PPDM Model dependent Bayesian Network for XML Association Rules Mining

Khalid Iqbal[1], Sohail Asghar[2], Simon Fong[3]
[1]Department of Computer Science
SZABIST, Islamabad
Pakistan
[2]Department of Computer Science
Mohammad Ali Jinnah University
Islamabad, Pakistan
[3]Department of Computer and Information Science
University of Macau
Macau SAR
mykhalidiqbal@yahoo.com, sohail.asghar@Jinnah.edu.pk, ccfong@umac.mo

ABSTRACT: *In Association Rule Mining (ARM) rules are the core which determine the mining process and its effectiveness. In the ARM, a central issue is the sensitivity which is either ignored or not addressed by the researchers in data mining. It is important to avoid senstive information disclosure in ARM. We propose to use Bayesian Network which is dependent on PPDM model and can reliably hide away sensitive rules in ARM. Additionally we advoate that the XML domain of PPDM can be reinforced with the use of sensitivity. We document clearly the efficacy of PPDM model with empirical validity in the current study. We also expect that the future research will address more issues of ARM based on PPDM model.*

Keywords: XML domains, Bayesian Network, Association Rule Mining, K2 Algorithm, PPDM Model

Received: 25 May 2011, Revised 3 July 2011, Accepted 6 July 2011

1. Introduction

Large number of recent studies in Data Mining use massive datasets to unearth knowledge in terms of trends and patterns analysis. For doing this exercise, a number of algorithms are formulated, such as classification, clustering and association rules mining for extracting significant patterns. Association rule mining (ARM) has its unique advantage in picking up rules that reveal strong association of items from complex datasets. In such analysis, user defines arbitrary minimum values for *c*% (confidence) and *s*% (support) for thresholds. Such threshold values are used to control how many transactional items from the original data source (*D*) which are represented as association rules that we would be interested in. The representation of rule can take a form such as $X \rightarrow Y$, where *X* and *Y* are the antecedent and consequent respectively. The problem in such presentation is the potential risk of disclosing the sensitive information to a third party while sharing data sources because the identities of both *X* and *Y* are clearly revealed. Consequently, Privacy Preserving in Data Mining emerges as a popular research topic, and it has a large relevancy to association rules mining.

In association rule mining, a question arises - how to identify the sensitive items in the original data source? For this purpose, researchers have proposed various methodologies such as in [1, 2]. Unfortunately, these techniques are unable to identify the sensitive item(s) based on just the antecedent or consequent of the rules after they are statistically transformed from the amount of transactional records from the original data source. In such transformation, another question rises - how many item(s) should be considered as reliable, for declaring them to be sensitive hence a recommendation/justification for a modification of database should be made? These two questions pertaining to reliability and privacy hiding need to be addressed even if the problem is NP-hard [4] with the minimal effect to the database.

To address the above-mentioned issues in ARM, Bayesian Network (BN) is attempted by the authors in handling uncertain situations of the ARM rules for achieving a high degree of reliability. This is technical possible because from literature, K2 algorithm [10] was used to generate BN from XML document. This document can also be used to generate XML association rules with the help of Apriori algorithm [11]. By using K2 algorithm, we record the occurrences of the transactional items according to their dependencies on each other, and then we compute the mode (the most frequent item(s)). Such item(s) can be used to modify the most frequent transactions of the original data source. The modified data source *D'* is later used for minimizing the disclosure risk involved in association rule mining. Our proposed method is verified to be feasible with testing data, in hiding sensitive association rules.

The structure of the remaining paper is as follows. Section II provides a literature review over PPDM, XARs and BN which are the core techniques used in our method. Section III presents the proposed PPDM model while Experimental Results are shown in Section IV. Finally, section V concludes and ponders on the future work.

