American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-6, Issue-11, pp-135-154
www.ajer.org
Research Paper Open Access

Developing Suitable Methods for Obtaining Better Accuracy in Privacy Preserving Association Rules Mining and Classification

Dr.Venkata Reddy Medikonda[1], Dr.Vignesh Janarthanan[2], Prof.Er.Dr.G.Manoj Someswar[3]

[1]*M.Tech., Ph.D., Associate Professor, Sreyas Institute Of Engineering & Technology, Hyderabad, Telangana State, India.*
[2]*M.E., Ph.D., Professor, Sreyas Institute Of Engineering & Technology, Hyderabad, Telangana State, India.*
[3]*B.Tech., M.S.(USA), Ph.D., D.Phil., PDF, Research Professor & Scientist 'H', Global Research Academy – Scientific & Industrial Research Organisation [Autonomous], Hyderabad, Telangana State, India.*

Abstract: *This proposition is committed to protection saving characterization and affiliation rules mining over concentrated information twisted with randomisation-based techniques which change singular esteems aimlessly to give a normal level of security. It is expected that exclusive contorted esteems and parameters of a misshaping methodology are known amid the way toward building a classifier and mining affiliation rules.*
In this research paper, we have proposed the improvement MMASK, which takes out exponential intricacy of evaluating a unique help of an item set as for its cardinality, and, in result, makes the security saving disclosure of continuous item sets and, by this, association rules plausible. It likewise empowers each estimation of every credit to have distinctive bending parameters. We demonstrated tentatively that the proposed improvement expanded the exactness of the outcomes for abnormal state of security. We have additionally displayed how to utilize the randomisation for both ordinal and number ascribes to change their qualities as per the request of conceivable estimations of these credits to both keep up their unique area and acquire comparable circulation of estimations of a characteristic after bending. What's more, we have proposed security protecting strategies for arrangement in light of Emerging Patterns. Specifically, we have offered the energetic ePPCwEP and apathetic lPPCwEP classifiers as security protecting changes of enthusiastic CAEP and sluggish DeEPs classifiers, individually. We have connected meta-figuring out how to protection safeguarding arrangement. Have we utilized packing and boosting, as well as we have joined contrast meant likelihood conveyance of estimations of characteristics remaking calculations and recreation sorts for a choice tree keeping in mind the end goal to accomplish higher exactness of order. We have demonstrated tentatively that meta-learning gives higher precision pick up for security safeguarding classification than for undistorted information.
The arrangements displayed in this proposition were assessed and contrasted with the current ones. The proposed techniques got better exactness in protection saving affiliation rules mining and arrangement.
Additionally, we decreased the time of multifaceted nature for finding affiliation rules with protected security.
Keywords: *Privacy Preservation, Secure Multiparty Computations (SMC), Apriori-MMASK, Apriori-rMMASK, Relaxation, EMMASK*

Date of Submission: 04-11-2017 Date of acceptance: 17-11-2017

I. INTRODUCTION

Since security concerns identified with a conceivable abuse of learning found by methods for information mining strategies have been raised, many endeavours have been made to give protection safeguarding systems to information mining. Hence, another (sub) domain of information mining, protection saving information mining, rose in the most recent decade. With a specific end goal to give adequate privacy insurance in information mining, a few strategies for joining security have been created. Security itself isn't a simple term to characterize and can be saved on various levels in various situations. Regardless of huge decent variety in security parts of information mining, three fundamental methodologies can be recognized: heuristic-based, recreation based and cryptography-based.[1]

In the principal approach, the heuristic calculations are utilized to conceal learning an association does not have any desire to uncover, for example, singular esteems in information are changed by a heuristic calculation to shroud touchy information, for example, and critical guidelines on account of affiliation rules mining.

The remaking based approach is utilized to join security on an individual level by changing unique individual esteems (for example, clients' answers) arbitrarily by methods for a randomisation-based strategy and uncovering just adjusted esteems.[2] The contorted information and in addition parameters of a randomisation-based strategy used to mutilate them can be distributed or gone to an outsider. Knowing mutilated individual esteems and parameters of a randomisation-based strategy, one can perform information mining undertakings. To this end, first unique dispersions of estimations of traits are remade (assessed) in light of the contorted esteems and the parameters of the mutilation technique, and an information mining model is assembled in view of the misshaped information. The making of a model is completed without the need to get to unique individual information.

The third approach, which depends on cryptography, utilizes secure multiparty computations (SMC) to do information mining errands in view of appropriated information, that is, information controlled by various associations that would prefer not to uncover their private information. Besides, encryption strategies which empower one to perform calculations over scrambled information without having the capacity to decode can be utilized as a part of security saving. The heuristic approach is intended for incorporated information.[3] The cryptography-based approach is utilized for the circulated information, while the reproduction based approach can be connected to both disseminated and brought together information.

The execution is one of the benefits of the recreation based approach contrasted with cryptography-based arrangements. In any case, this method accomplishes it at the cost of precision of the outcomes.

As the remaking based approach utilizes individual randomized values, there is an exchange off between the level of security and acquired exactness of the information mining comes about. The more elevated amount of security is utilized, the higher loss of precision of the information mining comes about is gotten.

The cryptography-based strategies don't experience the ill effects of exactness misfortune, yet their fundamental disadvantage is elite cost, particularly when many gatherings are associated with the procedure. Both the reproduction based and cryptography-based methodologies can be utilized to safeguard security on an individual level. The last expect that there are numerous data suppliers which need to communicate to play out an information mining assignment. Subsequently, each time an information mining errand is per-framed, data suppliers ought to be prepared to communicate.[4] The recreation based approach is intended to be utilized for both conveyed and brought together information. The misshaped records can be put away in an incorporated database and can be utilized ordinarily in a procedure of building diverse models without co-operations with data suppliers. Moreover, the reproduction based approach empowers associations to give information that one can use to find concealed learning from it without uncovering singular attributes of items.

Considering Internet reviews, which are exceptionally prominent these days and likely will be as yet prevalent later on, we may state that the reproduction based approach fits such applications better since, as opposed to the cryptography-based approach, information suppliers don't take part during the time spent building a model, i.e., they don't acquire the outcomes or halfway consequences of an information mining assignment. Clients' worries about a conceivable abuse of gave information can be decreased by methods for a randomisation-based technique, which misshapes unique information and stores just twisted esteems in a brought together database. The contortion technique, which alters singular estimations of the question, does not rely upon any data about different items, in this manner it can be performed at the protest site. On account of Internet reviews, the mutilation should be possible at a client machine, which makes a randomisation-based strategy simple to join into a web program.

In addition, in the reproduction based approach extra protests gave by new clients can be added to a concentrated database and an information mining model can be modified without the need to associate with other information suppliers. In this proposition, we examine the utilization of randomisation-based strategies and the reproduction based systems in creating security protecting information digging for brought together information. In this work, we concentrate on the accompanying way to deal with protection saving information mining: 1) information is twisted by methods for randomisation-based strategies to save security and is put away in a brought together database; 2) the reproduction based procedures are utilized to play out a mining assignment on put away contorted information, and, on account of characterization, on unique estimations of items to be arranged.[5]

The principle objective of this postulation is to discover answers for protection saving grouping and affiliation rules mining over concentrated information misshaped with randomisation-based strategies. The proposed arrangements should consider the accompanying:

— Fundamental sorts of properties, to be specific nonstop, whole number, ostensible,
and ordinal qualities,

— Abnormal state of security,
— time many-sided quality of security saving information mining undertakings,
— A level of exactness of the outcomes.

The principle challenge is to diminish or dispose of the loss of precision of the outcomes caused by applying the randomisation-based techniques. Second undertaking of this theory is to make security protecting information mining reasonable practically speaking from time intricacy perspective.

Statements of the Research Work

The following statements are made in this work:

1. In privacy preserving classification and association rules mining, continuous, integer, nominal, and ordinal attributes can be distorted by means of the randomisation-based methods according to attributes characteristics.
2. The accuracy loss which comes from incorporating privacy on an individual level by means of the randomisation-based methods for centralised data in privacy preserving classification and association rules mining can be reduced or eliminated.
3. The exponential complexity of estimating support of an item set with respect to its cardinality in the process of privacy preserving association rules mining over centralised data distorted by means of the randomisation-based methods can be reduced to the constant complexity.

