Privacy-Utility Feature Selection as a Privacy Mechanism in Collaborative Data Classification

Mina Sheikhalishahi, Fabio Martinelli
Istituto di Informatica e Telematica Consiglio Nazionale delle Ricerche (IIT-CNR), Pisa, Italy
Email: name.surname@iit.cnr.it

***Abstract*—This paper presents a novel framework for privacy aware collaborative information sharing for data classification. Data holders participating in this information sharing system, for global benefits are interested to model a classifier on whole dataset, but are ready to share their own table of data if a certain amount of privacy is guaranteed.**

To address this issue, we propose a privacy mechanism based on privacy-utility feature selection, which by eliminating the most irrelevant set of features in terms of accuracy and privacy, guarantees the privacy requirements of data providers, whilst the data remain practically useful for classification. Due to the fact that the proposed trade-off metric is required to be exploited on whole dataset, a distributed secure sum protocol is utilized to protect information leakage in each site. The proposed approach is evaluated and validated through standard *Tumor* dataset.

***Index Terms*—Privacy, Utility, Accuracy, Feature Selection, Classification, Collaborative, Distributed, Information Sharing;**

I. Introduction

Facing the new challenges brought by a continuous evolving Information Technology (IT) market, large companies and small-to-medium enterprises found in *Information Sharing* a valid instrument to improve their key performance indexes. Sharing data with partners, authorities for data collection and even competitors, may help in inferring additional intelligence through collaborative information analysis [1] [2] [3]. Such an intelligence could be exploited to improve revenues, e.g. through best practice sharing [4], market basket analysis [5], or prevent loss coming from brand-new potential cyber-threats [6]. Other applications include analysis of medical data, provided by several hospitals and health centers for statistical analysis on patient records, useful, for example, to shape the causes and symptoms related to a new pathology [7].

Independently from the final goal, unfortunately information sharing brings issues and drawbacks which must be addressed. These issues are mainly related to the information privacy. Shared information may contain the sensitive information, which could be potentially harming the privacy of physical persons, such as employee records for business applications, or patient records for medical ones. Hence, the most desirable strategy is the one which enables data sharing in secure environment, such that the individual privacy requirement is satisfied, while at the same time data are still practically useful.

In present work, we assume that the data providers are interested to model a classifier on shared data. However, for privacy concerns they admit to share their own dataset only if some level of privacy is guaranteed. To this end, we propose a privacy-aware feature selection framework which by secure removing the most *irrelevant* features, from all datasets, increases privacy gain while slightly modifies the data utility.

Generally, *feature selection* is based on the notion that a subset of features from the input can effectively describe the data [8]. This means that the information content can be obtained from the smaller number of variables which represent more discrimination information about the classes. On the other side, removing a set of variables increases the uncertainty of new dataset comparing to the original one, i.e. it increases the *privacy gain*. The rational behind it comes simply from the fact that each feature carries some information about data, such that by its removal the data become less indicative. However, the optimum set of features in terms of data accuracy are not necessarily equivalent to the best set of feature in terms of data privacy. To this end, we propose a feature selection approach based on a trade-off between *feature utility* (class discrimination) and *feature privacy* (record discrimination).

In this study, since we assume more than two parties are involved, it is required that some secure multi-party computation protocols be exploited. To this end, we utilize a distributed secure sum protocol to find the proper subset of features where the required privacy of each data provider is preserved, and the modified shared information sustain data utility for classification.

The contribution of this paper can be summarized as follows: 1) We propose privacy-utility feature selection as a privacy mechanism in a distributed architecture; 2) We present in detail the formulas and algorithm for secure computation of total accuracy and privacy, when a feature is removed from distributed parties' datasets; 3) Finally, we evaluate and validate the proposed methodology on a standard dataset.

The rest of the paper is structured as follows. Section III presents some preliminary notations exploited in this study. Section IV proposes the privacy-utility trade-off score. Section V describes the proposed framework, detailing the secure computation of total utility score and average privacy score. Section VI reports the experimental analysis to evaluate the effectiveness of proposed approach on real data. Related work is presented in Section II. Finally Section VII briefly concludes proposing future research directions.