2. Literature Review

Digital data has been increasing enormously. Therefore, privacy issues are considered by the researchers in a wide variety of domains. Association Rule Mining as a popular topic in data mining needs especially to be addressed with the privacy issues. In such association rule mining process from large databases, there is a high danger of revealing valuable information to external parties. In this paper, we focus on the privacy preservation in Association Rule Mining in order to reveal non-obvious information from data sources. As such, the literature is presented mainly on Privacy Preservation in Data Mining as below.

Weng et al. proposed an efficient algorithm for fast hiding sensitive association rules named as FHSAR [1]. This algorithm considers a database *D* and SAR (Sensitive Association Rules) with minimum support and minimum confidence to hide. For this purpose, the FHSAR generates a released database *D'*. Afterwards, *D'* is used to generate rules. Consequently, sensitive rules are entirely hidden with the minimized side effects.

The implementation of this algorithm is carried out in two stages. In first stage, *FHSAR* scans the database *D* one time and it gathers correlated information of the transactions and sensitive rules. This correlation is represented by a graph *G* [1]. Moreover, a transaction t_1 has i_k items and can be represented as $< R_k , |R_k| >$; where $R_k = \{ j \mid SAR_i \subseteq t_1, i_k \in SAR_j \}$. In addition to t_i (transactions), a prior weight w_i is associated with each edge (*u*, *v*) as heuristic to estimate side effects. It can be computed by a simple formula such as $w_i = MIC_i / 2^{(|ti| = 1)}$; where $MIC_i = \max (|R_k|)$ [1]. In the second stage, transactions (t_i) are modified one after another until the entire set of sensitive rules is hidden. This modification of transactions is carried out according to the prior weight (w_i) associated with the transactions. Therefore,the proposed steps are repeated until $SAR = \Phi$. These steps include selection of transactions, deletion of items, function for checking-and-removing item for the sake of avoiding hidden rules, computation of w_k for modified item and placing it in PWT (Prior Weighted Transaction), modifying $||SAR_j||$ and $||L(SAR_j)||$ and removal of SAR_j from SAR [1]. Thus, their proposed algorithm contributes in hiding the entire sensitive rules with minimum side effects. The main strength of FHSAR is the outclass performance over the other techniques. Moreover, it modifies the original data source into a release database with minimum side effects. The limitation of the proposed algorithm is over-hiding a significant number of nonsensitive rules.

To restrict the sensitivity issue in quantitative analysis data using association rules, Krishna et al. proposed a novel method to mine statistical and fuzzy association rules from quantitative data [2]. Before the use of Apriori algorithm [11] on the data, they are booleanized by changing the quantity of an attribute to 1. Also for the quantity of zero for attribute is taken as 0. Thus, these booleanized data are passed to the Apriori algorithm [11, 12] to generate BARs (Booleanized Association Rules). Later on, SARs (Statistical Association Rules) and FARs (Fuzzy Association Rules) are generated from the quantitative data using other relationship measures. The measures used in place of support and confidence for SARs are Mean ($\overline{X}$) and Standard Deviation (SD or σ). These measures are calculated with the following formulas in the equations below [2].

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow (2.3); \qquad \sigma = \sqrt{\frac{1}{n}(X_i - \mu)^2} \rightarrow (2.4)$$

Let *D* be a database that contains set of attributes {*A*, *B*, *C*, … , *P*} and n transactions. Therefore, *D* presents paired entries based on attributes and transactions over a set of real numbers *R*. Hence, the SARs can be presented in the form of association rules as $A(\overline{X}_{1a}, \sigma_{1a}), B(\overline{X}_{1b}, \sigma_{1b}) \Rightarrow C(\overline{X}_{1c}, \sigma_{1c})$ where $\overline{X}$ and σ represent the Mean and Standard deviation respectively, and *A*, *B*, *C* are attributes in *D*. FARs is then mined from the quantitative database using fuzzy logic. Fuzzy logic suggests the belonging of an element to a membership value of either 1 or 0. The membership value is assigned by the membership function as $mf_x(x)$: $D \rightarrow [0,1]$[2]. Moreover, FAR is presented in pairwise form such as <*attribute*, *linguistic term*> which is useful for the assessment and understanding by users. Thus the membership function can be built with fuzzy sets that are comprehensible, such as Low, Medium, High and Very High. Hence, FAR is presented in the form *A*(*Low /* 0.8, *Medium /* 0.2), *B*(*Medium / 0.7, High /* 0 .3) ⇒ *C*(*High /* 0 .1, *VeryH igh /* 0 .9)[2]. This rule can be interpreted in a simplier form such as *A* (*Low* / 0.8), *B*(*Medium* / 0.7) ⇒ *C*(*VeryHigh* /0 .9) . Thus the theoretical interpretation is based on higher values.