Contribution

We made the following contribution:

1. Our proposed optimisations of support estimation of an item set in the process of privacy preserving association rules mining can substantially reduce the time complexity of this process and make it viable in practice.
2. We showed that the randomisation-based attribution of data can be applied for ordered
3. Attributes according to an order and domains of attributes.
4. We presented that privacy preserving classifiers based on (Jumping) Emerging Patterns can reduce the trade-off between privacy and accuracy. Both the eager and lazy approach to classification with Emerging Patterns can be used in an effective way.
5. We proposed how to use meta-learning to combine results obtained for different reconstruction types and algorithms of reconstructing an original distribution of values of an attribute in order to significantly reduce the accuracy loss caused by incorporating privacy by means of the randomisation-based methods.
6. Optimisation for MASK Scheme in Privacy Preserving Association Rules Mining,

In order to eliminate exponential complexity in estimating a support of an item set with respect to its cardinality, we propose the optimisation for MASK scheme.[6] Instead of $O(2^n)$ complexity, where n is the number of all possible items in a database, we have $O(2^{rThreshold})$, where r Threshold, r Threshold < n, is a constant. We will call MASK scheme with our proposed optimisation as MMASK (Modified MASK).

Reducing Number of Items in Estimating n-item sets Support

The reduction of a number of items in estimating the original support of an n-item set X can be obtained by choosing for an item set X a subset of distorted transactions for estimation of the true support.[7] Let reduction Threshold (r Threshold) denote the maximal length of an item set used in estimating the support of an n-item set X, rT hreshold < n.

The true support of X can be estimated as the support of a reduced item set R X in transactions which support X n R in a true database, jRj < rT hreshold.

Example Let rT hreshold = 3. We would like to estimate the support of the item set X = fa; b; c; dg using no more than rT hreshold items. Thus, the support of the item set X is estimated as the support of, e.g., the reduced item set R = fdg in transactions which support the item set X n R = fa; b; cg.

As there is no access to a true database T , the subset of the chosen distorted transactions D_R from the distorted database D, D_R D, should support X nR in the true database T with a high probability. The CTS algorithm for choosing distorted transactions which support a given item set X n R in the true database T with a high probability is proposed. A probability that a distorted transaction supports a given item set in the true database is estimated based on the distorted set of transactions.

Process of Mining Frequent Item sets with MMASK

The Privacy Preserved Apriori-MMASK (PP Apriori-MMASK) algorithm for mining frequent item sets which uses MMASK scheme estimates original supports of candidate m-item sets like the Apriori algorithm modified to use MASK, PP Apriori-MASK, until m is less than or equal to r Threshold.[8] When m is greater than rT hreshold, the support of the m-item set X is determined by estimating the support of a reduced item set R in the set D_R of distorted transactions which support X n R, jX n Rj = rT hreshold, in the true database with a high probability.

The true supports for item sets with higher length are estimated based on the transaction set D_R D until the length of a reduced item set exceeds rT hreshold. Then a subset of transactions D_R0 is chosen (by means of the CTS algorithm, for details refer to the subset D_R. The chosen D_R0 subset of transactions support the subset X^0 n R^0 of the item set candidate X^0, $|X^0nR^0|$ = rT hreshold, X X^0, with a high probability. The true support of the item set X^0 is estimated as the support of an item set R^0 in the subset of transactions D_R0.

The true supports for longer candidate item sets are estimated based on the transaction set D_R0 until the length of a reduced candidate item set exceeds r Threshold.[9] Then the subset of transactions D_R00 is chosen (by means of the CTS algorithm, for details refer to the process is continued.

We will use the following notation for MMASK :

— X_m denotes candidate m-item sets, which are potentially frequent.

— F_m are frequent m-item sets based on estimations of original supports of item sets.

— X[i] is the i-th item in the item set X.

— X[1] X[2] X[3] : : : X[m] denotes m-item set, which consists of X[1]; X[2]; X[3]; : : : ; X[m].

— T is the original data set.

— D is the data set distorted according to the MASK scheme and each item i is distorted according to the matrix M_i.

— X:R is the reduced item set of X.

— $X:R:C^D$ is the support vector field of the reduced item set X:R in the distorted data set D.

— $X:R:C^T$ is the support vector field of the reduced item set X:R in the true data set.

— $X:R:C^T_j$ is the j-th element of the vector $X:R:C^T$.

— $X:R:C^D_j$ is the j-th element of the vector $X:R:C^D$.

— X:R:M is a M matrix (see Equation 3.3) for the reduced item set X:R.

— $X:R:M^{1}$ is an inverted M matrix for the reduced item set X:R.

— $X:R_F$ is the lexicographically first ancestor of the reduced item set X:R.

— $X:D_R$ is the subset of transactions from the database D which support X:R with a high probability.

Algorithm 1: The PP Apriori-MMASK algorithm, Apriori algorithm modified to use MMASK

input: minimumSupport

input: D // binary distorted data set

input: rT hreshold, rT hreshold > 0

F_1 =f1-itemsets which are frequent based on estimations of original support of singletonsg for all X 2 F_1 do begin

$X:D_R = D$
$X:R = X$

end

index = 1
for (m = 2; $F_{m\ 1}$ 6= ;; m++) do begin
index++

X_m = aprioriGen($F_{m\ 1}$) //generate new candidates if index > rT hreshold then begin
for all X 2 X_m do begin

$X:D_R$ = fO 2 $X:D_R$jO supports $X:R_F$ with a high probability in the true data setg end

end
supportCount(X_m)

F_m = fX 2 X_mjX:R:C^T_2m $_1$ if index > rT hreshold then

end
S
return $_m F_m$

minimumSupport g index = 1

The PP Apriori-MMASK algorithm generates candidates for frequent sets with a given length in the Apriori-like fashion. In every iteration of candidates generation when the reduced item-set used to estimate an original support of a candidate in the true database based on a support counted in distorted transactions exceeds r Threshold, the reduction of the set of transactions which is used to estimate an original support of the candidate is performed.[10] Then, having estimated the original supports of candidates, the minimum support condition is checked and candidates which are not frequent are removed. As the result of the PPApriori-MMASK algorithm, the item sets with the estimated support greater than or equal to minimum Support are provided.

Example Let rT hreshold = 3. We would like to estimate the support of the candidate X = fa; b; c; dg based on distorted database D. This candidate was generated in the Apriori-like fashion from the frequent sets fa; b; cg and fa; b; dg. We can use either fa; b; cg or fa; b; dg as a condition to reduce the transaction set. Let us assume that we choose the lexicographically first set, fa; b; cg^1, thus, R = fdg, X n R = fa; b; cg. Then the subset D_R of the distorted transactions which support X n R with a high probability is chosen (we use the CTS algorithm for choosing transactions subset described in the next section to achieve this goal). Having chosen the subset D_R of the transactions, we estimate the true support of the candidate X = fa; b; c; dg as the estimated true support for the singleton R = fdg based on the subset D_R of distorted transactions because all distorted transactions in D_R used to estimate the support of = fdg support the frequent set X n R = fa; b; cg in the true database with a high probability. We compute the support for supersets of X n R = fa; b; cg
in the same manner until candidates with the length greater than jfa; b; cgj + r Threshold, for instance X^0 = fa; b; c; d; e; f; gg, appear.

Given the subset D_R, we can choose from this subset the transactions which support fd; e; fg set with a high probability and obtain D_R0 subset, which contains distorted transactions. Those transactions support X^0 n R^0 = fa; b; c; d; e; fg item set with a high probability in the true database, because the distorted transactions in D_R already support X n R = fa; b; cg in the true database with a high probability. Then, we estimate the support of X^0 = fa; b; c; d; e; f; gg set like we compute the singleton R^0 = fgg support based on distorted transactions in D_R0.