II. Related Work

Several work have been devoted in secure feature selection in multiparty data analysis framework. As a remarkable study, in [14], a secure distributed protocol is proposed which allows feature selection for multiple parties without revealing their own dataset. The proposed approach is based on *virtual dimension reduction* for selecting the subset of features in image processing. In this paper, the methodology is designed to privately select the most appropriate subset of features. However, differently from our approach, privacy gain does not come under consideration in feature selection. Moreover, the dimension reduction technique is exploited for unsupervised algorithms, e.g clustering, rather than classification. In [15] feature selection is combined with anonymization techniques to remove redundant features in order to publish a secure dataset. The result is evaluated through computing the accuracy of classification on original and sanitized datasets applying UCI benchmark datasets. The authors show that in some cases the accuracy even improves comparing to the original dataset. However, in this work, differently from our approach, the dataset is centralized, and just one data holder is involved. In [16] a framework is proposed to select the few features which contain maximum discriminative information of classes. The classification accuracy and privacy gain are combined in feature selection process. However, the proposed approach, differently from our methodology, does not consider the case of privacy-aware feature selection in distributed parties, but in centralized architecture. In [17] the optimum subset of features in terms of privacy-utility trade-off is selected when data is partitioned between two data holders. Secure two-party computation protocols are exploited to find the desired set of features. Two-party protocols are not applicable when data is divided among more than two parties.

To the best of our knowledge, it is among the very first work which incorporates the usefulness and the amount of privacy that a feature carries to shape optimum feature selection as a service when data are desired to be shared to model a classifier.

III. Preliminary Notations

In the following, we present the *distributed secure sum protocol* proposed in [9], proven to be resistant against colluding.

Assume that N parties $P_1, P_2, \ldots, P_N$ involve in a cooperative secure sum computation, where each party is able to break his private number into a fixed number of segments, such that the addition of segments is equal to her private number. In the proposed protocol the number of segments is equal to the number of parties (N). The values of each segment is randomly selected by the associated party and it is secret from other parties. Then, each party holds one segment of her data and sends $N-1$ other segments to the other $N-1$ parties. In this way, at the end each party holds N segments, where only one belongs to the party and the others are collected from remaining parties, one from each. Now, the *secure sum protocol* can be applied to obtain the sum of all the segments. According to this protocol, one of the parties is required to be selected as the protocol initiator party that starts the computation by sending the data segment to the next party in the ring. The receiving party adds his data segment and to the received partial sum and then sends the result to the next party in the ring. This process is repeated till all the segments of all the parties are added and the sum is announced by the initiator party. For the sake of simplicity, in the rest of this paper, we represent the call of *distributed secure sum protocol* on N numbers x_i $(1 \leq i \leq N)$ as $\mathcal{DSS}_{i=1}^{N} x_i$, where without loss of generality x_1 is the *initiator*.

IV. Privacy-Utility Trade-off Score

In what follows we present the general definitions of *feature utility, feature privacy,* and *privacy-utility trade-off* scores.

a) ***Feature Utility Score:*** The aim of *feature selection* in data mining algorithms is to obtain an *optimal* set of features such that it contains all *relevant* features which are not *redundant* in terms of identifying the class labels of a dataset [8]. There are many potential benefits of feature selection, spanning from facilitating data visualization and understanding, reducing the measurement and storage requirements, reducing training and utilization times, to defying the curse of dimensionality to improve prediction performance [10]. The *filter* techniques in feature selection, with the use of general characteristics of the data, rank the features based on their relevance. More precisely, *feature ranking* creates a scoring function, say $\upsilon(j)$, computed from the impact of a feature A_j in discriminating class labels. Generally, a high score is indicative of a valuable feature. In this study, any feature ranking function satisfying the following properties can be utilized to measure *feature utility score*: 1) it can be measured through statistical information of a dataset; 2) it returns a real number in $[0,1]$, such that the higher score for a feature indicates that the feature is more relevant in identifying the class labels.

b) ***Feature Privacy Score:*** Data privacy is quantified as the degree of uncertainty that the original data can be inferred from the sanitized one [11]. Generally, reducing the information which a dataset carries will increase privacy gain. Hence, removing a feature completely from a dataset can be considered as a tool which increases this uncertainty.

