International Journal of Computer Science
Oddity...Probe...Reviste...

http://www.ijcsjournal.com Volume 2, Issue 2, No 1, 2014. ISSN: 2348-6600.
Reference ID: IJCS-052. PAGE NO: 297-305.

# *A Dynamic Programming Approach Privacy Preserving Collaborative Data Publishing*

***Ms.S.Saranya***
***PG Student***
***Department of CSE***
***Bharathiyar Institute of Engineering for Women***
***Salem, TamilNadu, India***
**Saranyait777@gmail.com**

***Abstract*— A user wants to store his files in an encrypted form on a remote file server. Later the user wants to efficiently retrieve some of the encrypted files containing (or indexed by) specific keywords, keeping the keywords themselves secret and not jeopardizing the security of the remotely stored files. In this problem under well-defined security requirements and the global distribution of the attributes for the privacy preserving. Our schemes are efficient in the sense that over all efficient attribute table the sensitive attribute in any equivalence class to be distribution of the attribute in the overall records. They are also incremental, in that can submit new files which are secure against previous queries but still searchable against future queries.**

**Keywords- Index Terms—Privacy, security, integrity, and protection, distributed databases.**

## I. Introduction

There is an increasing need for sharing data that contain personal information from distributed databases. For example, in the healthcare domain, a national agenda is to develop the Nationwide Health Information Network (NHIN)1 to share information among hospitals and other providers, and support appropriate use of health information beyond direct patient care with privacy protection.Privacy preserving data analysis, and data publishing [2]–[4] have received considerable attention in recent years as promising approaches for sharing data while preserving individual privacy. In a non-interactive model, a data provider (e.g., hospital) publishes a "sanitized" version of the data, simultaneously providing utility for data users (e.g., researchers), and privacy protection for the individuals represented in the data (e.g., patients). When data are gathered from multiple data providers or data owners, two main settings are used for anonymization [3],[5]. One approach is for each provider to anonymize the data independently (anonymize-and-aggregate, . A more desirable approach is *collaborative data publishing* [3],[5]–[7], which anonymizes data from all providers as if they would come from one source (aggregate-and-anonymize, Distributed data publishing settings for four providers.lishing setting with horizontally distributed data across multiple data providers, each contributing a subset of records *Ti*. Each record has an owner, whose identity should be protected. Each record attribute is either *an identifier*, which directly identifies the owner, or *a quasiidentifier* (QID), which may identify the owner if joined with a publicly known dataset, or a sensitive attribute,which should be also protected. As a special case, a data provider could be the data owner itself who is contributing its own records. A data recipient may have access to some background knowledge , which represents any publicly available information about released data, e.g.,Census datasets.Our goal is to publish an anonymized view of the integrated data, $T*$, which will be immune to attacks. Attacks are run by *attackers*, i.e., a single or a group (*a coalition*) of external or internal entities that wants to breach privacy of data using background knowledge, as well as anonymized data. Privacy is breached if one learns anything about data.

**Existing Solutions.** Collaborative data publishing can be considered as a multi-party computation problem, in which multiple providers wish to compute an anonymized view of

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 1, 2014. ISSN: 2348-6600.
Reference ID: IJCS-052. PAGE NO: 297-305.

their data without disclosing any private and sensitive information. We assume the data providers are semihonest [8], [9], commonly used in distributed computation setting. A trusted third party (TTP) or Secure Multi-Party Computation (SMC) protocols [6] can be used to guarantee there is no disclosure of *intermediate* information *during* the anonymization. However, neither TTP nor SMC protects against inferring information using the anonymized data.The problem of inferring information from anonymized data has been widely studied in a single data provider settings [3]. A data recipient that is an attacker, e.g., *P*0, attempts to infer additional information about data records using the published data, $T*$, and background knowledge, *BK*. For example, *k*-anonymity [10], [11] protects against identity disclosure attacks by requiring each quasiidentifier equivalence group (QI group) to contain at least *k* records. *l*-Diversity requires each QI group to contain at least *l* "well-represented" sensitive values [12]. Differential privacy [2], [4] guarantees that the presence of a record cannot be inferred from a statistical data release with little assumptions on an attacker's background knowledge.

