A Details Study of Data Transformation for Privacy Preserving in Data Mining

Syed Md. Tarique Ahmad, Shameemul Haque, SM Faizanut Tauhid

Abstract— The field of privacy has seen rapid advances in recent years because of the increases in the ability to store data. In particular, recent advances in the datamining field have lead to increased concerns about privacy. While the topic of privacy has been traditionally studied in the context of cryptography and informationhiding, recent emphasis on data mining has lead to renewed interest in the field. The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. In this paper, we study the previos work regarding Privacy Preserving Data Mining.

Index Terms— Privacy Preserving, Data mining, Privacy Preserving, Data Stream, Clustering.

1 Introduction

RECENT developments in information technology have made possible the collection and analysis of millions of transactions containing personal data. These data include shopping habits, criminal records, medical histories, and credit records, among others [1]. This progress in the storage and analysis of data has led individuals and organizations to face the challenge of turning such data into useful information and knowledge.

Data mining is a promising approach to meet this challenging requirement. The area of data mining, also called Knowledge Discovery in Databases (KDD), has received special attention since the 1990s. This new research area has emerged as a means of extracting hidden patterns or previously unknown implicit information from large repositories of data [2]. The fascination with the promise of analysis of large volumes of data has led to an increasing number of successful applications of data mining in recent years. Undoubtedly, these applications are very useful in many areas such as marketing, business, medical analysis, and other applications in which pattern discovery is paramount for strategic decision making.

Despite its benenifits in various areas, the use of data mining techniques can also result in new threats to privacy and information security. The problem is not data mining itself, but the way data mining is done [3]. As Vaidya & Clifton [4] state, "Data mining results rarely violate privacy, as they generally reveal high-level knowledge rather than disclosing instances of data." However, the concern among privacy advocates is well founded, as bringing data together to support data mining projects makes misuse easier. Thus, in the absence of adequate safeguards, the use of data mining can jeopardize the privacy and autonomy of individuals.

More serious is the privacy invasion occasioned by secondary usage of data when individuals are unaware of "behind the scenes" use of data mining techniques [5]. As an example, Culnan [6] made a particular study of secondary information use, which she defined as "the use of personal information for the other purposes subsequent to the original transaction between an individual and an organization when the information was collected." The key finding of this study was that concern over secondary use was correlated with the level of control the individual has over the secondary use.

Even though many nations have developed privacy protection laws and regulations to guard against private use of personal information, the existing laws and their conceptual foundations have become outdated because of changes in technology [7, 8, 9, 10]. As a result, these personal data reside on thousands of file servers, largely beyond the control of existing privacy laws, leading to potential privacy invasion on a scale never before possible.

Complex issues, such as those involved in privacy-preserving data mining, cannot simply be addressed by restricting data collection or even by restricting the secondary use of information technology [11, 12, 13]. Moreover, there is no exact solution that resolves privacy preservation in data mining. An approximate solution could be sucient, depending on the application since the appropriate level of privacy can be interpreted in different contexts [14, 15]. In some applications (e.g., association rules, classification, or clustering), an appropriate balance between a need for privacy and knowledge discovery should be found.

2 Data Mining

Data mining is an innovative way of gaining new and valuable business insights by analyzing the information held in your company database. These insights can enable you to identify market niches, and they support and facilitate the making of well-informed business decisions. Essentially, data mining is a ground-breaking way to leverage the information that your company already has in order to plan a business strategy for the future.

- *Syed Md. Tarique Ahmad is currently pursuing Ph.d in Computer Science in Pacific University Udaipur Rajasthan, India, E-mail: tariquemca1@gmail.com*
- *Shameemul Haque is currently working with Dept. of Computer Science, Veer Kunwar Singh University Ara, Bihar, India, E-mail: shameem32123@gmail.com*
- *SM Faizanut Tauhid is currently working with Dept. of Computer Science, Pacific University Udaipur Rajasthan, India. E-mail: tauhidfaiz@gmail.com*

Data mining uncovers this in-depth business intelligence by using advanced analytical and modeling techniques. With data mining, you can ask far more sophisticated questions of your data than you can with conventional querying methods. The information that data mining provides can lead to an immense improvement in the quality and dependability of business decision making.