In this study any function having the following properties can be utilized to measure the feature privacy score: 1) it can be measured through statistical information of a dataset; 2) it returns a real number in $[0,1]$, such that the higher privacy score for a feature (a set of features) indicates that the more privacy is obtained by removing that feature (those features).

c) ***Privacy-Utility Trade-off Score:*** The typical feature selection approaches exploit the impact of features in discriminating the class labels. However, the optimal set of features in terms of classification accuracy, is not necessarily equivalent to the optimal set of features in terms of privacy gain. Hence, a trade-off metric between privacy gain and data utility is in high demand when both criteria are vital.

Generally *ranking feature selection* approach computes a score for each feature, named *utility score*, which represents the relevance of the feature in characterizing the classes. The features with the lower scores are the candidates to be removed first [8]. On the other hand, removing a feature completely from a dataset can be considered as a tool which increases the degree of uncertainty, i.e. privacy gain, in a dataset. On the base of the aforementioned concepts, we define a *trade-off score* as the output of a function of privacy and utility scores of a feature, in a way that the trade-off score increases when utility and privacy scores are maximized. Hence, the best feature to be removed first, is the one with the minimum privacy-utility trade-off score, which has negligible effect in discriminating class labels and in preserving the privacy.

Several functions and optimization models match the aforementioned definition and can be used according to the preferences of the specific system implementation. In particular it is possible to use the trade-off score functions where privacy gain and utility score have the same impact on the result; otherwise it is possible to give different weights to the two values. Formally, we represent privacy-utility trade-off score of a feature A_j with $\tau(\upsilon(A_j),\rho(A_j))$, as a function of feature's utility and privacy scores.

The problem of privacy-utility feature selection becomes more challenging when the data is distributed across different parties, and the parties are not interested to share their original datasets unless that some level of privacy is guaranteed. Under this condition, applying the above strategy, separately on each dataset, may lead to different sets of published features for each data holder. Therefore, it is vital to find *securely* a *homogenized* optimum set of features, which provides the highest possible data utility and privacy gain for all parties.

V. Problem Statement

Let consider that $N(>2)$ data holders are interested to shape a classifier on whole of their datasets. It is assumed that the data are distributed *horizontally* among parties. This means that each data holder involved in data sharing has information about all the features but for different collection of objects.

In this scenario, the set of features applied to describe the records are known beforehand. Let $\mathcal{A} = \{A_1, A_2, \ldots, A_t\}$ be the set of t features all used to express each record of data, and the class labels come from the set $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$. Therefore, each record is a $t+1$ dimensional vector $z_i = (a_{i1}, a_{i2}, \ldots, a_{it}, C_i)$, where the first t components correspond to the features describing the record z_i, i.e. $a_{ij} \in A_j$, for $1 \leq j \leq t$, and the last component presents the class label of z_i, i.e $C_i \in \mathcal{C}$.

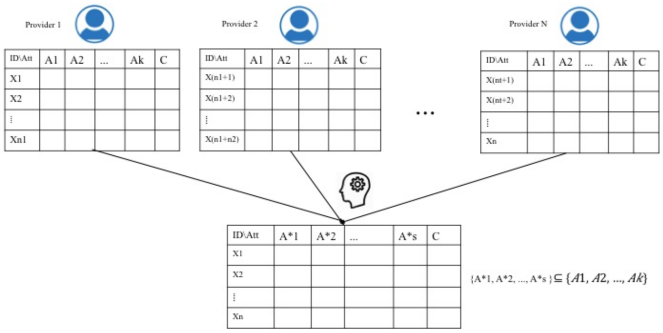
Let $P_1, P_2, \ldots, P_N$ be N distributed parties, such that each party P_s $(1 \leq s \leq N)$ holds dataset D_s, such that $D_i \cap D_j = \emptyset$, for all i, j, and it is desired that a classifier to be trained on $D = \cup_{s=1}^{N} D_s$. However, the data holders accept to share their own dataset only if a minimum amount of privacy is guaranteed.

In this study, we propose privacy-utility feature selection as a tool, which by removing the most irrelevant set of features in terms of accuracy and privacy, produces sanitized datasets, which satisfy the data holders' privacy requirement.