**New Challenges.** Collaborative data publishing introduces a new attack that has not been studied so far. Compared to the attack by the external recipient in the second scenario, each provider has additional data knowledge of its own records, which can help with the attack. This issue can be further worsened when multiple data providers collude with each other.In the social network or recommendation setting, a user may attempt to infer private information about other users using the anonymized data or recommendations assisted by some background knowledge and her own account information. Malicious users may collude or even create artificial accounts as in a shilling attack [13]. Assume that hospitals *P*1, *P*2, *P*3,and *P*4 wish to collaboratively anonymize their respective patient databases *T*1, *T*2, *T*3, and *T*4. In each database,**Name** is an identifier, *{***Age**, **Zip***}* is a quasi-identifier (QI),and **Disease** is a sensitive attribute. Note that one record, owned by Olga, is contributed by two providers *P*2 and *P*4, and is represented as a single record in anonymized dataset. $T*$ *a* is one possible anonymization that guarantees *k*-anonymity and *l*-diversity ($k = 2$, $l = 2$), i.e., each QI group contains records with at least *l* different sensitive values. However, an attacker from the hospital *P*1 may remove all records from *P*1. In the first QI group there will be only one remaining record, which belongs to a patient between 20 and 30 years old. By joining this record with the background knowledge *BK* (e.g., part of the Census database) using quasi-identifier attributes, *P*1 can identify Sara as the owner of the record (highlighted in the table) and her disease Epilepsy. In practice, the attacker would use more attributes as a QI and maximal *BK* to mount the linking attack [14]. In general, multiple providers may collude with each other, hence having access to the union of their data, or a user may have access to multiple databases, e.g., a physician switching to another hospital, and using information about her former patients.

**Contributions.** We define and address this new type of insider attack" by data providers in this paper. In general, we define an *m-adversary* as a coalition of *m* colluding data providers or data owners, and attempts to infer data records contributed by other data providers. Note that 0-adversary models the external data recipient, who has only access to the external background knowledge. Since each provider holds a subset of the overall data, this inherent data knowledge has to be explicitly modeled, and considered when the data are anonymized.We address the new threat introduced by *m*-adversaries,

and make several important contributions. First, we introduce the notion of *m*-privacy that explicitly models the inherent data knowledge of an *m*-adversary, and protects anonymized data against such adversaries with respect to a given privacy constraint. For example, $T*$ *b* is an anonymized table that satisfies *m*-privacy ($m = 1$) with respect to *k*-anonymity and *l*-diversity ($k = 2$, $l = 2$).Second, for scenarios with a TTP, to address the challenges of checking a combinatorial number of potential *m*adversaries,we present heuristic algorithms for efficiently verifying *m*-privacy given a set of records. Our approach utilizes effective pruning strategies exploiting the equivalence group monotonicity property of privacy constraints and adaptive ordering techniques based on a novel notion of privacy fitness.We also present a data *provider-aware* anonymization algorithm with adaptive strategies of checking *m*-privacy, to ensure high utility and *m*-privacy of sanitized data with efficiency. Compared to our preliminary version [1], our new contributions

extend above results. First, we adapt privacy verification and anonymization mechanisms to work for *m*-privacy w.r.t. to any privacy constraint, including nonmonotonic ones. We list all necessary privacy checks and prove that no fewer checks is enough to confirm *m*-privacy.Second, we propose SMC protocols for secure *m*-privacy verification and



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 1, 2014. ISSN: 2348-6600.
Reference ID: IJCS-052. PAGE NO: 297-305.

anonymization. For all protocols we prove their security, complexity and experimentally confirm their efficiency.