Conventional methods can tell a bank, for example, which of the bank account types that it provides is the most profitable. However, only data mining enables the bank to create profiles of the customers who already have this type of account. The bank can then use data mining to find other customers who match that profile, so that it can accurately target a marketing campaign to them.

Data mining can identify patterns in company data, for example, in records of supermarket purchases. If, for example, customers buy product A and product B, which product C are they most likely to buy as well? Accurate answers to questions like these are invaluable aids to marketing strategies.

Data mining can identify the characteristics of a known group of customers, for example, those who have a proven record as poor credit risks. The company can then use these characteristics to screen new customers and to predict if they also will be poor credit risks.

Data mining tools ease and automate the process of discovering this kind of information from large stores of data.

3 Data Mining Process

Data mining is an iterative process that typically involves the following phases:

3.1 Problem definition

A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition.

3.2 Data exploration

Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital.

In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

3.3 Data preparation

Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value.

In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

3.4 Modeling

Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model.

In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required.

The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

3.5 Evaluation

Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions:

I. Does the model achieve the business objective?
II. Have all business issues been considered?

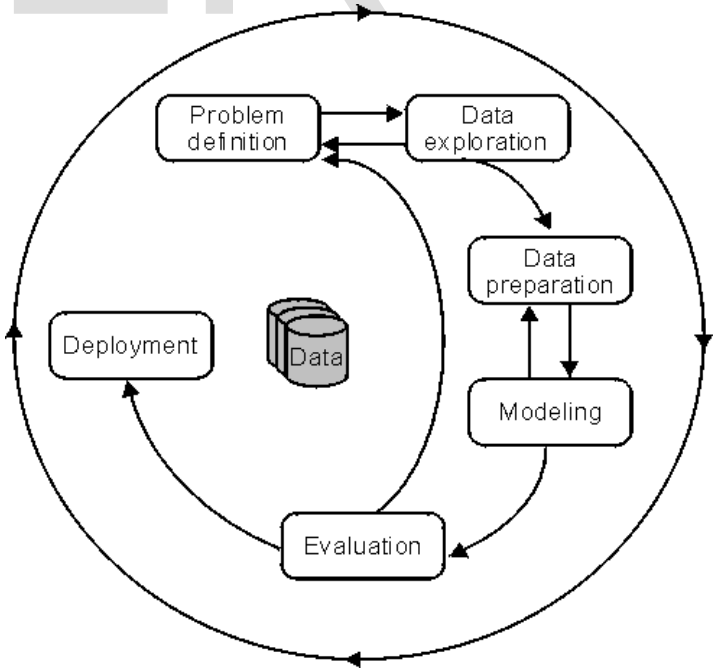
At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

3.6 Deployment

Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets.

Given below figure shows the Data Mining Process

Fig. 1. Data Mining Process Model



4 Major Task Of Data Mining

The tasks of data mining are very diverse and distinct because there are many patterns in a large database. Different kinds of

methods and techniques are needed to find different kinds of patterns. Based on the kinds of patterns we are looking for, tasks in data mining can be classified into summarization, classification, clustering, association, and trend analysis [16, 17].

4.1 Summarization

Summarization is the abstraction or generalization of data. A set of task-relevant data is summarized and abstracted, resulting a smaller set which gives a general overview of the data and usually with aggregation information. For example, the long distance calls of a costumer can be summarized into totabninutes, total_spending, total_calls, etc. Such high-level, summary information, instead of detailed calls, is presented to the sales managers for costumer analysis.