To obtain the optimal subset of features, say $\{A_1^*, A_2^*, \ldots, A_p^*\} \subseteq \mathcal{A}$, first all parties set together the minimum amount of *privacy gain*, say θ, desired to be preserved through removing a subset of features.

To obtain the utility and privacy scores of each feature on all parties' datasets, two metrics named *total feature utility score* and *average privacy score* are computed with the use of distributed secure sum protocol. The details of the proposed private computations will be presented in Section IV.

Finally, after secure computation of feature utility and privacy scores, the trade-off score for each feature is computed to find the optimum subset of features to be removed. In present study, the feature with the minimum privacy-utility trade-off score is the first candidate to be eliminated. If after removing the first feature, the privacy requirement of all data holders satisfies (i.e. *privacy gain*$> \theta$), then that specific feature is removed and the datasets are published. Otherwise, the next feature in terms of minimizing the privacy-utility trade-off score is removed. In the case that by removing half of the features the privacy requirements of parties are not satisfied, it is asked to the data holders to refine the restriction of privacy gain threshold. Figure 1 depicts a higher level representation of the final result of privacy-utility feature selection in horizontal distributed framework. As it can be observed, N parties are participated in the process of data sharing for modeling a classifier. Each data holder owns a table of data, such that all records are expressed with the same set of features $\{A_1, A_2, \ldots, A_t\}$ but different records. With the use of distributed secure sum protocol, an optimal subset of features in terms of privacy-utility trade-off $\{A_1^*, A_2^*, \ldots, A_p^*\} \subseteq \mathcal{A}$ is selected.



ID\Att	A1	A2	...	Ak	C
X1					
X2					
⋮					
Xn1					

ID\Att	A1	A2	...	Ak	C
X(n1+1)					
X(n1+2)					
⋮					
X(n1+n2)					

ID\Att	A1	A2	...	Ak	C
X(nt+1)					
X(nt+2)					
⋮					
Xn					

ID\Att	A*1	A*2	...	A*s	C
X1					
X2					
⋮					
Xn					

Fig. 1: Feature Selection in Horizontal Distributed Data

A. Privacy-Utility Feature Selection

To find the optimal set of features, it is required to compute the total privacy-utility trade-off score for each feature on all datasets. It is worth noting that the best feature in terms of data utility (in distributed architecture) is the one which most effectively discriminates the classes on *whole* dataset $D = \cup_{s=1}^{N} D_s$. On the other hand, differently from utility score, the best feature in terms of privacy gain (in distributed architecture) is the one that maximizes the average privacy gain on all parties' datasets. The reason is that at the end each data holder publishes her own dataset, and it is possible that the maximum privacy gain on whole data does not satisfy the minimum privacy requirement of each party. In both *total utility* and *average privacy* scores computation, distributed secure sum protocol is exploited. In this way, no one knows about the privacy and utility scores of a specific feature in the other party's dataset, but the sum is public. In what follows, we explain in detail the process of secure computing of total utility score and average privacy score for each feature in horizontal distributed system.

Total Utility Score: As discussed in Section IV, the *utility score* in present study is required to be computed based on statistical information of data. Thence, if it is possible to obtain securely the statistical information of whole data, it would be possible to extract the feature utility score on D. To this end, each party P_s, on her own dataset D_s, extracts the number of records respecting k'th value of j'th feature which is labeled C_i, for all k, j, i. The result denoted as $\eta_s(v_k \cdot A_j, C_i)$ is summed up with the use of distributed secure sum protocol, proven to be secure in [9]. Algorithm 1 details the process of secure utility computation.

```
Algorithm 1: Secure computation of total feature utility scores
Data: Each party has the statistical information of her own dataset, ranking feature function f has been set among parties
Result: Total feature utility scores on whole dataset are securely obtained
1  initialization;
2  $SUM(\mathcal{A}, \mathcal{C}) = 0$
3  for $1 \le i \le m$ do
4      for $1 \le j \le t$ do
5          for $1 \le k \le |A_j|$ do
6              for $1 \le s \le N$ do
7                  $P_s$: $\eta_s(v_k \cdot A_j, C_i) \leftarrow$ the number of records in dataset $D_s$ respecting $k$'th value of $j$'th feature labeled $C_i$
8              end
9              $Sum(v_k \cdot A_j, C_i) \leftarrow \mathcal{DSS}_{s=1}^{N}(\eta_s(v_k \cdot A_j, C_i))$
10             $SUM(\mathcal{A}, \mathcal{C}) = SUM(\mathcal{A}, \mathcal{C}) + Sum(v_k \cdot A_j, C_i)$
11         end
12     end
13 end
14 return $f(SUM(\mathcal{A}, \mathcal{C}))$
```