The reduced subset of distorted transactions is chosen (for each candidate) only in those passes in which candidates have length k rT reshold + 1; k = 1; 2::: In other passes (for a given candidate) the reduced subset of distorted transactions from the latest pass is used. Thus, the number of reduced sets of distorted transactions is less than or equal to the number of candidates in the current pass. Instead of MASK $O(2^n)$ complexity of estimating an original support of item sets with respect to n, the number of all possible items in a database, MMASK has $O(2^{r\ \text{Threshold}})$ complexity, where r Threshold, r Threshold < n, is a constant. We will illustrate the reduction of complexity of estimating an original support of item sets with respect to their length

on the following example. Example Let rT hreshold = 3 and we would like to estimate the support of the candidate item sets X = fa; b; c; dg and X^0 = fa; b; c; d; e; f; gg based on distorted database D, like in the previous example. In future work, we plan to investigate which subset should be chosen to achieve the best accuracy.

Algorithm 2: The candidate generation algorithm for MMASK

function apriori Gen(var F_m)

for all Y; Z 2 F_m do begin
if Y [1] = Z[1] ^ : : : ^ Y [k 1] = Z[k 1] ^ Y [k] < Z[k] then begin
X = Y [1] Y [2] Y [3] : : : Y [k 1] Y [k] Z[k]
$X:D_R = Y:D_R$
if index > rT hreshold then begin
$X:R_F = Y:R$
X:R = Z[k]
end else begin

X:R = Y:R [Z[k]
end
add X to X_{m+1}
end
end
for all X 2 X_{m+1} do begin
for all m-itemsets Z X do begin
if 62 F then X
m+
Z m delete X from 1
end

end
return X_{m+1}
end

For better visualisation of the reduction of complexity of estimating an original support of Item sets with respect to their length, we consider also the estimation of the support of some subsets of the candidate item sets X = fa; b; c; dg and X^0 = fa; b; c; d; e; f; gg, for instance, fag, fa; b; cg, etc. For the item set fag, the vector C^D and the estimated vector C^T have $2^1 = 2$ elements and these vectors are the same for MASK and MMASK:

$$C^D = \begin{matrix} 2 \\ C_1^D \end{matrix} \quad D \quad \begin{matrix} 3 \\ \end{matrix} \; ; C^T = \begin{matrix} 2T \\ C_1^T \end{matrix} \quad \begin{matrix} 3 \\ \end{matrix} \quad : \qquad (4.1)$$

$$\begin{matrix} 4 \\ \end{matrix} \quad {}^C0 \quad \begin{matrix} 5 \\ \end{matrix} \qquad \begin{matrix} 4 \\ \end{matrix} {}^C0 \quad \begin{matrix} 5 \\ \end{matrix}$$

For the item set fa; b; cg, the vector C^D and the estimated vector C^T have 2^3 = 8 elements and these vectors are also the same for MASK and MMASK, because jfa; b; cgj = r Threshold = 3:

Algorithm 3: The support count algorithm for MMASK

procedure supportCount(var X_m)

for all transactions T 2 D do begin
for all candidates X 2 X_m do begin

if T 2 $X:D_R$ then $X:C^D_j$++ //j is the number which has a binary form (in m digits) // of X:R in the transaction T

end

end
for all candidates X 2 X_m do begin
$X:R:C^T = X:R:M\ ^1X:R:C^D$
end

end

$$C^D = \begin{matrix} 2\ C_2^D 3_{\ 1=7} & 3 \\ 6\ :\ & 7 \\ 6\ :\ & 7 \\ 6 & 7;\ C^T = \\ 6\ :\ & 7 \\ 6 & 7 \\ 6 & 7 \\ 6\ C^D & 7 \\ 6\ \ 1 & 7 \\ 6 & 7 \\ 6\ {}_C D & 7 \\ 6\ \ 0 & 7 \\ 4 & 5 \end{matrix} \quad \begin{matrix} 2\ C_2^T 3_{\ 1=7} & 3 \\ 6\ :\ & 7 \\ 6\ :\ & 7 \\ 6 & 7 \\ 6\ :\ & 7 \\ 6 & 7 \\ 6 & 7 \\ 6\ C^T & 7 \\ 6\ \ 1 & 7 \\ 6 & 7 \\ 6\ {}_C T & 7 \\ 6\ \ 0 & 7 \\ 4 & 5 \end{matrix} : \qquad (4.2)$$

For the item set X = fa; b; c; dg, the vector C^D and the estimated vector C^T for MASK have $2^4 = 16$ elements and the vector C^D and the estimated vector C^T for MMASK have $2^1 = 2$ elements, because jX = fa; b; c; dgj > r Threshold = 3 and the support of the item set x= fa; b; c; dg is calculated as the estimation of the support of the reduced item set R = fdg based on the subset D_R of the distorted transactions which support X n R = fa; b; c; dg n fdg = fa; b; cg with a high probability in the true database:

$$\begin{matrix} {}^D\text{MASK} \\ \\ = \end{matrix} \begin{matrix} 2\ C_2^D 4_{\ 1=15} & 3 \\ 6\ :\ & 7 \\ 6\ :\ & 7 \\ 6 & 7 \\ 6\ :\ & 7 \\ 6 & 7 \\ 6 & 7 \\ 6\ C^D & 7 \\ 6\ \ 1 & 7 \\ 6 & 7 \\ 6\ C^D & 7 \\ 6\ \ 0 & 7 \\ 4 & 5 \end{matrix} \quad \begin{matrix} {}^T\text{MASK} = \\ \\ \end{matrix} \begin{matrix} 2\ C_2^T 4_{\ 1=15} & 3 \\ 6\ :\ & 7 \\ 6\ :\ & 7 \\ 6 & 7 \\ 6\ :\ & 7 \\ 6 & 7 \\ 6 & 7 \\ 6\ C^T & 7 \\ 6\ \ 1 & 7 \\ 6 & 7 \\ 6\ C^T & 7 \\ 6\ \ 0 & 7 \\ 4 & 5 \end{matrix} ;$$

$$C_{MMASK}{}^D = \begin{matrix} 2 \\ 4 \end{matrix} \begin{matrix} D \\ {}_C D \\ 1 \end{matrix} \begin{matrix} 3; \\ 5 \end{matrix} \quad C_{MMASK}{}^T = \begin{matrix} 2 \\ 4 \end{matrix} \begin{matrix} T \\ {}_C T \\ 1 \end{matrix} \begin{matrix} 3: \\ 5 \end{matrix} \qquad (4.3)$$

C_0 C_0

For the item set fa; b; c; d; e; fg, the vector C^D and the estimated vector C^T for MASK have $2^6 = 64$ elements and the vector C^D and the estimated vector C^T for MMASK have $2^3 = 8$ elements, because jfa; b; c; d; e; fgj > rT hreshold = 3 and the support of the item-set fa; b; c; d; e; fg is calculated as the estimation of the support of reduced item set fd; e; fg based on the subset D_R of the distorted transactions which support fa; b; c; d; e; fg n fd; e; fg = fa; b; cg = X n R with a high probability in the true database.[13] For the item set X^0 = fa; b; c; d; e; f; gg, the vector C^D and the estimated vector C^T for MASK have $2^7 = 128$ elements. For MMASK, the reduced item set jfd; e; f; ggj would have the length greater than r Threshold = 3, hence, the reduction is performed for the second time and the support of the item set X^0 = fa; b; c; d; e; f; gg is calculated as the estimation of the support of reduced item set R^0 = fdg based on the subset D_R0 D_R of the distorted transactions which support X^0 n R^0 = fa; b; c; d; e; f; gg n fgg = fa; b; c; d; e; fg with a high probability in the true database. The distorted transactions in subset D_R0 are chosen from the subset D_R on the condition that those chosen transactions support fd; e; fg with a high probability in the true database.[14] Please, observe that the distorted transactions from the subset D_R support n R = fa; b; cg with a high probability in the true database, hence, by choosing the distorted transactions from D_R that support fd; e; fg with a high probability in the true database, we find the distorted transactions that support X^0 nR^0 = fa; b; c; d; e; f; ggnfgg = fa; b; c; d; e; fg with a high probability in the true database.

In each pass, the vector C^D and the estimated vector C^T for MMASK have at most $2^{rT\ hreshold}$ elements, where rT hreshold is a constant, contrary to MASK, where the number of elements is equal to 2^k, where k is the length of an item set.