## 2. *M*-Privacy Definition

We first formally describe our problem setting. Then, we present our *m*-privacy definition with respect to a privacy constraint to prevent inference attacks by *m*-adversary,followed by properties of this new privacy notion.Let $T = \{t1, t2, \ldots\}$ be a set of records with the same attributes gathered from *n* data providers $P = \{P1, P2, \ldots, Pn\}$, such that $Ti \subseteq T$ are records provided by *Pi*. Let *AS* be a sensitive attribute with a domain *DS*.If the records contain multiple sensitive attributes then,we treat each of them as the sole sensitive attribute, while remaining ones we include to the quasi-identifier [12].However, for our scenarios we use an approach, which preserves more utility without sacrificing privacy [15].Our goal is to publish an anonymized table $T*$ while preventing any *m*-adversary from inferring *AS* for any single record. An *m*-adversary is a coalition of data users with *m* data providers cooperating to breach privacy of anonymized records.

## 3. Verification Of *M*-Privacy

Checking whether a set of records satisfies *m*-privacy creates a potential computational challenge due to the combinatorial number of *m*-adversaries. In this section,we first analyze the problem by modeling the adversary space. Then, we present heuristic algorithms with effective pruning strategies and adaptive ordering techniques for efficiently checking *m*-privacy w.r.t. an EG monotonic constraint *C*. Implementation of introduced algorithms can be run by a trusted third party (TTP). For scenarios without such party, we introduce secure multi-party (SMC) protocols.Finally, in Appendix B.1 we present modifications of TTP heuristics and SMC protocols to verify *m*-privacy w.r.t.non-EG monotonic privacy constraints.

### 3.1 Adversary Space Enumeration

Given a set of *nG* data providers, the entire space of *m*adversaries (*m* varying from 0 to $nG-1$) can be represented using a lattice. Each node at layer *m* represents an *m*-adversary of a particular combination of *m* providers. The number of all possible *m*-adversaries is given by $\_nG\ m\ \_$. Each node has parents (children) representing their direct super- (sub-) coalitions. For simplicity the space is depicted as a *diamond*, where a horizontal line at a level *m* corresponds to all *m*-adversaries, the bottom node to 0-adversary (external data recipient), and the top line to $(nG - 1)$-adversaries. In order to verify *m*-privacy w.r.t. a constraint *C* for a set of records, we need to check fulfillment of *C* for all records after excluding any possible subset of *m*-adversary records. When *C* is EG monotonic, we only need to check *C* for the records excluding all records from any *m*-adversary (Observation 2.3), i.e., adversaries on the horizontal line.Given an EG monotonic constraint, a *direct* algorithm can sequentially generate all possible $\_nG\ m\ \_$ *m*-adversaries,and then check privacy of the corresponding remaining records. In the worst-case scenario, when $m = nG/2$, the number of checks is equal to the central binomial coefficient $\_ nG\ nG/2\_ = O(2nGn-1/2\ G\ )$. Thus, the *direct* algorithm is not efficient enough.

### 3.2 Heuristic Algorithms for EG Monotonic Constraints

In this section, we present heuristic algorithms for efficiently checking *m*-privacy w.r.t. an EG monotonic constraint.Then, we modify them to check *m*-privacy w.r.t.a non-EG monotonic constraint.The key idea of our heuristics for EG monotonic privacy constraints is to efficiently search through the adversaryspace with effective pruning such that not all *m*-adversaries need to be checked. This is achieved by two different pruning strategies, an adversary ordering technique, and a set of search strategies that enable fast pruning.

**Pruning Strategies.** The pruning is possible thanks to the EG monotonicity of *m*-privacy .If a coalition is not able to breach privacy, then all its subcoalitions will not be able to do so as well, and hence do not need to be checked (downward pruning). On the other hand, if a coalition is able to breach privacy, then all its super-coalitions will be able to do so as well, and hence do not need to be checked (upward pruning). In fact, if a sub-coalition of an *m*-adversary is able to breach privacy, then the upward pruning allows the algorithm to terminate immediately as the *m*-adversary will be able to breach privacy (*early stop*).