Theorem V.1. *Algorithm 1 reveals nothing to any data holder except the total statistical information of the whole dataset.*

Proof. The only communication among parties occurs at line 9 which consists of a call to the distributed secure sum function ($\mathcal{DSS}$), proven to be secure in [9]. □

Example V.1. Mutual information *of an attribute A_j and class labels $\mathcal{C} = \{C_1, C_2, \dots, C_m\}$, in a labeled dataset, is a common approach to rank features based on their relevance in class discrimination with the use of statistical information of dataset. Formally, the* Shanon entropy *of class labels $\mathcal{C} = \{C_1, C_2, \dots, C_m\}$ is defined as:*

$$H(\mathcal{C}) = -\sum_{C_i \in C} p(C_i) \log_2(p(C_i)) \quad (1)$$

where $p(C_i)$ is the number of records in D labeled C_i divided to the total number of records in D. The conditional entropy *of the output $\mathcal{C}$ to feature A_j is given by:*

$$H(\mathcal{C}|A_j) = -\sum_{v_k \in A_j} \sum_{C_i \in C} p(v_k \cdot A_j, C_i) \log_2(p(C_i|v_k \cdot A_j)) \quad (2)$$

where $p(v_k \cdot A_j, C_i)$ represents the number of elements respecting k'th value of feature A_j and having the class label C_i divided to the whole number of records in D. Then, the mutual information *of A_j and $\mathcal{C}$ is computed through $\upsilon(\mathcal{C}, A_j) = H(\mathcal{C}) - H(\mathcal{C}|A_j)$. In this example, we call $\upsilon(\mathcal{C}, A_j)$, or simply $\upsilon(A_j)$ (when $\mathcal{C}$ is known from the context), the* utility score *of feature A_j. It can be inferred that if each party P_s ($1 \le s \le N$) computes on her own site the number of records in D_s which respects k'th value of attribute A_j ($1 \le j \le t$) and labeled C_i ($1 \le i \le m$), denoted as $\eta_s(v_k \cdot A_j, C_i)$, then by summing up all these numbers* securely, *it is possible to derive $\upsilon(A_j)$ on whole data. More precisely, let the following variables to be derived securely through distributed secure sum protocol among s parties.*

$$\eta(v_k.A_j, C_i) = \sum_{s=1}^{N} \eta_s(v_k.A_j, C_i) \quad (3)$$

$$\eta(C_i|v_k.A_j) = \sum_{s=1}^{N} \eta_s(C_i|v_k.A_j) \quad (4)$$

and $n = \sum_{s=1}^{N} n_s$, where n_s is the number of records in D_s. Then the utility score *of A_j, based on* mutual information, *is obtained as the following:*

$$\upsilon(A_j) = \sum_{v_k \in A_j} \sum_{C_i \in C} \frac{-1}{n} \eta(v_k.A_j, C_i) \log_2(\frac{1}{n} \eta(C_i|v_k.A_j)) \quad (5)$$

Due to the fact that the communications among parties occur only through secure sum protocol *(proven to be secure), no one learns about other party dataset. The feature with the lower score is the one desired to be first removed in terms of feature utility, i.e. it has the minimum relevance in identifying the class labels of the records.*

Average Feature Privacy Score: We recall that the average privacy score is not computed on whole dataset. The rational behind it is that at the end each party will publish her own sanitized dataset, and it is possible the privacy gain on whole data does not satisfy all parties' privacy threshold. This means that the amount of privacy gain needs to be computed on each site separately instead of on the whole dataset. Afterwards, the best feature in terms of privacy gain, is the one which maximizes the average privacy gain among all parties. Hence, each party finds the amount of feature privacy score resulted through removing each feature on her own dataset. More precisely, s'th ($1 \le s \le N$) party calculates the following vector on her own dataset as $\vec{\Omega}_s = (\rho_s(A_1), \rho_s(A_2), \dots, \rho_s(A_t))$, where