Consequently, the authors converted quantified data into BARs and generated association rules using Apriori algorithm [11, 12]. Furthermore, the quantified database is used for the generation of SARs and FARs using the commodity dataset and the results are compared with the BARs. The main strength of the proposed approach is to describe the behavior of each attribute in the form of association rules. Additionally the clustering technique dependency is removed by using cross validation to cluster data in an optimal and automated way. Moreover, dissimilarity among values of clusters is calculated with the use of coefficient variation which is the ratio of σ and *x*. In spite of reasonable benefits, the measures used in the proposed approach are highly influenced either from the very lowor very high value of commodities. Furthermore, the proposed method does not hide the sensitive data and display all the patterns regardless of interesting or not. Finally, the output results are not very easy to interpret especially when the values of the measures in the rule are chosen to be very high or very low.

With the new metric introduction for not to disclose sensitive information, Saygin et al. [3] presented algorithms for privacy preserving by demonstration of security issues related to association rules. They extended their approach by making some modifications in the original dataset. This modification is carried out with an introduction of new symbol "?" mark. The question mark in the transactions neither represents the presence nor the absence of an item. With this modification in the original dataset, support and confidence are also affected and modified definitions are presented. For this purpose, the support for an itemset is taken in interval form such as [*minSupport* (*A*), *maxSupport* (*A*)] [3]. Moreover, the confidence interval is presented as [*minConfidence*(*antecedent consequent*), *maxConfidence* (*antecedent*→ *consequent*)]. The following equations (2.5) and (2.6) represent the *minConfidence*(*antecedent*→ *consequent*), *maxConfidence* (*antecedent* → *consequent*).

$$\mathit{minconfidence}\,(A \Rightarrow B) = \min \mathit{Support}\,(A \cap B) \mathrm{X}100/ \max \mathit{Support}\,(A) \rightarrow (2.5)$$
$$\mathit{maxconfidence}\,(A \Rightarrow B) = \max \mathit{Support}\,(A \cap B) \mathrm{X}100/ \min \mathit{Support}\,(A) \rightarrow (2.6)$$

From the above equations, they noted some interesting properties: *minConfidence* (*A* ⇒ *B*) = *maxconfidence* (*A* ⇒ *B*) and min *Support* (*A* ∩ *B*) = max *Support* (*A* ∩ *B*). The problem with the placement of "?" mark in sanitization process is the deviation from the Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT) values which increase the degree of uncertainty of the rule called as Safety Margin (SM). Therefore, the introduction of "?" mark in place of 1 does not reduce the support and leads to confidence reduction. This reduction is made by placing "?" in place of 0 in an interleaved fashion in the process of hiding association rules. Consequently, half of the sensitive association rules are hidden with the reduction of support and the rest of the half is hidden with confidence reduction.