**Adaptive Ordering of Adversaries.** In order to facilitate the above pruning in both directions, we adaptively order the coalitions based on their attack powers . This is motivated by



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 1, 2014. ISSN: 2348-6600.
Reference ID: IJCS-052. PAGE NO: 297-305.

following observations. For downward pruning, super-coalitions of *m*-adversaries with limited attack powers are preferred to be checked first as they are less likely to breach privacy, and hence increase the chance of downward pruning. In contrast, sub-coalitions of *m*-adversaries with significant attack powers are preferred to be checked first as they are more likely to breach privacy, and hence increase the chance of the early stop.(a) Adaptive ordering. (b) First steps of the *binary* algorithm with verified coalitions depicted as numbered red dots.To quantify privacy fulfillment by a set of records, which is used to measure the attack power of a coalition and privacy of remaining records, we introduce a privacy fitness score w.r.t. *C*. It also used to facilitate the anonymization, which we will discuss in the following section. The privacy fitness score quantifies also the attack power of attackers. The higher their privacy fitness scores are,

the more likely they are able to breach the privacy of the remaining records. In order to maximize the benefit of both pruning strategies, the super-coalitions of *m*-adversaries are generated in the order of ascending fitness scores (ascending attack powers), and the sub-coalitions of *m*adversaries are generated in the order of descending fitness scores . Now we present several heuristic algorithms that use

different search strategies, and hence utilize different pruning directions. All of them use the adaptive ordering of adversaries to enable fast pruning.

**The Top-Down Algorithm.** The *top-down* algorithm checks the coalitions in a top-down fashion using downward pruning, starting from $(nG - 1)$-adversaries, and moving down until a violation by an *m*-adversary is detected or all *m*-adversaries are pruned or checked.

**The Bottom-Up Algorithm.** The *bottom-up* algorithm is similar to the *top-down* algorithm. The main difference is in the sequence of coalition checks, which is in a bottom up fashion starting from 0-adversary, and moving up. The algorithm stops if a violation by any adversary is detected (early stop) or all *m*-adversaries are checked.

**Algorithms.** Each of the above algorithms focuses on different search strategy, and hence utilizes different pruning. Which algorithm to use is largely dependent on the characteristics of a given group of providers.

Intuitively, the privacy fitness score, which quantifies also the level of privacy fulfillment of the group,may be used to select the most suitable algorithm. The higher the fitness score, the more likely *m*-privacy will be satisfied, and hence the *top-down* algorithm with downward pruning will significantly reduce the number of adversary checks. We utilize such strategy in the anonymization algorithm (discussed later), and experimentally evaluate it.

## II Proposed System

### 4. Correlation Based Filter Approach

In this section, we discuss how to evaluate the goodness of features for classification. In general, a feature is good if it is relevant to the class concept but is not redundant to any of the other relevant features. If we adopt the correlation between two variables as a goodness measure, the above definition becomes that a feature is good if it is highly correlated to the class but not highly correlated to any other features. In other words, if the correlation between a feature and the class is high enough to make it relevant to (or predictive of) the class and the correlation between it and any other relevant features does not reach a level so that it can be predicted by any of the other relevant features, it will be regarded as a good feature for the classification task.

There exist broadly two approaches to measure the correlation between two random variables. One is based on classical linear correlation and the other is based on information theory. Under the first approach the most well-known measure is linear correlation coefficient. There are several benefits of choosing linear correlation as a feature goodness measure for classification.

First, it helps removes feature with near zero linear correlation to the class. Second, it helps to reduce redundancy among selected features. It is known that if data is linearly separable if all but one of a group of linearly dependent features is removed. However it is not safe to always assume linear correlations that are not linear in nature. Another limitation is that the calculation requires all features contain numerical values.