$\rho_s(A_j)$ is the feature privacy score of attribute A_j in dataset D_s. Then, the average privacy score for all features is reported through the following relation:

$$\vec{\Omega} = \frac{1}{N}\sum_{s=1}^{N}\vec{\Omega}_s = \frac{1}{N}(\sum_{s=1}^{N}\rho_s(A_1),\ldots,\sum_{s=1}^{N}\rho_s(A_t)) \qquad (6)$$

As it can be inferred from equation 6, the final result of privacy score of each feature is computed through summing up the results of all parties divided to N. To compute *securely* the result of equation 6, simply distributed secure sum protocol is exploited among parties.

Example V.2. *Let D and $D^{'}$ represent the tables of data before and after sanitization (removing a feature or a set of features), respectively. Then, we define the* privacy score *resulting from removing a feature (a set of features) from original dataset D and obtaining the sanitized dataset D', denoted by $\rho(D,D^{'})$, as follows:*

$$\rho(D,D^{'}) = \tfrac{1}{M_D}(-\textstyle\sum_{j=1}^{t}\sum_{k=1}^{|A_j|}(p'(v_k\cdot A_j)\log_2(p'(v_k\cdot A_j)) - p(v_k\cdot A_j)\log_2(p(v_k\cdot A_j)))$$

where $|A_j|$ is the number of possible values for the j'th attribute, $p(v_k\cdot A_j)$ and $p'(v_k\cdot A_j)$ denote the number of records that respects k'th value of j'th attribute divided to the number of records in the original dataset D and the sanitized dataset $D^{'}$, respectively. M_D denotes the maximum amount of entropy *of dataset D, which is achieved when the values of each attribute are equally distributed [1].*

For the sake of simplicity, when the privacy gain score is computed for one specific attribute, say A, we denote the privacy score of feature A as $\rho(A)$ instead of $\rho(D,D')$, where D' is obtained from D by removing just A.

Average Privacy-Utility Score The result of secure computation of *total feature utility* and *average feature privacy* scores for each feature is applied to compute the privacy-utility trade-off score for each attribute A_j, $1 \leq j \leq t$, as $\tau(\upsilon(A_j),\rho(A_j))$, such that it increases when both feature utility and privacy scores increase. The feature $A^* \in \mathcal{A}$ respecting the following relation is the first candidate to be removed:

$$A^* = argmin_{A_j \in \mathcal{A}}\, \tau(\upsilon(A_j),\rho(A_j)) \qquad (7)$$

then, if after removing A^* the required privacy level (θ) of each user is satisfied, the process of feature removal finishes; Otherwise, the above process is repeated for the next attribute respecting relation 7. If after removing half of the features, i.e. $\lfloor \frac{t}{2} \rfloor + 1$, the privacy requirement of all parties is not satisfied, then it is required that they refine the privacy threshold.

Example V.3. *A simple expression matching trade-off score properties, which gives the same weight to feature privacy gain and feature utility, is the following:*

$$\tau(\upsilon(A),\rho(A)) = \frac{1}{2}(\upsilon(A)+\rho(A)) \qquad (8)$$

where $\rho(A)$ and $\upsilon(A)$ are the privacy and utility scores of feature A, and $\tau(\upsilon(A),\rho(A))$ is utilized to denote the privacy-utility trade-off score of A. From equation 8, the feature which gets the minimum score is the best candidate to be removed.

VI. Experimental Analysis

In this section we exploit our proposed methodology on real benchmark dataset. We show that the proposed approach will result in a sanitized dataset that while preserving the privacy requirements of each data holder, it does not have considerable negative impact on the performance of the trained classifier.