The summarization can go up to different abstraction levels and can be viewed from different angles. For example, the calling minutes and spending can be totaled along the calling period in weeks, months, quarters, or years. Similarly, the calls can be summarized into in-state calls, state-to-state calls, Asia calls, Europe calls, etc., which can be further summarized into domestic calls and international calls. Different com-binations of abstraction levels and dimensions reveal various kinds of patterns and regularities.

4.2 Classification

Classification is the derivation of a function or model which determines the class of an object based on its attributes. A set of objects is given as the training set in which every object is represented by a vector of attributes along with its class. A classification function or model is constructed by analyzing the relationship between the attributes and the classes of the objects in the training set. Such a classification function or model can be used to classify future objects and develop a better understanding of the classes of the objects in the database.

For example, from a set of diagnosed patients, who serve as the training set, a classifi-cation model can be built, which concludes a patient's disease from his/her diagnostic data. The classification model can be used to diagnose a new patient's disease based on the patient's diagnostic data, such as age, sex, weight, temperature, blood pressure, etc.

4.3 Association

Association is the discovery of togetherness or connection of objects. Such kind of togetherness or connection is termed as association rule. An association rule reveals the associative relationships among objects. i.e. the appearence of a set of objects in a database is Strongly related to the appearance of another set of objects. For eg. in a telecommunication database. an association rule that "call waiting" is associated with "call display", denoted as "call waiting --> call display", says if a customer subscribes to the "call waiting" service, he or she very likely also has "call display".

The association rules can be useful for marketing, commodity management, advertising, etc. For eg. a retail store may discover that people tend to buy soft drinks together with potato chips, and then put the potato chips on sale to promote the sale of soft drinks.

4.4 Clustering

Clustering is the identification of classes, also called clusters or groups, for a set of objects whose classes are uknown. The objected are so clustered that the intraclass similarities are maximized and the interclass similarities are minimized based on some criteria defined on the attributes of objects. Once the clusters are decided, the objects are labeled with their corresponding clusters, and common features of the objects in a cluster are summarized to form the class description.

For eg. a bank may cluster its customers into several groups based on the similarities of their age, income, residence, etc., and the common characteristics of the customers in a group can be used to describe that group of customers. The clusters will help the bank to understand its costumers better and thus provide more suitable products and customized services.

4.4 Trend analysis

A lot of data available now are time series data that are accumulated over time. For example, a company's sales, a costumer's credit card transactions, and stock prices, are all time series data. Such kind of data can be viewed as objects with an attribute time, and the objects are the snapshots of entities with values that changes over time. It is interesting to find the patterns and regularities in the data evolutions along the dimension of time.

Trend analysis discovers interesting patterns in the evolution history of the objects. One topic in trend analysis is the identification of patterns in an object's evolution, such as up, down, peak, valley, etc. A model or function is constructed to simulate the behaviors of the object, which can be used to predict the future behaviors. For example, we can estimate this year's profit of a company from its last year's profit and the estimated annual increasing rate.

Another topic in trend analysis is the matching of the objects changing trends, such as increasing streaks, decreasing streaks, etc. By comparing two or more objects historical changing curves or tracks, similar and dissimilar trends can be discovered which will help us to understand the behaviors of the objects. For example, a company's sales and profit figures can be analyzed to find the disagreeing trends and search for the reasons behind such disagreements.

5 LITERATURE SURVEY

The study of Privacy-Preserving Data Mining techniques started extensively since 2000 [18], covering development approximately in two categories: Perturbation-Base technique [18, 19] and Secure Multi-Party Computation Base technique [20, 21]. The main idea of Perturbation-Based technique involves increasing a noise in the raw data in order to perturb the original data distribution and to preserve the content of hidden raw data. Geometric Data Transformation Methods (GDTMs) [22] is one simple and typical example of data perturbation technique, which perturbs numeric data with confidential attributes in cluster mining in order to preserve privacy. Nonetheless Kumari et al. [19] proposed a privacypreserving clustering technique of Fuzzy Sets, transforming confidential attributes into fuzzy items in order to preserve privacy.