As most of the literature do not provide an accurate solution for addressing sensitivity of association rules, Atallah et al. [4] focused on association rules *R* mined from the source database *D*. The problem focused is how to transform a database *D* into released database *D'*. Moreover, hidden rules (R_h) can be observed through the modified database by reducing the support of rules. Thus, transformation is referred to as sanitization of *D* from which the knowledge is preserved from disclosing to the public eyes. Such optimal transformation is of a NP-hard problem [4]. To solve this problem approximately, let *A* and *B* be the two itemsets which are "good" and "bad" respectively [4]. From these itemsets, we are not interested in restricting the occurrence of *A* but we want to ensure little occurrence of *B*. Both occurrence and non-occurrence of itemsets are incompatible. Besides this incompatibility, NPhardness problem comprises of three Optimization Problems such as Problem 1, Problem 2 and Problem 3. To solve these problems, the HITTING SET instances are taken into account for each of the proposed problems [4]. Moreover, heuristic approach is suggested for optimal sanitization with illustration based on preliminary definitions and data structure required for algorithm are described [4].

The reliability estimation is carried out by Doguc et al. [5] who came up in literature with a generic approach to estimate the trustworthiness of a system. Such estimation of an approach in terms of reliability is presented by BN (Bayesian Network) Model. In such estimation, historical dataset is used based on the edges and nodes. Edges represent the relationship among nodes which is uncertain. For this purpose, Bayesian theorem is used through which an i^{th} node occurrence can be measured with a-priori occurrences of j^{th} nodes. Therefore K2 algorithm [5, 10] can be used to measure the reliability of occurrences of nodes at any position in an incremental fashion. Moreover, K2 algorithm [10] quantifies associations and ranks the parent set with a reduction of search space heuristically and with the use of scoring function. On the down side, however, the limitation of K2 algorithm [10] is the fixed order including the first node in order as parent which can benefit in mining process of preserving association rule privacy. This may reduce the n attributes set to $(n-1)^{th}$ attributes set.

Furthermore Bayesian Networks can be used in more complex environment for reliability analysis. In such case, Richiardi et al. [6] suggested an approach to measure the modality reliability information. Such reliability can be obtained through Bayesian Network by combining the acoustic environment as well as classifier behavior under noisy acoustic conditions. Consequently, the overall average accuracy and variability results show the effectiveness of measuring reliability through Bayesian Network.

Similarly to diagnose the abnormalities of a system, Doguc et al. [7] presented a SOE (System Operation Effectiveness) assessment through Bayesian Network model. This model has nodes which show sub-systems with a related dependency between them. With the evaluation of dependency, a problematic node/variable/item is identified for review. This problematic review of node/item/variable helps in re-designing of the problematic part of the system. In this way, the overall SOE effectiveness is improved.

In relation to Privacy Preservation, Vaidya et al. [8] applied a method to preserve information using Naïve Bayes Classifier for vertically partitioned data. In this data, the numbers of items/column are variable but the numbers of entities are the same. Such situation can be observed while sharing information between the insurance companies and the drug companies. Thus, to hide information from being revealed to others, model parameters are computed for the purpose of sharing and classified as new instance.

Similarly for preserving information, Wright et al. [9] presented a privacy-preserving protocol with the use of BN for the distributed heterogeneous data. In this case, confidential database can be shared by using the BN structure between the databases of the two parties. To construct a BN structure, K2 algorithm [10, 9] is used. The construction of BN structure produces joint data of the two parties. The production of joint data is computed by summing all the intermediate values in a cryptographic manner. In this way the proposed methodology becomes secured against passive adversaries. Despite cryptographic security, the suggested methodology does not offer any detail on the irrelevant data as well as the loss of data. So there is no easy way to understand the generated reliable output. For this reason, the output may not give the relevant results for which both parties share their databases.

3. Proposed PPDM Model

Public concerns regarding privacy are on the rise when association rules are being generated especially from database that contains personal records. Privacy Preserving Data Mining (PPDM) has been studied extensively by researchers and practitioners mainly for upholding privacy and security. One way is to counter the interference from data mining by hiding sensitive information from the data. That is, by modifying a database and the output representations of association rules. Here we focus on data in XML domain that are common in private and public organizations, and address security issues in the context of XML association rules. We propose a PPDM model as shown in Figure 1. Although various table text styles are provided, the formatter will need to create these components, incorporating the applicable criteria that follow.