CTS Algorithm for Choosing Transactions Subset

Having a vector C^D and an estimated vector C^T (for detailed information about the structure of vectors C^D and C^T refer to Section 3.6.1), we know an estimated support $C_2^T n_1$ of a candidate item set C (for instance, the candidate fa; b; cg). Thus, $C_2^T n_1$ transactions should be kept in the subset D_R^2. Randomisation factors in the range h0:7; 0:9i, which are usually used, let us infer that it is very probable that transactions which support a candidate item set C in a distorted database also support this set in a true database.[15] Thus, we want to keep these transactions in our subset. There are two possible cases:

We focus only on the $C_2^T n_1$ and $C_2^D n_1$ values because we are interested in the support of the set fa; b; cg, but not its subsets.

— $C_2^T n_1 C_2^D n_1$ – in this case we choose the first $C_2^T n_1$ transactions[3] from a distorted database which support a candidate item set C (for instance, fa; b; cg)[4].

— $C_2^T n_1 > C_2^D n_1$ – all $C_2^D n_1$ distorted transactions which support a candidate item set C (for instance, fa; b; cg) in a distorted database are kept in the subset C_A. Then we choose i (0 i 2^n 2) for which there is the highest probability[5] that a true tuple which supports a candidate item set set C (for instance, fa; b; cg) was distorted to value i and C_i^D is greater than zero.

Example Let $C_2^T n_1$ $C_2^D n_1$, rT hreshold = 2, and the candidate is fa; b; cg.

$$C^D = \begin{matrix} 2 & 172 & 3 \\ 6 & : & 7 \\ 6 & : & 7 \\ 6 & & 7 \\ 6 & : & 7 \\ 6 & & 7 \\ \\ 4 & & 5 \end{matrix} \quad ; \quad C^T = \begin{matrix} 2 & 168 & 3 \\ 6 & : & 7 \\ 6 & : & 7 \\ 6 & & 7 \\ 6 & : & 7 \\ 6 & & 7 \\ \\ 4 & & 5 \end{matrix}$$

The first 168 distorted transactions from 172 transactions which support fa; bg in the distorted database D are chosen. Example Let $C_2^T n_1 > C_2^D n_1$, rT hreshold = 2, the candidate is fa; b; cg.

$$C^D = \begin{matrix} 2 & 172 & 3 \\ 6 & 20 & 7 \\ 6 & 9 & 7 \\ 6 & & 7 \\ 6 & : & 7 \\ 6 & & 7 \\ \\ 4 & & 5 \end{matrix} ; C^T = \begin{matrix} 2 & 191 & 3 \\ 6 & 23 & 7 \\ 6 & 11 & 7 \\ 6 & & 7 \\ 6 & : & 7 \\ 6 & & 7 \\ \\ 4 & & 5 \end{matrix}$$

All 172 distorted transactions which support fa; bg in the distorted database are chosen. The additional 191 - 172 = 19 transactions are needed. Then the highest probable value of i (possible values are: 0, 1 or 2) is found. The probabilities for different values of i are computed based on M. Let us assume that i = 1. Then we choose 9 distorted transactions which support fbg in the distorted database. There are still 10 transaction left to choose. Let assume that the second highest probability that a distorted transaction comes from an original transaction which support fa; bg is for i = 2. There are 20 distorted transactions for i = 2, that is, the distorted transactions which support fag in the distorted database. The first 10 of them are chosen. Thus, distorted transactions are chosen. Considering only items from the set fa; b; cg, all $C_2^D n_1$ transactions have the same probability to support fa; b; cg in the true database. We plan to determine the best way to choose $C_2^T n_1$ transaction. We use M matrix to compute this probability.

Error Measures

We use two kinds of mining errors presented in Support Error and Identity Error, in our experiments:

— Support Error ( ):

This measure reflects the average relative error in the reconstructed support values for those item sets that are correctly identified to be frequent. Denoting the reconstructed support by rec Support and the actual support by act Support, the support error is computed over all frequent item sets F as follows:

$$= \frac{1}{jF\,j} \sum_{x2F} \frac{jrecSupport_x \quad actSupport_x j}{actSupport_x} \quad 100\ [\%]:$$

We compute this measure separately for each length of item sets, that is, for 1-itemsets, 2-item- sets, etc.

— Identity Error ( ):

This measure reflects the percentage error in identifying frequent item sets and has two components: $^+$ indicating the percentage of false positives, and indicating the percentage of false negatives.[16] Denoting the reconstructed set of frequent item sets with R and the correct set of frequent item sets with F , these measures are computed as follows:

$$^+ = \frac{jR \cap F\,j}{jF\,j} \quad 100\ [\%]; \qquad = \frac{jF \cap Rj}{jF\,j} \quad 100\ [\%]:$$

An additional measure to those used in calculation in the experiments.

— Accuracy of Identity (f): This measure reflects the accuracy of identifying frequent item sets and shows how many sets are correctly identified to be frequent.

f $= jF \setminus Rj$

This measure is equal to the number of all frequent sets mined in a true database minus false negatives.

f $= jF\,j^{jF\,j\ 100}$

Relaxation

As the false negative error causes the mining process to miss some frequent item sets, it is possible to use an effect of marginally relaxing minimum Support. The relaxation proposed in can decrease the false negative error component, which decreases the number of true frequent item sets that are missed. As a result of the relaxation the false positive error components goes up, inevitably attracts not frequent sets. The relaxation can be also applied to the MMASK scheme.

Algorithm 4: The PP Apriori-rMMASK algorithm, the apriority algorithm modified to use the rMMASK scheme (i.e., MMASK and relaxation)

input: minimum Support

input: D // binary distorted data set

input: rT hreshold, rT hreshold > 0

input: relax

input: rrelax
F_1 =f1-itemsets which are frequent based on estimated original support of singletons,
that is, have estimated support greater than or equal to ^{minimumSupport} g
1+relax
for all X 2 F_1 do begin
$X:D_R = D$
$X:R = X$

end

index = 1

currentRelax = relax // relaxation
for (m = 2; $F_{m\ 1}$ 6= ;; m++) do begin
index++
X_m = aprioriGen($F_{m\ 1}$) //generate new candidates
if index > rT hreshold then begin
for all X 2 X_m do begin
$X:D_R$ = fO 2 $X:D_R$jO supports $X:R_F$ with a high probability in the true databaseg end
end
supportCount(X_m)
F_m = fX 2 X_mjX:R:C^T_2m $_1$ $^{minimumSupport}_{1+currentRelax}$ g
if index > rT hreshold then begin
index = 1
currentRelax = currentRelax + rrelax // reduction relaxation
end
end

return $^S_m F_m$

Let relax be the parameter of the relaxation, for instance, relax = 0:02 = 2%. The relaxed minimum support (minimumSupport^0) is calculated as follows:

$$\text{minimumSupport}^0 = \frac{\text{minimumSupport}}{1 + \text{relax}} = \frac{\text{minimumSupport}}{1 + 0{:}02} :$$

Moreover, the relaxation can be repeated every time the reduction of the distorted transaction set is performed. We will call it the reduction relaxation (with the rrelax parameter). The modified minimum support ($minimumSupport^0$) in the reduction relaxation is calculated as follows:

$$minimumSupport^0 = \frac{minimumSupport}{1 + rrelax}:$$

These two relaxations can be combined at the same time in the MMASK scheme, we will call this scheme rMMASK (please, see the PPApriori-rMMASK algorithm, Algorithm 13).

At the beginning, the PP Apriori-rMMASK algorithm finds frequent sets with relaxation equal to the parameter relax.[17] Then, the relaxation is increased by the parameter rrelax in each iteration if the length of the reduced item set is greater than rT hreshold. As the result of the PP Apriori-rMMASK algorithm, the item sets with the estimated support greater than or equal to relaxed minimumSupport are provided.

Experimental Evaluation

In this section, we present the results of the experiments conducted to check the accuracy and the efficiency of the modified MASK scheme.

Data Sets

Our experiments were carried out on the following databases:

— A database, Led-24 from UCI Machine Learning Repository, which contains in-formation about Light Emitting Diode. There are about 3200 tuples with 25 attributes, 24 of them are binary. One attribute was not distorted because it is a nominal attribute and was treated as 0 for value of 0 and 1 for the remaining values to transform it to a binary attribute.

