
- Performance Comparison.

1.) Preprocessing of Data Set

A data set is a collection of data, usually presented in tabular form. Each column represents a particular variables and a data set has several characteristics

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 1, No 1, 2014. ISSN: 2348-6600
Reference ID: IJCS-031 PAGE NO: 166-171

which define its structure and properties. These include the number and types of the attributes or variables.

2.) Deriving the Classification Tree

In this modules, are employing a decision tree inducer (denoted by CTI) to generate a decision tree denoted by CT. The tree can be derived using various inducers. It concentrate on top-down univariate inducers which are considered the most popular decision tree inducers and include the well-known algorithms. Top-down inducers are greedy by nature and construct the decision tree in a top-down recursive manner (also known as divide and conquer). Univariate means that the internal nodes are split according to the value of a single attribute.

The decision tree is trained over the projection of the quasi-identifiers, i.e., $(CT \leftarrow CTI(\pi_{Q \cup y}(S))$. The wrapped inducer CTI should be differentiated from the target inducer I. Inducer I is applied on the anonymous data set (i.e., after applying k-anonymization process). The aim of the CTI is to reveal which quasi-identifier is more relevant for predicting the class value. Any internal node (non leaf) with less than k instances cannot be used by itself for generating the anonymous data set. if such a node is provided in the classification tree it can be pruned in advance. In many decision trees inducers. The user can control the tree growing process by setting the algorithm's parameters. Specifically, the parameter MinObj ("minimum number of instances") indicates the number of instances that should be associated with a node in order it to be considered for splitting. By setting MinObj to k, one ensures that there are no non complying internal nodes that are needed to be pruned.

3.) Implementing K-Anonymity Process

The classification tree inducer is used to induce a classification tree from the original data set which indicates how the attribute values affect the target class. The classification tree can be easily interpreted by a machine in order to perform the k-anonymity process. In this phase, we use the classification tree that was created in the first phase to generate the anonymous data set. We assume that the classification tree complies with the following properties:

- The classification tree is univariate, i.e., each internal node in the tree refers to exactly one attribute.
- All internal nodes refer to a quasi-identifier attributes. This is true because the decision tree was trained over the projection of the quasi-identifier set $(\pi_{Q \cup y}(S))$.
- Assuming a top-down inducer, the attributes are sorted (from left to right) according to their significance for predicting the class (where the rightmost relates to the least significant attribute).
- Complete Coverage: Each instance is associated with exactly one path from root to leaf.

Our supervised k-anonymity process is described in procedure Anonymize .The input to the Anonymize procedure includes the original data set S, quasi-identifier set Q, classification tree CT, and the anonymity threshold k. The Output of the algorithm is a k-anonymous data set denoted by S'.

4.) Applying Suppression for Numeric Attributes

In this module relax the assumption that the quasi identifier includes only categorical attributes. For this, need to revise only the suppressed function in the algorithm. The Suppress(R, Q, and V).the input will be R (Data Set), Q (The Quasi-identifier set, V (Antecedents) and output: R' (Suppressed dataset)

In this function, if the attribute a is included in v but it appears with an inequality predicate (i.e., it is a numeric attribute), then fill all the instances of the attribute A in R with a value that is equal to the mean of all a values in R.

3.3.5 Applying Suppression

Suppression refers to removing a certain attribute value and replacing occurrences of the value with a special value "?," indicating that any value can be placed instead.

a)Single Dimension Database

In this module by using a single-dimension suppression operator for attributes for which domain taxonomy does not exist. The performance of a single-dimension suppression operator is quite limited. Because it suppresses a certain value in all tuples without considering values of other attributes. This "oversuppression" may lead to unnecessary loss of data.

b)Multi Dimension Database

Algorithm uses multidimensional recoding, i.e., suppression of values is applied only on certain tuples, depending on other attribute values. It suggest a practical and effective approach that provides an improved predictive performance in comparing to existing approaches and does not require the generation of manual domain generalization taxonomy.