The application of the proposed approach will be evaluated on *Tumor* dataset, provided in [12]. This example enlightens the real use case of our proposed approach, since the data providers could be considered as health care centers or hospitals, desired to shape a classifier to prevent a disease outbreak on time, or to find new disease patterns and treatments established on more instances, so more reliable. After removing the outliers, we randomly divide the dataset D of *Primary Tumor* dataset into three subsets, named D_1, D_2 and D_3, containing respectively 110, 135, and 68 elements, such that $D_1 \cup D_2 \cup D_3 = D$.

In this case study, we exploit the same feature utility, feature privacy, and trade-off scores proposed in Examples V.1, V.2, and V.3, respectively.

The minimum amount of privacy gain, on each party's dataset, is set to be $\theta = 0.1$. This means that all parties agree that data to be published only if after removing the selected features, the sanitized dataset has the minimum privacy of 0.1 comparing to the original dataset.

The result of *total utility* and *average privacy* scores of each feature, computed by formulas 5 and 6 respectively, have been reported in Table I.

TABLE I: Total utility score, average privacy score, privacy-utility trade-off score on *Primary Tumor Repository*

Attribute	**Total Utility**	**Average Privacy**	**Privacy-Utility Trade-off**
Age	0.1547	0.6213	0.3880
Sex	0.3354	0.6462	0.4908
Histologic Type	0.5524	0.6884	0.6204
Degree of Diffe	0.3791	0.8795	0.6293
Bone	0.2125	0.8517	0.5321
Bone Marrow	0.0204	0.1450	**0.0827**
Lung	0.1009	0.7623	0.4316
Pleura	0.0679	0.7623	0.4151
Peritoneum	0.2205	0.8557	0.5381
Liver	0.1998	0.9060	0.5529
Brain	0.0671	0.3351	**0.2011**
Skin	0.0603	0.2221	**0.1412**
Neck	0.2915	0.5567	0.4241
Supraclavicular	0.1272	0.6798	0.4035
Axillar	0.2459	0.3085	0.2772
Mediastinum	0.1843	0.8433	0.5138
Abdominal	0.1701	0.9239	0.5470

As it can be inferred from Table I, the feature "*Bone Marrow*" has the minimum trade-off score, and so it is the first candidate to be removed. The result of its removal does not provide the required privacy of any party (Table II, first column). Hence, the next feature having the minimum privacy-utility trade-off score, i.e. "*Skin*", is selected to be removed. Still the minimum privacy threshold is not respected for any party (Table II, second column). Hence, the next feature, i.e.

"*Brain*" is removed. After eliminating this feature, the privacy requirement of all parties, i.e. $\theta \geq 0.1$, is satisfied (Table II, third column). This means that the data of each party, after removing those features, can be published without violating their privacy requirements.

TABLE II: Privacy gain after removing a set of features on each party's dataset.

Feature	$\{BoneM\}$	$\{BoneM, Skin\}$	$\{BoneM, Skin, Brain\}$
Party 1	0.0346	0.0637	**0.1010**
Party 2	0.0434	0.0803	**0.1164**
Party 3	0.0473	0.0835	**0.1355**

Now, it is the time to evaluate how much classification accuracy has been affected by removing a set of features. Table III reports the *False Positive Rate* (FPR) and *True Positive Rate* (TPR) of the four well-known classifiers (*K-Star, C4.5, Naive Bayes, Random Forest*) through 5-fold cross validation on *Primary Tumor* dataset, before and after removing the features {*Bone Marrow, Skin, Brain*}.

As it can be observed from Table III, for exploited well-known classifiers the True Positive Rate has been even slightly improved or remained the same (negative results shows the improvement of accuracy). On the other hand, False Positive Rate for all classifiers has changed negligibly (less than 0.006). From the results of Tables II and III, it can be inferred that by removing a set of *irrelevant* features in terms of privacy-utility trade-off, general privacy gain improves, while at the same time the dataset remain practically useful.

TABLE III: Classification results evaluated on 5-fold cross validation on Primary Tumor dataset.

Algorithm	**K-star**		**C4.5**		**NaiveBayes**		**RandomForest**	
Result	*TPR*	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*
Before	0.413	0.066	0.401	0.082	0.475	0.070	0.431	0.066
After	0.425	0.068	0.407	0.081	0.469	0.070	0.431	0.066
Difference	-0.012	-0.006	-0.005	-0.001	-0.006	0.000	0.000	0.000

It is noticeable that according to standard dimensioning technique, proposed in [13], the minimum size for a dataset to produce reliable result is to dimension it as six times the number of used features. The utilized dataset in our experiment containing 313 elements, already matches this condition ($6 \times 17 < 313$). However, we expect that the proposed approach outcomes the better result on larger datasets with large number of features.