The steps on how our proposed model function are described as follow:

1. Read XML Document. The document all the transactions in XML format.

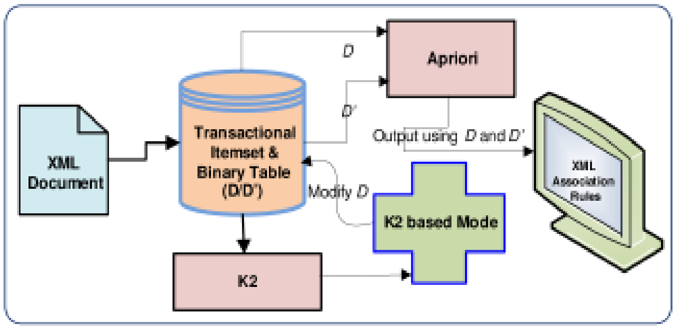

Figure 1. Proposed Mode based PPDM Model

2. Form transactional itemset and binary table from the inputted document. Transactional Symbolized Items are a group of symbolized items that forms a transaction based on XML document items. Binary Table of Transaction is a table containing 1's and 0's to represent the presence or absence of an item in a transaction.

3. Apply Apriori algorithm on the transactional itemset from D to generate association rules. Apriori algorithm is suggested in [11] for effectively generating XML association rules after preprocessing in step 2.

4. From the binary table of transactional itemset, use K2 to generate a Bayesian Network. BN and K2 Algorithm produce a useful graphical model that trains and displays interesting relationship among nodes in a probabilistic manner. The combined usage of BN and K2 is suggested by [10].

5. Item# column is read to identify Mode using Conditional Probability Table (CPT). This table contains items and their conditional probabilities according to their dependency in Bayesian Network.

6. Modify the transactional itemset based on Mode that is obtained in step 5. From CPT, the most frequent item(s) are identified for the modification of transactions. This kind of frequent item identification is called Mode.

7. Apply Apriori algorithm again on the modified transactional itemsets. Then output the results in XARs.

4. Experiment

In the proposed PPDM model, XML document is formed from the primary-tumor dataset [12]. From this dataset, a sample of 15 transactions with 14 variables of different diseases including sex is taken in the XML document. The reason for taking a small number of transactions is to help researchers to easily observe compact results with a concise dataset. Therefore, this would also be helpful while performing the comparison of their research methodologies. Table 1 shows the experimental dataset with variables (first row), alphabetical order of variables (second row) and numerical order corresponding to second row. The purpose of such arrangement order is to maintain the consistency among the analysis of dataset after passing to the Apriori algorithm [11] and K2 algorithm [10]. In this way, the same order gives more accuracy and reliability in the generated results.

The dataset in Table 1 is being used to generate the Bayesian Network (BN). BN is generated by using the original data source (Table 1) and a snapshot of the resultant BN is shown in Figure 2. The BN shows dependency of items/variables for which an item/variable has the maximum probability in an incremental fashion with the given order. This order limits the generalization of K2 algorithm [10]. But despite this, K2 gives great reliability because of the computation of probabilities which enhances the accuracy.

Besides BN, mode is computed based on BN. In this case, 'N' is computed as mode because this item/variable/node was the most frequent one. So, this item is being modified in the original data source. From the primary-tumour dataset as in Table 1, XML Association Rules are generated from Original Data Source as shown in Table 2.

sex	bone	marrow	lung	pleura	m	liver	brain	skin	neck	cular	axillar	um	abdominal
L	C	D	G	K	J	F	E	M	I	N	B	H	A
2	3	4	7	1	0	6	5	3	9	4	2	8	1
0	0	0	0	0	0	1	0	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0	1	0	1	0
0	0	0	0	1	0	0	0	1	0	0	1	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	1	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	1	0	0	0
1	0	0	0	1	0	0	0	0	1	0	0	0	0
0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	0	0	0	0	0	0	1
1	1	0	0	0	0	0	0	1	1	1	0	0	0