To overcome these shortcomings, in our solution we adopt the other approach and choose a correlation measure based on the information theoretical concept of entropy, a measure of the



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 1, 2014. ISSN: 2348-6600.
Reference ID: IJCS-052. PAGE NO: 297-305.

uncertainty of a random variable. Using symmetrical uncertainty (SU) as the goodness measure, we are now ready to develop a procedure to select good features for classification based on correlation analysis of features for (including the class). This involves two aspects (1) how to decide whether a feature is relevant to the class or not; and (2) how to decide whether such a relevant feature is redundant or not when considering it with other relevant features.

The answer to the first question can be using a user defined threshold SU value, as the method used by many other feature weighting algorithms (e.g., Relief). More specifically, suppose a dataset S contains N features and C class. Let $SU_{ic}$ denote the SU value that measures the correlation between a feature Fi and class C (named C-correlation), then a subset S' of relevant features can be decided by a threshold SU value

The answer to the second question is more complicated because it may involve analysis of pairwise correlation between all features (named F-correlation), which results in time complexity of $0(N^2)$ associated with the number of features N for most existing algorithms.

Since F-correlation is also captured by SU values, in order to decide whether a relevant feature is redundant or not, we need to find a reasonable way to decide the threshold level for F-correlations as well. In other words, we need to decide whether the level of correlation between two features in S' is high enough to cause redundancy so that one of them may be removed from S'. In the context of a set of relevant features S' already identified for the class concept, when we try to determine the highly correlated features for a given feature Fi and the class concept, $SU_{i,c}$ as a reference. Therefore, even the correlation between this feature and the class concept is larger than some threshold value and therefore making this feature relevant to the class concept, this correlation is by no means predominant.

**Definition4**: (Predominant Feature). A feature is predominant to the class, if its correlation to the class is predominant or can become predominant after removing its redundant peers. According to the above definitions, a feature is good if it is predominant in predicting the class concept, and feature selection for classification s a process that identifies all predominant features and remove redundant ones among all relevant features, without having to identify all the redundant peers for every feature in S', and thus avoids pairwise analysis of F-correlation between all relevant features.

## 5. Empirical Study

We run two sets of experiments for *m*-privacy w.r.t. *C* with the following goals: 1) to compare and evaluate the different *m*-privacy verification algorithms, and 2) to evaluate and compare the proposed anonymization algorithm with the baseline algorithm in terms of both utility and efficiency.

All experiments have been run for scenarios with a trusted third party (TTP), and without it (SMC protocols). Due to space restrictions all experiments for a TTP setting are in the previous version of the paper [1] and in Appendix C.

### 5.1 Experimental Setup

We merged the training and testing sets of the Adult dataset2. Records with missing values have been removed.All remaining 45,222 records have been randomly distributed among *n* providers. As a sensitive attribute *AS* we chose *Occupation* with 14 distinct values.To implement SMC protocols, we have enhanced the SEPIA framework [21], which utilizes Shamir's secret sharing scheme [19]. Security of communication is guaranteed by the SSL using 128-bit AES encryption scheme. For the secure *l*-diversity protocol we have used commutative Pohlig-Hellman encryption scheme with a 64-bit key [29].

**Privacy Constraints.** The EG monotonic privacy constraint is defined as a conjunction of *k*-anonymity [11] and *l*diversity [12]. Experiment settings and default values of SMC protocols.**Name Description Verification Anonymization** *m* Power of *m*-privacy 3 3 *n* Number of data providers – 10 *nG* Number of data providers contributing to a group 10 – |*T*| Total number of records – 1000 |*TG*| Number of records in a group 150 –*k* Parameter of *k*-anonymity 30 30 *l* Parameter of *l*-diversity 3 3 All experiments have been performed on the local network of 64 HP Z210 with 2 quad-core CPUs, 8 GB of RAM, and running Ubuntu 2.6 each. The efficiency of protocols is measured by their computation time.