— A real database, Dna from UCI Machine Learning Repository, which contains information about DNA sequences.[18] There are 2000 tuples with 180 binary attributes and 1 nominal attribute. One attribute was not distorted because it is a nominal attribute and was treated as 0 for value of 0 and 1 for the remaining values to transform it to a binary attribute.

— A synthetic database, T10I8D100kN100, generated from the IBM Almaden generator. The data set was created with parameters T=10 (the average size of the transactions), I=8

Table 1: The results of mining the frequent sets in the set T10I8D100kN100 with parameters p = 0.5, q = 0.97, rThreshold = 3, minimumSupport = 0.005

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	98	97	2.5	1.0	0.0	97	97	2.5	1.0	0.0	97
2	2522	2520	7.2	3.6	3.5	2432	2520	7.2	3.6	3.5	2432
3	10930	11553	11.6	9.3	15.0	9910	11553	11.6	9.3	15.0	9910
4	10185	12758	17.1	17.6	42.9	8389	7474	13.6	29.2	2.6	7214
5	2021	4499	26.3	26.0	148.6	1495	665	21.5	67.2	0.1	662
6	24	440	49.4	58.3	1791.7	10	2	0.7	95.8	4.2	1
7	0	4	-	-	-	0	0	-	-	-	0

Table 2: The results of mining the frequent sets in the set T10I8D100kN100 with parameters p = 0.5, q = 0.87, rThreshold = 3, minimumSupport = 0.005

L.	jF oj	jF rj	r	$_r$	$+_r$	^{f}r	jF rmj	rm	rm	$^+$rm	^{f}rm
1	98	98	5.4	1.0	1.0	97	98	5.4	1.0	1.0	97
2	2522	2704	20.4	10.8	18.0	2250	2704	20.4	10.8	18.0	2250
3	10930	16780	37.5	25.9	79.5	8094	16780	37.5	25.9	79.5	8094
4	10185	22411	62.4	40.1	160.2	6098	7787	16.4	36.3	12.7	6490
5	2021	5810	115.9	57.9	245.4	851	1129	16.6	57.3	13.2	862
6	24	200	318.5	79.2	812.5	5	28	6.8	70.8	87.5	7

(the average size of the maximal potentially frequent item sets), D=100k (the number of transactions), N=100 (the number of items). More information on a generator and naming convention can be found in [6]). The data set contains about 100000 tuples with each customer purchasing about ten items on average.

— A synthetic database, T20I16D50kN200, generated from the IBM Almaden generator [6]. The data set was created with parameters T=20 (the average size of the transactions), I=16 (the average size of the maximal potentially large item sets), D=50k (the number of trans-actions), N=200 (the number of items). It contains about 50000 tuples with each customer purchasing about 20 items on average.

Accuracy vs Privacy

The first experiment was conducted on the synthetic database with the distortion parameters of p = 0:5 and q = 0:97, and no relaxation. The results of this experiment are shown in Table 1. The level (denoted as L. in tables), which corresponds to the consecutive iterations in Apriori-like algorithms, indicates the length of frequent item sets, jF_0j indicates the number of frequent item sets at a given level, jF_rj ($jF_{rm}j$) shows the number of mined frequent item-sets from the distorted database using MASK (MMASK).

Table 3: The results of mining the frequent sets in the set T10I8D100kN100 with parameters p = 0.5, q = 0.77, rThreshold = 3, minimumSupport = 0.005

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	98	95	8.3	4.1	1.0	94	95	8.3	4.1	1.0	94
2	2522	2733	47.3	20.3	28.7	2010	2733	47.3	20.3	28.7	2010
3	10930	19677	127.7	42.8	122.9	6248	19677	127.7	42.8	122.9	6248
4	10185	20248	305.9	68.9	167.7	3168	8764	31.9	62.6	48.7	3809
5	2021	1515	594.0	92.5	67.5	151	1319	30.9	69.2	34.5	622
6	24	3	-	100.0	12.5	0	74	99.7	70.8	279.2	7

The results indicate that, firstly, for MASK the support error ( ) is less than 10% for the two first levels. The support error grows up to about 50% for 6-itemsets. The modified algorithm achieves the same results for level 1-3 because rT hreshold is 3. For higher levels is less than 22%, which is more than 2 times less than for the original algorithm. Secondly, the negative identity errors are high for both algorithms, especially for level 6 - about 50% for the original algorithm and 95% for the modified (see the experiment with the relaxation). The positive identity errors are very high for MASK (about 150% for level 5 and more than 10 times higher for level 6). For levels 1-3 the highest positive error is for level 3, namely 15%. For higher levels the positive identity error for MMASK does not exceed 5%. The comparison measures f_r and f_{rm} show that for levels higher than 3 MMASK discovers less true frequent sets.[20]

Summarising, MMASK for p = 0:5; q = 0:97 has lower support error and positive error and higher negative error. This makes f_r higher than f_{rm}.

Tables 3 and 4 show the results of the experiment for the set T10I8D100kN100 with lower q (higher privacy), q = 0:87 and q = 0:77, respectively.
Decreasing value of q (p is constant and equal to 0.5) results in increasing privacy. For = 0:97; 0:87 and 0:77 Basic Privacy is equal to 63.8%, 81% and 86.1%, respectively. Thus, a drop of q from 0.97 to 0.87 causes Basic Privacy to increase by more than 17%.

As stated in [98], MASK performs much more worse with lower probabilities for p = q. Conducted experiments confirmed this property of MASK (constant p and variable q).

MMASK performs significantly better for lower probabilities. The support error for MASK is as high as 300-600% for levels 5-6, when for MMASK does not exceed 17% for levels 4-6 with q = 0:87 and 32% for levels 4-5 with q = 0:77. Level 6 is critical for q = 0:77, because support error is 99.7%, but it is still better than MASK, because the original algorithm has not discovered any true frequent set.[21] There is the only one case when MMASK performs worse than MASK the positive error for level 6 and q = 0:77.

Table 4: The results of mining the frequent sets in the set Dna with parameters p = 0.5, q = 0.97, rThreshold = 3, minimumSupport = 0.05

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	181	181	4.4	0.0	0.0	181	181	4.4	0.0	0.0	181
2	13126	11715	13.7	17.2	6.5	10867	11715	13.7	17.2	6.5	10867
3	11118	23213	17.2	30.5	139.3	7723	23213	17.2	30.5	139.3	7723
4	1403	9314	25.7	36.3	600.1	894	4118	22.0	47.3	240.8	739
5	174	1245	34.9	49.4	664.9	88	91	13.0	74.1	26.4	45
6	4	69	57.8	50.0	1675.0	2	0	-	100.0	0.0	0

Table 5: The results of mining the frequent sets in the set Dna with parameters p = 0.5, q = 0.87, rThreshold = 3, minimumSupport = 0.05

L.	jF oj	jF rj	r	$_r$	$+_r$	^{f}r	jF rmj	rm	rm	$^+$rm	^{f}rm
1	181	181	7.2	0.0	0.0	181	181	7.2	0.0	0.0	181
2	13126	10467	29.1	29.4	9.1	9266	10467	29.1	29.4	9.1	9266
3	11118	89833	40.2	44.3	752.3	6197	89833	40.2	44.3	752.3	6197
4	1403	73920	58.3	57.1	5225.8	602	22635	39.5	57.8	1571.1	592
5	174	6554	60.2	82.2	3748.9	31	153	16.6	88.5	76.4	20
6	4	87	-	100.0	2175.0	0	0	-	100.0	0.0	0
7	0	1	-	-	-	0	0	-	-	-	0

To sum up, MMASK is significantly better with higher privacy (lower probability q). The accuracy error is always better for MMASK (for levels greater than 3).

Appendix A.1 presents the results of the experiments for the set T20I16D50kN200 with q = 0:87 and q = 0:77, where p is constants and equals to 0:5.

The results of the experiments with p = q for synthetic data sets are quite similar. The modified algorithm accomplishes better results than MASK for p = 0:8 and p = 0:7.

The experiments on the real database (either with p = q or different p and q) lead to the same conclusions (results for the set Dna are shown in Tables 4, 5, 6).