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 1, No 1, 2014. ISSN: 2348-6600
Reference ID: IJCS-031 PAGE NO: 166-171

4. EXPERIMENTAL RESULT

The kACTUS' predictive performance is better than that of existing k-anonymity algorithms. In order to evaluate the performance of the proposed method for applying k-anonymity to a data set used for classification tasks, a comparative experiment was conducted on benchmark data sets. The new method also shows a higher predictive performance when compared to existing state-of-the-art methods. Examining kACTUS with other decision trees inducers; revising kACTUS to overcome its existing drawbacks; extending the proposed method to other data mining tasks (such as clustering and association rules) and to other anonymity measures (such as l-diversity) which respond to different known attacks against k-anonymity, such as homogeneous attack and background attack. kACTUS scales well with large data sets and anonymity levels. When compared to the kADET algorithm, kACTUS is not restricted to a decision tree classifier, and its output can be used by any induction algorithms

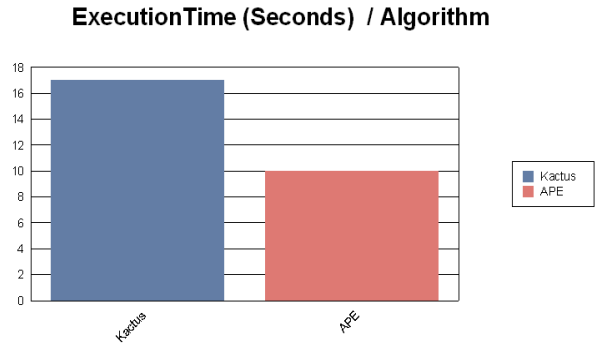

Fig.1 Execution Time of kACTUS and APE Algorithms

5.CONCLUSION

In this research presented a new method for preserving the privacy in classification tasks using k-anonymity. The proposed method requires no prior knowledge regarding the domain hierarchy taxonomy and can be used by any inducer. The new method also shows a higher predictive performance when compared to existing state-of-the-art methods. Additional issues to be studied further include: Examining kACTUS with other decision trees inducers revising.

kACTUS to overcome its existing drawbacks; extending the proposed method to other data mining tasks (such as clustering and association rules) and to other anonymity measures (such as l-diversity) which respond to different known attacks against k-anonymity, such as homogeneous attack and background attack. Further, *l*-diversity is insufficient to prevent similarity attack. Limitation of *t*-closeness: The *t*-closeness model protects against sensitive attributes disclosure by defining semantic distance among sensitive attributes. further include: Examining kACTUS with other decision trees inducers; revising kACTUS to overcome its existing drawbacks; extending the proposed method to other data mining tasks (such as clustering and association rules) and to other anonymity measures (such as l-diversity) which respond to different known attacks against k-anonymity, such as homogeneous attack and background attack.

REFERENCES

[1] M. Kantarcioglu, J. Jin, and C. Clifton, "When Do Data Mining Results Violate Privacy?" *Proc. 2004 Int'l Conf. Knowledge Discovery and Data Mining, pp. 599-604, 2004.*

[2] L. Rokach, R. Romano, and O. Maimon, "Negation Recognition in Medical Narrative Reports," Information Retrieval, *vol. 11, no. 6, pp. 499-538, 2008.*

[3] M.S. Wolf and C.L. Bennett, "Local Perspective of the Impact of the HIPAA Privacy Rule on Research," Cancer-Philadelphia Then Hoboken, *vol. 106, no. 2, pp. 474-479, 2006.*

[4] P. Samarati and L. Sweeney, "Generalizing Data to Provide Anonymity When Disclosing Information," Proc. 17th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems, *vol. 17, p. 188, 1998.*

[5] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness, and Knowledge-Based Systems, *vol. 10, no. 5, pp. 557-570, 2002.*

[6] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," Int'l J. Uncertainty, Fuzziness, and Knowledge-Based Systems, *vol. 10, no. 5, pp. 571-588, 2002.*

[7] B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. 21st IEEE Int'l Conf.Data Eng. (ICDE '05*), pp. 205-216, Apr. 2005.*

[8] K. Wang, P.S. Yu, and S. Chakraborty, "Bottom-Up Generalization: A Data Mining Solution to Privacy Protection," Proc. Fourth IEEE Int'l Conf. Data Mining, *pp. 205-216, 2004.*