Furthermore, the largest issue encountered when implementing a perturbation technique is the inaccurate mining result from a perturbed data. In view of this issue, the technique of Random-data perturbation introduced by Agrawal and Skrikant [18] was the first study addressed. Whereas the technique derives the original data distribution using a random noise for data distribution, and constructs a result similar to the original data, it finally use this similar result to execute mining. This method could construct a more accurate data mining model, while reducing mining errors. In addition, usually the perturbation technique that has higher privacy preservation comes with a lower level of mining accuracy, whereas most of the perturbation techniques today belong to the one-size-fits-all and are relatively inflexible. To resolve this issue, Liu and Thuraisingham [23] developed the two-phase perturbation technique which frames different intervals according to different user demand, and directly obtain sample data from a specific interval to derive the original data distribution. In the study on Secure Multi-Party Computation Base technique, Vaidya and Clifton [24] proposed the method of privacy preserving clustering technique over vertically partitioning data, whereas data with different attributes and different locations are considered as the same data set, all data could perform K-means under preserving privacy. On the contrary, Meregu and Ghosh [21] proposed the method of privacy preserving cluster mining over horizontally data partitioning, whereas it is framework of "Privacy-preserving Distributed Clustering using Generative Model." In this framework, each data independently owns an individual source, using local data to train generative models, and delivers model parameters to the central combiner responsible for model integration, hence avoiding direct contact between data source and combiner in order to accomplish privacy preserving through this method.

Among the cluster mining algorithms, K-means is one of the most popular and well-know methods mainly due to its simple concept, easy implementation and comprehensible mining result. Although the method has its own drawbacks [25], most of the existing data stream clustering algorithm are nonetheless developed based on studies of this method. In literature [26, 27], a machine learning algorithm names, Very Fast machine Leaning (VFML) has been proposed, whereas this method depends on determining an upper boundary to be applied as data items test in each step of the algorithm. Subsequently, Very Fast K-Means (VFML) clustering and Very Fast Decision Tree (VFDT) classification techniques have been developed based on the concept of VFML, and applied on the data stream of artificial and real network. On the other hand, Ordonez [28] developed an incremental K-means algorithm to improve the problems of clustering binary data streams with Kmeans. Incremental K-means not only real-time processing and artificial datasets, but simplification of data processing for binary data could also eliminate the need for data normalization. The concept of this algorithm is based on the updating cluster center and weight immediately following examining a batch of data, in order to perform fast clustering. Furthermore, Aggarwal et al. [29] proposed another CluStream which is applicable in data stream clustering, using summarized statistical information of data streams to cluster according to the user desired cluster numbers. On the other hand, Gaber et al. [30] has developed a Lightweight Clustering algorithm to handle high speed data stream. This algorithm is based on the concept of Algorithm Output Granularity, which is mainly used to adjust the minimal boundary value of distance among datasets representing different clusters, then controls the output-input ratio according to available resources, and to output a combined clustering result when the memory space is full. More recently, Yang and Zhou [31] further developed an HCluStream data stream clustering algorithm which processes combined attributes based on CluStream algorithm in order to solve the weakness of inability to perform non-numerical data mining by CluStream.

The solutions presented in [32, 33] aim at mining globally valid results from distributed data without revealing information that compromises the privacy of the individual sources. In particular, the work in [32] addresses secure mining of association rules over horizontally partitioned data. This approach considers the discovery of associations in transactions that are split across sites, without revealing the contents of individual transactions. In this model, the data available in all parties have the same schema, and it is assumed that three or more parties are involved to minimize the leakage of information. The solution is based on secure multi-party computation to minimize the information shared, while adding overhead to the mining task. On the other hand, the work in [33], addresses the problem of association rule mining in which transactions are distributed across sources. Each site holds some attributes of each transaction, and the sites wish to collaborate to identify globally valid association rules. In this model, two parties are involved, one party being designated as the primary, which is the initiator of the protocol. The other party is the responder. There is a join key present in both databases. The goal is to and association rules involving attributes other than the join key.