### 5.2 Experimental Procedures

In order to make the best use of the data and obtain stable results,(M=5)*(N=10) cross validation strategy is used. That is, for each dataset, each classification algorithm, the 10-fold cross validation is repeated M=5 times, with each time the order of the instances of the dataset being randomized. Randomizing the order of the inputs can help diminish the



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 1, 2014. ISSN: 2348-6600.
Reference ID: IJCS-052. PAGE NO: 297-305.

order effects. In the experiment, for each feature subset selection algorithm, we obtain M N feature subset and the corresponding runtime time with each dataset .For each classification algorithm, we obtain M N classification accuracy for each feature selection algorithm and each dataset.

The Friedman test [24] can be used to compare k algorithms over N datasets by ranking each algorithm on each set separately. The algorithm obtained the best performance gets the rank of 11, the second best ranks 2, and so on. Then the average ranks of all algorithms on all data sets are calculated and compared. The Nemenyi test [21] compares classifiers in a pairwise manner. According to this test, the performances of two classifiers are significantly different if the distance of the average ranks exceed the critical distances. It is known that if data is linearly separable if all but one of a group of linearly dependent features is removed. However it is not safe to always assume linear correlations that are not linear in nature. Using symmetrical uncertainty (SU) as the goodness measure, we are now ready to develop a procedure to select good features for classification based on correlation analysis of features for (including the class).Another limitation is that the calculation requires all features contain numerical values.

Procedure Experimental Proces

```
1  M=5, N=10
2  DATA = {D1, D2, .., D35}
3  Learner= {NB, C4.5, IBI, RIPPER}
4  Feature Selector = {FAST, FCBF, Relief, CFS, Consist, FOCUS-SF}
5  for each data     DATA do
6      for each times     [1, N] do
7        randomize instance order for data
8       generate N bins from the randomized data
9        for each fold     [1, N] do
10         Test data=bin [fold]
11         training data=data – Test Data
12         for each selector     feature selectors do
13       (Subset, Time) = selector (Training Data)
14          Training Data' = select Subset from Training
15       Test Data'=select subset from test data
16            for each Learner     Learners do
17              classifier=learners (Training Data')
18              accuracy=apply classifier to Test Data'
```

### 5.3 Result and Analysis

In this section we present the experimental results in terms of the proportion of selected features, the time to obtain the feature subset, the classification accuracy.Generally all the six algorithms achieve significant reduction of dimensionality by selecting only a small portion of the original features. Fast on average obtains the best proportion of selected features.

In order to further explore feature selection algorithms whose reduction rates have statistically significant differences, we performed a Nemenyi test.The results indicate that the proportion of selected features of FAST is statistically smaller than those of Relief-F, CFS andFCBF, and there is no consistent evidence to indicate statistical difference between FAST, Consist, and FOCUS-SF, respectively.For real world data, we often do not have such prior knowledge about the optimal subset, so we use the predictive accuracy on the selected subset of feature.

#### 5.3.1Runtime

Table 1:Runtime (in ms) of the six feature selection algorithms

Dataset FAST FCBF CFS Relief Consist FOCUS-SF



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 1, 2014. ISSN: 2348-6600.
Reference ID: IJCS-052. PAGE NO: 297-305.

| | | | | | | |
|---|---|---|---|---|---|---|
| Chess | 105 | 60 | 345 | 12660 | 1990 | 653 |
| Coil | 1472 | 716 | 938 | 13918 | 3227 | 660 |
| Elephant | 866 | 875 | 1483 | 30416 | 53845 | 1282 |
| Arrhy | 783 | 312 | 905 | 1072 | 3492 | 1098 |
| Colon | 166 | 115 | 821 | 744 | 1360 | 2940 |
| Ar10p | 706 | 458 | 736 | 7945 | 1624 | 1032 |
| Pie10p | 678 | 1223 | 1224 | 3874 | 57934 | 960 |
| Oh0.wc | 5283 | 5990 | 6650 | 7636 | 3568 | 4689 |
| Oh10.wc | 5549 | 6033 | 1034 | 4898 | 4149 | 8446 |
| B-cell1 | 160 | 248 | 9345 | 5652 | 4882 | 3421 |
| b-cell2 | 626 | 1618 | 4524 | 4166 | 5102 | 1273 |
| b-cell3 | 635 | 2168 | 3415 | 5102 | 2914 | 8446 |
| Average | 3573 | 4671 | 5456 | 4580 | 7490 | 5227 |
| Win/draw/ loss | - | 22/0/3 | 31/0/1 | 20/0/6 | 35/0/0 | 34/0/1 |