VII. Conclusion

In this paper, we applied feature selection and privacy gain as an ensemble tool to find the best set of features in terms of privacy-utility trade-off in distributed data sharing architecture. The proposed approach, with the use of a distributed secure sum protocol, securely removed the set of irrelevant features to shape a tool for horizontal data sharing, with the aim of data classification. The experimental analysis on a benchmark dataset validates the effectiveness of the proposed technique.

In the future directions, we plan to generalize the proposed approach to a framework respecting different trade-off metrics of different privacy and utility measurements, e.g *differential privacy*. Also, we plan to evaluate the proposed methodology on different varieties of benchmark datasets, from *high dimensional* to *Big Data*. Moreover, we plan to solve the same issue, i.e. finding the optimum subset of features in terms of privacy-utility trade-off, in the case that data is partitioned vertically among different data holders.

Acknowledgment

This work was partially supported by the H2020 EU funded project NeCS [GA #675320] and by the H2020 EU funded project C3ISP [GA #700294].

References

[1] F. Martinelli, A. Saracino, and M. Sheikhalishahi, "Modeling privacy aware information sharing systems: A formal and general approach," in *15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, 2016.

[2] M. Sheikhalishahi, M. Mejri, N. Tawbi, and F. Martinelli, "Privacy-aware data sharing in a tree-based categorical clustering algorithm," in *Foundations and Practice of Security - 9th International Symposium, FPS 2016, Québec City, QC, Canada*, 2016, pp. 161–178.

[3] M. Sheikhalishahi and F. Martinelli, "Privacy preserving clustering over horizontal and vertical partitioned data," in *22nd IEEE Symposium on Computers and Communications, Crete, Greece*, 2017.

[4] E. Bogan and J. English, *Benchmarking for Best Practices: Winning Through Innovative Adaptation*, M. Hill, Ed., 1994.

[5] S. R. M. Oliveira and O. R. Zaane, "Privacy preserving frequent itemset mining," in *Proceedings of the IEEE International Conference on Privacy, Security and Data Mining - Volume 14*, 2002, pp. 43–54.

[6] C. Fung and R. Boutaba, "Design and management of collaborative intrusion detection networks," in *Integrated Network Management, IFIP/IEEE International Symposium on*, 2013, pp. 955–961.

[7] C. Artoisenet, M. Roland, and M. Closon, "Health networks: actors, professional relationships, and controversies," in *Collaborative Patient Centred eHealth*, vol. 141. IOSPress, 2013.

[8] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.

[9] R. Sheikh, B. Kumar, and D. K. Mishra, "A distributed k-secure sum protocol for secure multi-party computations," *CoRR*, vol. abs/1003.4071, 2010.

[10] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[11] E. Bertino, D. Lin, and W. Jiang, "A survey of quantification of privacy preserving data mining algorithms," in *Privacy-Preserving Data Mining*. Springer US, 2008, vol. 34, pp. 183–205.

[12] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition. CVPR*, vol. 1, 2001, pp. I–511–I–518.

[14] M. Banerjee and S. Chakravarty, "Privacy preserving feature selection for distributed data using virtual dimension," in *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM*, 2011, pp. 2281–2284.

[15] Y. Jafer, S. Matwin, and M. Sokolova, "Task oriented privacy preserving data publishing using feature selection," in *Advances in Artificial Intelligence - 27th Canadian Conference on Artificial Intelligence*, 2014, pp. 143–154.

[16] M. S. Y Jafer, S Matwin, "A framework for a privacy-aware feature selection evaluation measure," in *2015 13th Annual Conference on Privacy, Security and Trust (PST)*, July 2015, pp. 62–69.

[17] M. Sheikhalishahi and F. Martinelli, "Privacy-utility feature selection as a tool in private data classification," in *14th International Conference on Distributed Computing and Artificial Intelligence, Porto, Portugal*, 2017.