6 CONCLUSION

As most previous studies on privacy-preserving data mining placed specific importance on the security of massive amounts of data from a static database, consequently data undergoing privacy-preservation often leads to a decline in the accuracy of mining results. Furthermore, following by the rapid advancement of Internet and telecommunication technology, subsequently data types have transformed from traditional static data into data streams with consecutive, rapid, temporal, and unpredictable properties. Due to the increase of such data types, traditional privacy-preserving data mining algorithms requiring complex calculation are no longer applicable. The major objective of the review paper is to throw some light on the data transformation for privacy preserving data mining previous work. This article also discussed the various researches related to data transformation for privacy preserving data mining.

REFERENCES

[1] L. Brankovic and V. Estivill-Castro. Privacy Issues in Knowledge Discovery and Data Mining. In Proc. of Australian Institute of Computer Ethics Conference (AICEC99), Melbourne, Victoria, Australia, July 1999.

[2] U. Fayyad. Knowledge Discovery in Databases: An Overview, Chapter 2 of "Relational Data Mining", S. Dezeroski and N. Lavrac (eds.), Springer-Verlag, Germany, 2001, pages 28-47.

[3] M. Kantarcioglu, J. Jin, and C. Clifton. "When Do Data Mining Results Violate Privacy?", In Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 599-604, Seattle, WA, USA, August 2004.

[4] J. Vaidya and C. Clifton. "Privacy-Preserving K-Means Clustering Over Vertically Partitioned Data.", In Proc. of the 9th ACM SIGKDD Intl. Conf. on Knowlegde Discovery and Data Mining, pages 206-215, Washington, DC, USA, August 2003.

[5] G. H. John. "Behind-the-Scenes Data Mining." Newletter of ACM SIG on KDDM, 1(1):9-11, June 1999.

[6] M. J. Culnan. "How Did They Get My Name?: An Exploratory Investigation of Consumer Attitudes Toward Secondary Information.", MIS Quartely, 17(3):341-363, September 1993.

[7] K. C. Laudon. "Markets and Privacy.", Communication of the ACM, 39(9):92-104, September 1996.

[8] D. E. O'Leary. "Some Privacy Issues in Knowledge Discovery: The OECD Personal Privacy Guidelines", IEEE EXPERT, 10(2):48-52, April 1995.

[9] V. Estivill-Castro, L. Brankovic, and D. L. Dowe. "Privacy in Data Mining", Privacy Law and Policy Reporter, 6(3):33-35, September 1999.

[10] S. Cockcroft and P. Clutterbuck. "Attitudes towards Information Privacy", In Proc. of the 12th Australasian Conference on Information Systems, Coffs Harbour, NSW, Australia, December 2001.

[11] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. "Hippocratic Databas es", In Proc. Of the 28th Conference on Very Large Data Bases, Hong Kong, China, August 2002.

[12] L. Brankovic and V. Estivill-Castro. "Privacy Issues in Knowledge Discovery and Data Mining", In Proc. of Australian Institute of Computer Ethics Conference (AICEC99), Melbourne, Victoria, Australia, July 1999.

[13] S. R. M. Oliveira and O. R. Zaiane. "Foundations for an Access Control Model for Privacy Preservation in Multi-Relational Association Rule Mining", In Proc. of the IEEE ICDM Workshop on Privacy, Security, and Data Mining, pages 19-26, Maebashi City, Japan, December 2002.

[14] C. Clifton, W. Du, M. Atallah, M. Kantarcioglu, X. Lin, and J. Vaidya. "Distributed Data Mining to Protect Information Privacy", Proposal to the National Science Foundation, December 2001.