Table 6: The results of mining the frequent sets in the set Dna with parameters p = 0.5, q = 0.77, rThreshold = 3, minimumSupport = 0.05

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	181	181	12.1	0.0	0	181	181	12.1	0.0	0	181
2	13126	9423	67.0	38.9	11	8019	9423	67.0	38.9	11	8019
3	11118	105312	96.3	56.9	904	4789	105312	96.3	56.9	904	4789
4	1403	135880	122.3	79.5	9665	287	89646	84.1	76.0	6366	337
5	174	12075	100.0	92.0	6932	14	3807	27.9	87.4	2175	22
6	4	43	-	100.0	1075	0	6	-	100.0	150	0

Table 7: The results of mining the frequent sets with the relaxation in the set T10I8D100kN100 with parameters p = 0.5, q = 0.87, relax = 0.05, rThreshold = 3, minimumSupport = 0.005

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	98	98	5.4	1.0	1.0	97	98	5.4	1.0	1.0	97
2	2522	2802	20.5	9.6	20.7	2279	2802	20.5	9.6	20.7	2279
3	10930	17991	37.3	24.6	89.2	8244	17991	37.3	24.6	89.2	8244

4	10185	24610	61.8	38.9	180.6	6220	8688	16.6	32.0	17.3	6921
5	2021	6523	114.1	56.9	279.6	872	1397	17.1	51.6	20.7	979
6	24	239	318.5	79.2	975.0	5	39	7.6	62.5	125.0	9

Table 8: The results of mining the frequent sets with the reduction relaxation in the set T10I8D100kN100 with parameters p = 0.5, q = 0.87, rrelax = 0.05, rThreshold = 3, minimumSupport = 0.005

L.	jF oj	jF rj	r	$_r$	$+_r$	^{f}r	jF rmj	rm	rm	$^+$rm	^{f}rm
1	98	98	5.4	1.0	1.0	97	98	5.4	1.0	1.0	97
2	2522	2704	20.4	10.8	18.0	2250	2704	20.4	10.8	18.0	2250
3	10930	16780	37.5	25.9	79.5	8094	16780	37.5	25.9	79.5	8094
4	10185	22411	62.4	40.1	160.2	6098	8679	16.6	32.1	17.3	6916
5	2021	5810	115.9	57.9	245.4	851	1411	17.2	51.5	21.3	980
6	24	200	318.5	79.2	812.5	5	39	7.6	62.5	125.0	9

Table 9: The results of mining the frequent sets with both relaxations in the set T10I8D100kN100 with parameters p = 0.5, q = 0.87, relax = 0.01, rrelax = 0.02, rThreshold = 3, minimumSupport = 0.005

L.	jF oj	jF rj	r	$_r$	$+_r$	^{f}r	jF rmj	rm	rm	$^+$rm	^{f}rm
1	98	98	5.4	1.0	1.0	97	98	5.4	1.0	1.0	97
2	2522	2718	20.4	10.7	18.4	2253	2718	20.4	10.7	18.4	2253
3	10930	17000	37.5	25.7	81.2	8123	17000	37.5	25.7	81.2	8123
4	10185	22816	62.3	39.9	163.9	6123	8340	16.5	33.7	15.6	6756
5	2021	5925	115.3	57.6	250.8	857	1304	17.0	53.6	18.2	937
6	24	208	318.5	79.2	845.8	5	35	8.0	66.7	112.5	8

Table 10: The results of mining the frequent sets with different randomisation factors for items in the set T10I8D100kN100 (p = 0.5 .. 0.4, q = 0.87 .. 0.88, rThreshold = 3, minimumSupport = 0.005)

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	98	97	6.2	1.0	0.0	97	97	6.2	1.0	0.0	97
2	2522	3041	36.8	8.0	28.5	2321	3041	36.8	8.0	28.5	2321
3	10930	23321	65.9	26.3	139.7	8056	23321	65.9	26.3	139.7	8056
4	10185	31175	133.9	51.7	257.8	4916	7824	20.8	48.0	24.9	5293
5	2021	4645	307.2	83.3	213.1	338	1165	18.9	62.4	20.0	760
6	24	27	-	100.0	112.5	0	29	4.0	91.7	112.5	2

Relaxation

MASK and MMASK lead to the high false negative error. To reduce it, the relaxation can be used. Table 7 shows the results with 5% relaxation for MASK, i.e., rMASK and MMASK, i.e., rMMASK.

For rMASK and rMMASK compared to the case without the relaxation (compare the results presented in Table 2) the support error is similar. The false negative error is smaller than without the relaxation. As a negative result of the relaxation, the false positive error is higher. Lower minimum support causes the number of discovered frequent sets (true frequent sets also) to grow.

The results with 5% reduction relaxation are shown in Table 8. The support error and identity errors are quite similar compared to 5% relaxation (Table 7). f measure for levels 1-3 is lower because the reduction relaxation works for levels higher than rT hreshold. However, for levels 4-6 f measure is almost the same for both types of the relaxation. Second difference is that the reduction relaxation does not influence the original MASK scheme.

Both relaxations can be combined. The results with 1% relaxation and 2% reduction relaxation are shown in Table 9. This combined relaxation is not as strong as the relaxations presented above - the number of correctly discovered frequent item sets is lower compared with both relaxation applied separately for MMASK and levels.

By combining relaxations, we can control the number of discovered frequent item sets of particular length. A higher relaxation results in more discovered frequent item sets for all lengths of item sets. A reduction relaxation applied on a particular level makes the number of discovered frequent item sets higher for this and higher levels.[22]

Different Randomisation Factors for Items

Tables 10 and 11 show the results of the experiments with different randomisation factors for different items for T10I8D100kN100 and Led-24 databases, respectively. The half of the items was distorted with parameters p = 0:4, q = 0:88 and the remaining items with parameters p = 0:5, q = 0:87.

Table 11: The results of mining the frequent sets with different randomisation factors for items in the set Led-24 with parameters p = 0.5. 0.4, q = 0.87 . 0.88, rThreshold = 3, minimumSupport = 0.01

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	25	25	3.3	0.0	0.0	25	25	3.3	0.0	0.0	25
2	300	300	10.7	0.0	0.0	300	300	10.7	0.0	0.0	300
3	2279	1871	24.3	18.3	0.4	1862	1871	24.3	18.3	0.4	1862
4	4713	4053	30.7	47.8	33.8	2461	4033	26.1	40.5	26.1	2805
5	2654	2070	28.1	76.8	54.7	617	2381	23.8	60.4	50.2	1050
6	413	130	17.2	96.9	28.3	13	242	16.3	88.4	47.0	48
7	20	0	-	100.0	0.0	0	0	-	100.0	0.0	0

Table 12: The results of mining the frequent sets with different randomisation factors for items and the relaxation in the set Led-24 (p = 0.5 .. 0.4, q = 0.87 .. 0.88, relax = 0.01, rThreshold = 3, minimumSupport = 0.01)

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	25	25.0	3.3	0.0	0	25	25	3.3	0.0	0.0	25
2	300	300.0	10.7	0.0	0	300	300	10.7	0.0	0.0	300
3	2279	1882.0	24.3	17.8	0.39491	1873	1882	24.3	17.8	0.4	1873
4	4713	4134.0	30.6	47.1	34.861	2491	4134	26.0	39.7	27.4	2843
5	2654	2144.0	27.9	76.2	56.9706	632	2527	23.7	59.1	54.3	1086
6	413	141.0	17.2	96.9	30.9927	13	274	17.3	87.4	53.8	52
7	20	0.0	-	100.0	0	0	0	-	100.0	0.0	0

Table 13: The results of mining the frequent sets with different randomisation factors for items and the reduction relaxation in the set Led-24 (p = 0.5 .. 0.4, q = 0.87 .. 0.88, rrelax = 0.02, rThreshold = 3, minimumSupport = 0.01)

L.	jF oj	jF rj	r	$_r$	$+_r$	^{f}r	jF rmj	rm	rm	$^+$rm	^{f}rm
1	25	25	3.3	0.0	0.0	25	25	3.3	0.0	0.0	25
2	300	300	10.7	0.0	0.0	300	300	10.7	0.0	0.0	300
3	2279	1871	24.3	18.3	0.4	1862	1871	24.3	18.3	0.4	1862
4	4713	4053	30.7	47.8	33.8	2461	4161	25.9	39.5	27.8	2851
5	2654	2070	28.1	76.8	54.7	617	2591	23.6	58.6	56.3	1098
6	413	130	17.2	96.9	28.3	13	284	17.7	87.2	55.9	53
7	20	0	-	100.0	0.0	0	0	-	100.0	0.0	0