Generally the individual evaluation based feature selection algorithms of FAST, FCBF and Relief are much faster than the subset evaluation based algorithms of CFS, Consist and FOCUS-SF.FAST is consistently faster than all other algorithms. The runtime of FAST is only 0.1 % of that of CFS 2.4 % of that of Consist, 2.8 % of that of FOCUS-SF, 7.8% of that of Relief, and 76.5% of that of FCBF, respectively. The Win/Draw/Loss records show that FAST outperforms other algorithms as well.

From the analysis above we can know that FAST performs very well on the microarray data. The reason lies in both the characteristics of the dataset itself and the property of the proposed algorithm. Microarray data has the nature of the large number of features (genes) but small sample size, which can cause "curse of dimensionality" and over-fitting of the training data [22]. For the purpose of exploring the relationship between feature selection algorithms and data types, i.e. which algorithm are more suitable for which types of data, we rank the six feature selection algorithms according to the classification accuracy of the feature selection method. Therefore, selecting a small number of discriminative genes from thousands of genes is essential for successful sample classification [21], [24].

Our proposed FAST effectively filters out a mass of irrelevant features in the first step. This reduces the possibility of improperly bringing the irrelevant features into the subsequent analysis. Then in the second step FAST removes redundant features by using the redundant filtering mechanism. Our proposed FAST also requires a parameter that is the threshold of feature relevance. Different values might end with different classification results.

## 6. Conclusion

A solution for accomplishing anonymity when data from two organizations with common privacy policy are included. The explanation is a unassuming and effective technique as it uses T-closenessalgorithms for achieving anonymity. The solution projected in is constructed on tree data structure called TIPS. We base our solution on subset generation and selecting the most relevant subset. We are currently examining the feasibility of this approach for achieving anonymity on the fly in dynamically growing databases.

All algorithms have been implemented in distributed settings with a TTP and as SMC protocols. All protocols have been presented in details and their security and complexity has been carefully analyzed. Implementation of algorithms for the TTP setting is available on-line for further development and deployments. There are many potential research directions. For example, it remains a question to model and address the data knowledge of data providers when data are distributed in a vertical or ad-hoc fashion. It would be also interesting to investigate if our methods can be generalized to other kinds of data such as set-valued data.

## References



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 1, 2014. ISSN: 2348-6600.
Reference ID: IJCS-052. PAGE NO: 297-305.

[1] M.A. Hall.. "Correlation-Based Feature Selection for Discrete and Numeric class Machine Learning", In Proceeding of 17th International Conference on Machine Learning, pp 359-366,2000.

[2]D.A. Bell and H. Wang., "Formalism for Relevance and its Application in Feature Subset Selection," Machine Learning, 41(2), pp 175-195,2000.

[3] M. Dash, H. Liu and H. Motoda ., "Consistency based Feature Selection", In Proceeding of the fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109,2000.

[4] S. Das., "Filters, Wrappers And A Boosting Based Hybrid For Feature Selection", In Proceeding Of The 18th International Conference On Machine Learning, pp 74-81, 2001.

[5] I.S Dhillon., S .Mallela. andR .Kumar., "A Divise Information Theoretic Feature Clustering Algorithm For Text Classification," J. Mach. Learn. Res., 3,, pp 1265-1287, 2003.

[6] F .Fleuret., "Fast Binary Feature Selection with Conditional Mutual Information, Journal of Machine" Learning Research, 5,Pp 1531-1555, 2004.