Table 1. A Sample of Primary-Rumour Dataset

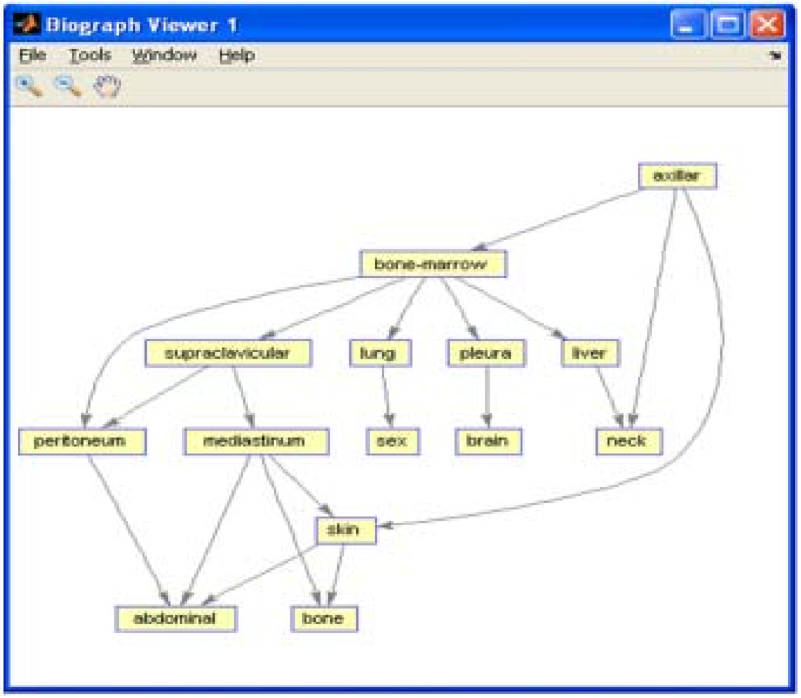

Figure 2. Bayesian Network formed from the dataset in Table 1

The above table presents rules that have some sensitive information which must be restricted from disclosure. Therefore, mode (computed from BN) is used to modify the largest transaction in Table 1 (extracted from the primarytumor dataset). This modified dataset is used to generate XML Association Rules using Apriori algorithm. This ensures to minimize the disclosure risk of sensitive rules as shown in Table 3.

With the use of Primary-tumor dataset [12], we pruned three rows of this dataset because data was not available in these rows. There were 18 variables in which only 14 were chosen for their relevancy as input. Hence, the overall hidden rules with no side effect (in terms of new rules, ghost rules) are shown based on 215 rows/transactions of the dataset in Table 4.

Rule #	XARs	s%	c%	LR
1	bone=>Sex	13.33	40	6.67
2	Liver=>peritoneum	13.33	40	8.00
3	Liver=>Sex	13.33	40	6.67
4	Liver=>Supraclavicular	13.33	40	10.00
5	Neck=>Sex	20	100	16.67
6	Neck=> Supraclavicular	13.33	66.67	16.67
7	Sex=> Supraclavicular	20	50	12.50
8	Neck, Sex=> Supraclavicular	13.33	66.67	16.67

Table 2. XML Association Rules From *D*

Rule #	XARs	s%	c%	LR
1	bone=>Sex	13.33	40	6.67
2	Liver=>peritoneum	13.33	40	8
3	Liver=>Sex	13.33	40	6.67
4	Liver=>Supraclavicular	13.33	40	10
5	Neck=>Sex	20	100	16.67
6	Neck=> Supraclavicular	13.33	33.33	8.33

Table 3. XML Association Rules Frome *D'*

5. Conclusion

We proposed a method for hiding sensitive XML association rules by using PPDM model and Bayesian Networks. Original data stored as a XML document is read automatically in the PPDM model and it is converted to structure transactional itemset and binary itemset. These structures are used for generation of association rule and BN by using Apriori and K2 algorithms. Our contribution in this paper is to compute the mode based on BN while declaring the item(s) as sensitive for modification of the D. This process enables us to put out of sight the identified sensitive XARs with reasonable reliability and accuracy. Furthermore, PPDM model modifies only one largest transaction in case of a single mode. Likewise, PPDM model also modifies the transaction which is the next largest in size even if two or more modes exist and declared as sensitive.
This method works well as we can just focus on the largest transaction; it will be identified and declared as sensitive, and subsequently get hidden away.