[15] C. Clifton. "Using Sample Size to Limit Exposure to Data Mining", Journal of Computer Security, 8(4):281-307, November 2000.

[16] G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. "From data mining to knowledge discov-ery: An overview", In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, pages 1-35. AAAI/MIT Press, 1996.

[17] M. S. Chen, J. Han, and P. S. Yu. "Data mining: An overview from a database perspective", IEEE Transactions on Knowledge and Data Engineering, 8:866-883, 1996.

[18] Agrawal, R. and Srikant, R., "Privacy-Preserving Data Mining,", Proceeding of the ACM SIGMOD Conference on Management of Data, Dallas, Texas, U.S.A., pp. 439-450 (2000).

[19] Kumari, P. K., Raju, K. and Rao, S. S., "Privacy Preserving in Clustering Using Fuzzy Sets," Proceedings of the 2006 International Conference on Data Mining, Las Vegas, Nevada, U.S.A., pp. 26□29 (2006).

[20] Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X. and Zhu, M. Y., "Tools for Privacy Preserving Distributed Data Mining," ACM SIGKDD Explorations Newsletter, Vol. 4, pp. 28□34 (2002).

[21] Meregu, S. and Ghosh, J., "Privacy-Preserving Distributed Clustering Using Generative Models," Proceedings of the 3th IEEE International Conference on Data Mining, Melbourne, Florida, U.S.A., pp. 211-218 (2003).

[22] Oliveira, S. R. M. and Zaïane, O. R., "Privacy Preserving Clustering by Data Transformation," Proceedings of the 18th Brazilian Symposium on Databases, Manaus, Brazil, pp. 304□318 (2003).

[23] Liu, L. and Thuraisingham, B., "The Applicability of the Perturbation Model-Based Privacy Preserving Data Mining for Real-World Data," Proceedings of the 6th IEEE International Conference on Data Mining, Hong Kong, China, pp. 507□512 (2006).

[24] Vaidya, J. and Clifton, C., "Privacy-Preserving KMeans Clustering over Vertically Partitioned Data," Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C., U.S.A., pp. 206□215 (2003).

[25] Chen, T. S., Lin, C. C., Chiu, Y. H. and Chen, R. C., "Combined Density-Based and Constraint-Based Algorithm for Clustering," Journal of Donghua University, Vol. 23, pp. 36□38 (2006).

[26] Hulten, G., Spencer, L. and Domingos, P., "Mining Time-Changing Data Streams," Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, U.S.A., pp. 97□106 (2001).

[27] Domingos, P. and Hulten, G., "Mining High-Speed Data Streams," Proceedings of the Association for Computing Machinery 6th International Conference on Knowledge Discovery and Data Mining, Boston, U.S.A., pp. 71□80 (2000).

[28] Ordonez, C., "Clustering Binary Data Streams with K-means," Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, San Diego, California, U.S.A., pp. 12□19 (2003).

[29] Aggarwal, C., Han, J., Wang, J. and Yu, P. S., "A Framework for Clustering Evolving Data Streams," Proceedings of the 29th International Conference on Very Large Data Bases, Berlin, Germany, pp. 81-92 (2003).

[30] Gaber, M. M., Krishnaswamy, S. and Zaslavsky, A., "On-Board Mining of Data Streams in Sensor Networks," Springer, Berlin Heidelberg, Germany, pp. 307-335 (2005).

[31] Yang, C. and Zhou, J., "HClustream: A Novel Approach for Clustering Evolving Heterogeneous Data Stream," Proceedings of the 6th IEEE International Conference on Data Mining, Hong Kong, China, pp. 682-688 (2006).

[32] M. Kantarcioglu and C. Clifton. "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", In Proc. of The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Madison, Wisconsin, June 2002.

[33] J. Vaidya and C. Clifton. "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", In Proc. of the 8th ACM SIGKDD Intl. Conf. on Knowlegde Discovery and Data Mining, pages 639-644, Edmonton, AB, Canada, July 2002.