[7] G. Forman., "An Extensive Empirical Study of Feature Selection Metrics for Text Classification".Journal of Machine Learning Research, 3, pp 1289-1305, 2003.

[8] F. Herrera, and S. Garcia., "An Extension on Statistical Comparison of Classifiers over Multiple Datasets for All Pairwise Comparisons". J.Mach. Learn. Res., 9,pp 2677-2694, 2008.

[9]I. Guyn .and A. Elisseeff ., "An Introduction to Variable and Feature Selection". Journal of Machine Learning Research, 3 pp 1157-1182, 2003.

[10] M.A Hall ., "Correlation Based Feature Subset Selection for Machine Learning", Ph.D Dissertation Waikato, New Zealand: Univ. Waikato,1999.

[11] C. Krier C. D. Francois., F. Rossi, and M. Verleysen., "Feature Clustering And Mutual Information For The Selection Of Variables In Spectral Data". In Proc European Symposium On Artificial Neural Networks Advances In Computational Intelligence And Learning, pp 157-162, 2007.

[12] M. Last., A. Kandel. AndO. Mainmom., "Information Theoretic Algorithm For Feature Selection, Pattern Recongnitition Letters", 22(6-7),pp 799-811,2001.

[13] L.C Monila,L. Belanche. And A. Nebot., "Feature Selection Algorithms: A Survey and Experimental Evaluation", In Proc. IEEE Int.Conf. Data Mining.,pp 306-313, 2002.

[14] H. Park, and H. Kwon., "Extended Relief Algorithm In Instance Based Feature Filtering", In Proceeding Of The Sixth International Conference On Advanced Language Processing And Web Information Technology (ALPTI 2007), pp 123-128, 2007.

[15]B. Raman. And T.R Loerger,,"Instance Based Filter for Feature SelectionJournal of Machine Learning Research, 1, pp 1-23, 2002.

[16] M. Robnik-Sikonja. andI. Kononenko., "Theoritic and Empirical Analysis of Relief and Relief", Machine Learning Research, 53, pp 23-69, 2003.

[17] P. Scanlon., G. Potamiano., "Mutual information based visual feature selection for lip-reading, in int. conf. on spoken language processing, 2004.

[18] Scherf M. and Brauer W., Feature Selection By Means Of A Feature Weighting Approach, Technical Report FKI-221-97, Institute Fur Informatics, and Technics Universidad Munched 1997.

[19] C. Sha, X.Quiu. And A. Zhou., "Feature Selection Based On A New Dependency Measure," 2008 Fifth International Conference On Fuzzy Systems And Knowledge Discovery, 1, pp 266-270, 2008.

[20] J. Souza., "Feature Selection with A General Hybrid Algorithm, Ph.D, University Of Ottawa, Ottawa, Ontario, Canada, 2004.

[21] G. Van Dijk. andM.M Van Hulle .," Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis,



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 1, 2014. ISSN: 2348-6600.
Reference ID: IJCS-052. PAGE NO: 297-305.

"International Conference on Artificial Neural Networks, 2006.

[22]G.I Webb.,"Multi boosting for Combining Boosting and Wagging,Machine Learning, 40(2), pp 159-196, 2000.

[23] E. Xing., M. Jordan. And R. Karp, "Feature Selection for High Dimensional Genomic Microarray Data, "In Proceedings of the Eighteenth International Conference on Machine Learning, pp 601-608, 2008.

[24] J. Yu., S.S.R. Adipi. And P.H Artes., "A Hybrid Feature Selection Strategy For Image Defining Features: "Towards Interpretation Of Optic Nerve Images, In. Proceedings Of 2005 International Conference On Machine Learning And Cybernetics, 8, pp 5127-5132, 2005.

[25] L. Yu. And H. Liu H., "Feature Selection For High Dimensional Data: Fast Correlation Based Filter Solution, "In Proceedings of 20th International Conferences On Machine Learning, 20(2), pp 856-863,2003.



Published by SK Research Group of Companies (SKRGC).