6. Acknowledgment

We acknowledge Lowel Guandi for K2 algorithm and Sadik Hava for Apriori implementation in Matlab.

Rule #	XARs	s%	c%	LR
14	Axillar => Brain	1.40	13.04	0.82
15	Axillar => Liver	1.40	13.04	0.18
77	Abdominal, Axillar => Liver	1.40	42.86	0.60
115	Axillar, Brain => Mediastinum	1.40	100	1.75
116	Axillar, Liver => Mediastinum	1.40	100	1.75
117	Axillar, Lung => Mediastinum	1.40	60	1.05
123	Axillar, Mediastinum => Sex	1.40	33.33	0.29
195	Abdominal, Axillar, Bone => Mediastinum	1.40	75	1.32
197	Abdominal, Axillar, Liver ==> Mediastinum	1.40	100	1.75

Table 4. Hidden XML Association Rules (Summary)

References

[1] Chih-Chia Weng, Shan-Tai Chen, Hung-Che Lo, (2008). A Novel Algorithm for Completely Hiding Sensitive Association Rules, Eighth International Conference on Intelligent Systems Design and Applications, 26-28 Nov., p.202-208.

[2] Vijay Krishna, G., Radha Krishna, P., (2008). A Novel Approach for Statistical and Fuzzy Association Rule Mining on Quantitative Data, *Journal of Scientific and Industrial Research,* 67 July, p. 512-517.

[3] Yucel Sayg1n, Vassilios S. Verykios, Ahmed K. Elmagarmid, (2002).Privacy Preserving Association Rule Mining, *In*: RIDE '02 Proceedings of the 12th International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems (RIDE 2002), IEEE Computer Society Washington, DC, USA, p.151-158.

[4] Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., Verykios, V. (1999). Disclosure Limitation of Sensitive Rules, *In*: Proceedings of the Workshop on Knowledge and Data Engineering Exchange, IEEE Computer Society Washington, DC, USA, p.45-52.

[5] Doguc, O., Ramirez-Marquez, J.E. (2009). A generic method for estimating system reliability using Bayesian networks, *Reliability Engineering & System Safety*, 94 (2) February, Elsevier, 542-550.

[6] Jonas Richiardi, Plamen Prodanov, Andrzej Drygajlo. (2005). A probabilistic measure of modality reliability in speaker verification, Winner of Best Student Paper Competition, IEEE ICASSP , p.709-712

[7] Ozge Doguc, Wei Jiang, (2005). A Bayesian Network (BN) Model for System Operational Effectiveness Assessment and Diagnosis, 26th ASEM National Conference Proceedings, October.

[8] Jaideep Vaidya, Chris Clifton. (2004). Privacy Preserving Naive Bayes Classifier for Vertically Partitioned Data, SDM, SIAM, p.522-526.

[9] Rebecca Wright, Zhiqiang Yang. (2004). Privacy Preserving Bayesian Network Structure Computation on Distributed Heterogeneous Data, KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, August 22–25, USA, p.713-718.

[10] Cooper Gregory, F., Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data, Machine Learning, 9 (4) Kluwer Academic Publishers, The Netheriands, p.309-347.

[11] Agralwal, R., Imielinski, T., Swami, A. (1993). Mining associations between sets of items in large databases, *In:* Buneman, P., and Jajodia, S. editors, SIGMOD93, Washington, D.C, USA, May, p.207-216.

[12] Primary-tumor dataset, http://mlearn.ics.uci.edu/databases/primarytum/[Accessed date: Aug. 2011]