Table 14: The results of mining the frequent sets with different randomisation factors for items and both relaxations in the set Led-24 (p = 0.5 .. 0.4, q = 0.87 .. 0.88, relax = 0.01, rrelax = 0.02, rThreshold = 3, minimumSupport = 0.01)

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	25	25	3.3	0.0	0.0	25	25	3.3	0.0	0.0	25
2	300	300	10.7	0.0	0.0	300	300	10.7	0.0	0.0	300
3	2279	1882	24.3	17.8	0.4	1873	1882	24.3	17.8	0.4	1873
4	4713	4134	30.6	47.1	34.9	2491	4273	25.8	38.7	29.4	2888
5	2654	2144	27.9	76.2	57.0	632	2716	23.5	57.6	59.9	1125
6	413	141	17.2	96.9	31.0	13	310	18.4	85.5	60.5	60
7	20	0	-	100.0	0.0	0	1	-	100.0	5.0	0

Table 15: The results of mining the frequent sets with different randomisation factors for items in the set T10I8D100kN100, p=0.5 .. 0.4, q=0.87 .. 0.88, rThreshold = 2, minimumSupport = 0.005

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	98	97	6.2	1.0	0.0	97	97	6.2	1.0	0.0	97
2	2522	3041	36.8	8.0	28.5	2321	3041	36.8	8.0	28.5	2321
3	10930	23321	65.9	26.3	139.7	8056	10013	14.4	18.2	9.8	8937
4	10185	31175	133.9	51.7	257.8	4916	10174	19.7	34.5	34.4	6668
5	2021	4645	307.2	83.3	213.1	338	507	23.0	76.6	1.7	472
6	24	27	-	100.0	112.5	0	2	7.7	91.7	0.0	2

The results once again show that MMASK scheme is better than MASK for low probabilities. In both experiments, f measure is higher for the modified algorithm (for levels greater than rT reshold, namely 3 in this experiment). Only once the modified algorithm performs worse (the positive error for Led-24 database on level 6).

As shown in the experiments presented in this chapter, high false negative errors can be reduced by means of the relaxation. Tables 12, 13, and 14 show the results of the experiments with 1% relaxation (Table 12), 2% reduction relaxation (Table 13), and those relaxations combined (Table 14) for the set Led24 and different randomisation factors. We can notice the fall of the false negative error and the growth of the false positive error. The higher relaxation is, the higher f measure is achieved. The combined relaxation is stronger in reducing false negative errors than those single relaxations (the individual relaxation parameters are the same) as well as in increasing f measure.

The additional results for the experiments with applied relaxation are shown.

Summarising, MMASK is better for lower probabilities when different randomisation factors are used for items. The relaxation can be used to reduce the false negative error and boost measure.

Accuracy and rThreshold Parameter

Tables 15 and 10 show the results of the experiments on the synthetic database for rT hreshold = 2 and rT hreshold = 3, respectively. The results of the experiments for different values of rT hreshold were significantly worse. Analysing those two experiments, we can notice that the results for level 3 are better with rT hreshold = 2. For higher levels some measures are better for rT hreshold = 3, others for rT hreshold = 2.

Table 16: The results of mining the frequent sets with different randomisation factors for items in the set Led-24, p=0.5 .. 0.4, q=0.87 .. 0.88, rThreshold = 2, minimumSupport = 0.01

L.	jF oj	jF rj	$_r$	$_r$	$+_r$	f_r	jF rmj	rm	$_{rm}$	$+_{rm}$	^{f}rm
1	25	25	3.3	0.0	0.0	25	25	3.3	0.0	0.0	25
2	300	300	10.7	0.0	0.0	300	300	10.7	0.0	0.0	300
3	2279	1871	24.3	18.3	0.4	1862	2155	13.5	6.1	0.7	2139
4	4713	4053	30.7	47.8	33.8	2461	4644	16.7	22.7	21.3	3641

5	2654	2070	28.1	76.8	54.7	617	2583	19.2	45.3	42.6	1453
6	413	130	17.2	96.9	28.3	13	225	13.2	80.9	35.4	79
7	20	0	-	100.0	0.0	0	1	-	100.0	5.0	0

The experiments for Led-24 database (Tables 16 and 11) were conducted with the parameter rT hreshold = 2 and rT hreshold = 3 (no relaxation), respectively.

For Led-24 database the results were the best for the values of rT hreshold mentioned above, namely 2 and 3. The results for rT hreshold = 2 are better for level 4-6. For level 7 the false positive error is higher than for rT hreshold = 3.

Based on the performed experiments, we can say that the best value of rT hreshold depends on a given data set.[23]

Choosing rT hreshold = 2, we can expect shorter time of the mining process. But significant growth in the number of discovered frequent sets may lead to reversed proportion of processing time. On the synthetic database the process with rT hreshold = 2 is more than 4 times faster, but on the real database the process with rT hreshold = 2 is slower less than 20% (there are more frequent item sets for level 3-5 for r Threshold = 2).

Efficiency

Figures 1 and 2 show the running time of the original algorithm based on Apriori and the modified algorithm, as compared to Apriori itself, for various settings of the minimum support parameter for Led-24 and T10I8D100kN100 database, respectively. The experiments with the original algorithm and MMASK were conducted on the distorted database and Apriori was applied on the original database. Running time of rMMASK with the relaxation equal to 1% and the reduction relaxation equal to 2% was very close to MMASK time (e.g., for the set T10I8D100kN100 and rMMASK time was about 1-2% higher than for MMASK) and was omitted to better visualise differences in running time. Figures 1 and 2 show that there are huge differences in running times between MASK and Apriori algorithm – specifically, mining the distorted database with the original algorithm can take as much as three to four times more time than in the case when the original database is mined. Overheads between those two algorithms are larger for lower minimum support.

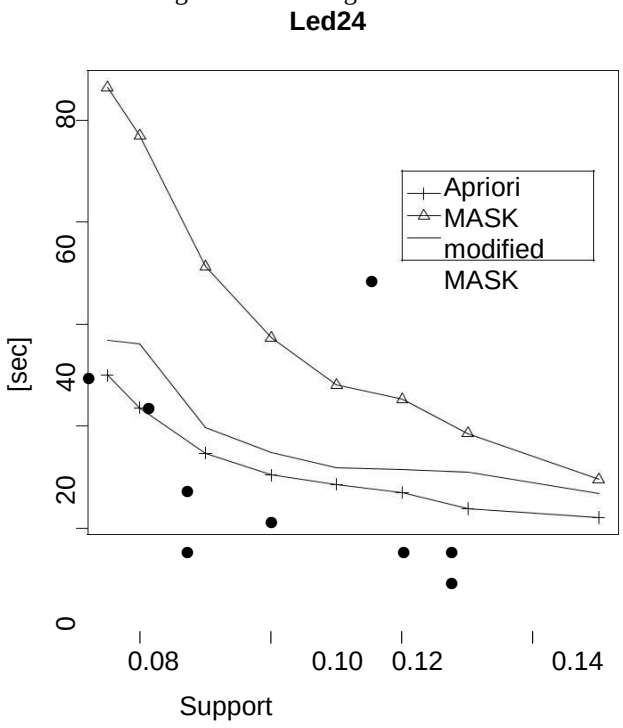

Figure 1: Time [s] vs support for mining the association rules in the set Led-24

The exhibited enhancement MMASK, which produces hopefuls in the Apriori-like fash-particle, influences the season of the mining to process practically as quick as Apriori. This is the favourable position, which makes MMASK reasonable by and by (EMASK can't utilize distinctive randomisation factors for 0's and 1's esteems).

The lessening in time cost could be additionally identified with bring down number of incessant things discovered by MMASK than MASK.