[9] L. Tiancheng and I. Ninghui, "Optimal K-Anonymity with Flexible Generalization Schemes through Bottom-Up Searching," Proc. Sixth IEEE Int'l Conf. Data Mining Workshops, *pp. 518-523,2006.*

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 1, No 1, 2014. ISSN: 2348-6600
Reference ID: IJCS-031 PAGE NO: 166-171

[10] S.V. Iyengar, "Transforming Data to Satisfy Privacy Constraints," Proc. Eighth ACM SIGKDD, pp. 279-288, 2002.
[11] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., *vol. 19, no. 5, pp. 711-725, May 2007.*
[12] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full Domain k-Anonymity," *Proc. 2005 ACM SIGMOD, pp. 49-60, 2005.*
[13] A. Friedman, R. Wolff, and A. Schuster, "Providing k-Anonymity in Data Mining," Int'l J. Very Large Data Bases, *vol. 17, no. 4, pp. 789-804, 2008.*
[13] A. Friedman, R. Wolff, and A. Schuster, "Providing k-Anonymity in Data Mining," Int'l J. Very Large Data Bases, *vol. 17, no. 4, pp. 789-804, 2008.*
[14] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, *vol. 97, nos. 1/2, pp. 273-324, 1997.*
[15] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-Art in Privacy Preserving Data Mining," ACM SIGMOD Record, *vol. 33, no. 1, pp. 50-57, 2004.*
[16] A. Agrawal and R. Srikant, "Privacy Preserving Data Mining," ACM SIGMOD Record, *vol. 29, no. 2, pp. 439-450, 2000.*
[17] B. Gilburd, A. Schuster, and R. Wolff, "k-TTP: A New Privacy Model for Large-Scale Distributed Environments," Proc. 10th ACM SIGKDD, *pp. 563-568, 2004.*
[18] Z. Yang, S. Zhong, and R.N. Wright, "Privacy-Preserving Classification of Customer Data without Loss of Accuracy," Proc. Fifth Int'l Conf. Data Mining, 2005.
J. Roberto, Jr. Bayardo, and A. Rakesh, "Data Privacy through Optimal k-Anonymization," Proc. Int'l Conf. Data Eng., *vol. 21, pp. 217-228, 2005.*
[20] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical Privacy: The SuLQ Framework," Proc. 24th ACM SIGMODSIGACT- SIGART Symp. Principles of Database Systems, *pp. 128- 138, June 2005.*
[21] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Toward Privacy in Public Databases," *Proc. Theory of Cryptography Conf., pp. 363-385, 2005.*
[22] K. Wang, B.C.M. Fung, and P.S. Yu, "Template-Based Privacy Preservation in Classification Problems," *Proc. Fifth IEEE Int'l Conf. Data Mining, pp. 466-473, 2005.*
[23] E. Bertino, B.C. Ooi, Y. Yang, and R.H. Deng, "Privacy and Ownership Preserving of Outsourced Medical Data," *Proc. Int'l Conf. Data Eng., vol. 21, pp. 521-532, 2005.*
[24] G. Aggarwal, A. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Approximation Algorithms for k-Anonymity," J. Privacy Technology, 2005.
[25] A. Meyerson and R. Williams, "On the Complexity of Optimal k-Anonymity," Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems, *pp. 223-228, 2004.*
[26] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., *vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.*
[27] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," Proc. 22nd Int'l Conf. Data Eng., p. 25, Apr. 2006.
[28] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization," Proc. 12th ACM SIGKDD, pp. 277-286, 2006. 346 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *vol. 22, no. 3, march 2010*
[29] L. Sweeney, "Datafly: A System for Providing Anonymity in Medical Data," Proc. IFIP TC11 WG11.3 11th Int'l Conf. Database Security XI: Status and Prospects, *pp. 356-381, 1997.*
[30] P. Sharkey, H. Tian, W. Zhang, and S. Xu, "Privacy-Preserving Data Mining through Knowledge Model Sharing," Privacy, Security and Trust in KDD, *pp. 97-115, Springer, 2008.*

Published by SK Research Group of Companies (SKRGC).