III. CONCLUSIONS AND FUTURE WORK

We explored the issue of the adequacy and productivity in the security protecting MASK conspire, and proposed the new advancement, MMASK, which is not quite the same as those exhibited in writing committed to protection saving information mining. MMASK gauges support of item sets in light of a decreased subset of twisted exchange. The primary favourable position of this enhancement is that it breaks the exponential intricacy of evaluating backing of an item set regarding its cardinality and makes finding successive item sets and, by this, affiliation rules with saving security reasonable by and by. The following favourable position is that the proposed optimisation can be utilized with various randomisation factors for 0's and 1's, that is, the point at which a thing is not present and is present in an original database. Moreover, it allows different items to have different randomisation factors.

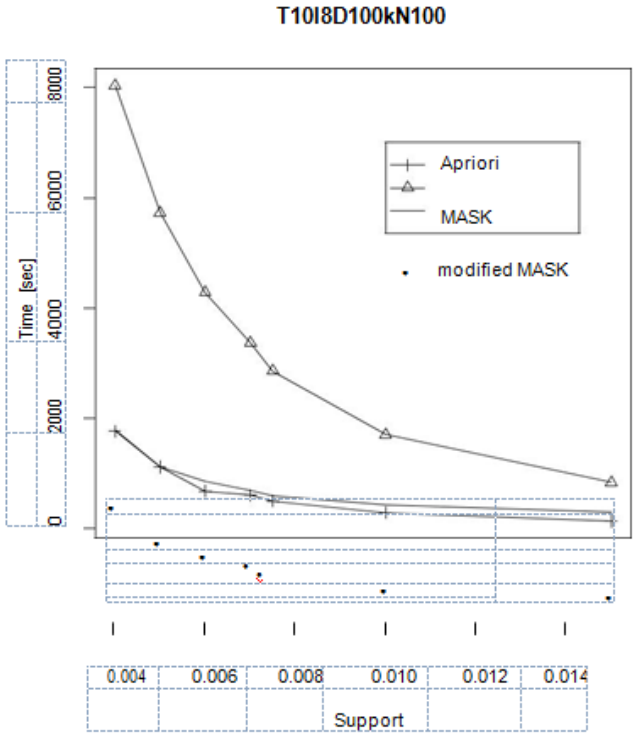

Figure 2: Time [s] vs support for mining the association rules in the synthetic set T10I8D100kN100

Besides, for abnormal amounts of protection it accomplishes altogether preferred outcomes over the first MASK plot. The adequacy and effectiveness of the new arrangement have been tried on the manufactured and genuine databases. In future work, we intend to examine the likelihood of augmentation of our outcomes to quantitative and summed up affiliation rules.

We likewise plan to explore which subset of a hopeful ought to be picked as a condition to decrease an exchange set to accomplish the best precision when a competitor length surpasses reductionT hreshold parameter. Another conceivable arrangement is to consolidate the outcomes got from subsets of an applicant.

We intend to decide the most ideal approach to pick C2Tn 1 exchanges while picking mutilated exchanges which bolster an applicant item set with a high likelihood in a genuine database. Presently the principal C2Tn 1 exchanges are picked.

We likewise plan to figure out what are the best esteems for the rThreshold and discover a govern to help a digger to pick the best estimation of the rT hreshold for a given set.

REFERENCES

[1] Charu C. Aggarwal and Philip S. Yu. Privacy-Preserving Data Mining: Models and Algorithms. Springer Publishing Company, Incorporated, 2008.

[2] Dakshi Agrawal and Charu C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 247–255, 2001.

[3] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, SIGMOD Conference, pages 207–216. ACM Press, 1993.

[4] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. Order-preserving encryption for numeric data. In Gerhard Weikum, Arnd Christian König, and Stefan Deßloch, editors, SIGMOD Conference, pages 563–574. ACM, 2004.

[5] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In Advances in Knowledge Discovery and Data Mining, pages 307–328. AAAI/MIT Press, 1996.

[6] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, edi-tors, VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile, pages 487–499. Morgan Kaufmann, 1994.
[7] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, editors, SIGMOD Conference, pages 439–450. ACM, 2000.
[8] Rakesh Agrawal, Ramakrishnan Srikant, and Dilys Thomas. Privacy preserving olap. In SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pages 251–262, New York, NY, USA, 2005. ACM.
[9] Shipra Agrawal, Vijay Krishnan, and Jayant R. Haritsa. On addressing efficiency concerns in privacy preserving data mining. CoRR, cs.DB/0310038, 2003.
[10] Shipra Agrawal, Vijay Krishnan, and Jayant R. Haritsa. On addressing efficiency concerns in privacy-preserving mining. In Yoon-Joon Lee, Jianzhong Li, Kyu-Young Whang, and Doheon Lee, editors, DASFAA, volume 2973 of Lecture Notes in Computer Science, pages 113–124. Springer, 2004.
[11] Leila N. Alachaher and Sylvie Guillaume. Variables interaction for mining negative and positive quantitative association rules. In ICTAI, pages 82–85. IEEE Computer Society, 2006.
[12] Piotr Andruszkiewicz. Privacy preserving data mining on the example of classification (in Polish). Master's thesis, Warsaw University of Technology, 2005.
[13] Piotr Andruszkiewicz. Optimization for mask scheme in privacy preserving data mining for association rules. In Marzena Kryszkiewicz, James F. Peters, Henryk Rybinski,´ and Andrzej Skowron, editors, RSEISP, volume 4585 of Lecture Notes in Computer Science, pages 465–474. Springer, 2007.
[14] Piotr Andruszkiewicz. Privacy preserving classification for continuous and nominal at-tributes. In Proceedings of the 16th International Conference on Intelligent Information Systems, 2008.
[15] Piotr Andruszkiewicz. Probability distribution reconstruction for nominal attributes in privacy preserving classification. In ICHIT '08: Proceedings of the 2008 International Conference on Convergence and Hybrid Information Technology, pages 494–500, Wash-ington, DC, USA, 2008. IEEE Computer Society.
[16] Piotr Andruszkiewicz. Classification with meta-learning in privacy preserving data min-ing. In Lei Chen, Chengfei Liu, Qing Liu, and Ke Deng, editors, DASFAA Workshops, volume 5667 of Lecture Notes in Computer Science, pages 261–275. Springer, 2009.
[17] Piotr Andruszkiewicz. Privacy preserving classification for ordered attributes. In James F. Peters Urszula Stanczyk´ Krzysztof A. Cyran, Stanisław Kozielski and Alicja Wakulicz-Deja, editors, Man-Machine Interactions, volume 59/2009 of Advances in Soft Computing, pages 353–360. Springer, 2009.
[18] Piotr Andruszkiewicz. Privacy preserving classification with emerging patterns. In Yücel Saygin, Jeffrey Xu Yu, Hillol Kargupta, Wei Wang, Sanjay Ranka, Philip S. Yu, and Xindong Wu, editors, ICDM Workshops, pages 100–105. IEEE Computer Society, 2009.
[19] Piotr Andruszkiewicz. Lazy approach to privacy preserving classification with emerg-ing patterns. In Dominik Ryzko,˙ Piotr Gawrysiak, Henryk Rybinski,´ and Marzena Kryszkiewicz, editors, Emerging Intelligent Technologies in Industry, volume 369 of Studies in Computational Intelligence. Springer, 2011.
[20] Arthur Asuncion and David J. Newman. UCI machine learning repository, 2007.
[21] Mikhail J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, M. Ibrahim, and Vassilios S. Verykios. Disclosure limitation of sensitive rules. In Proceedings of the IEEE Knowledge and Data Engineering Workshop (1999), pages 45–52, 1999.
[22] Yonatan Aumann and Yehuda Lindell. A statistical theory for quantitative association rules. In KDD, pages 261–270, 1999.
[23] Roberto J. Bayardo Jr., Bart Goethals, and Mohammed J. Zaki, editors. FIMI '04, Proceedings of the IEEE ICDM Workshop on Frequent Items et Mining Implementations, Brighton, UK, November 1, 2004, volume 126 of CEUR Workshop Proceedings. CEUR-WS.org, 2004.

Dr.Venkata Reddy Medikonda Developing Suitable Methods for Obtaining Better Accuracy in Privacy Preserving Association Rules Mining and Classification." American Journal of Engineering Research (AJER), vol. 6, no. 11, 2017, pp. 135-154.