Privacy-Utility Feature Selection as a tool in Private Data Classification

Mina Sheikhalishahi, Fabio Martinelli

Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Pisa, Italy
{name.surname}@iit.cnr.it

Abstract. This paper presents a novel framework for privacy aware collaborative information sharing for data classification. Two data holders participating in this information sharing system, for global benefits are interested to model a classifier on whole dataset, if a certain amount of privacy is guaranteed. To address this issue, we propose a privacy mechanism approach 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, secure weighted average protocol is utilized to protect information leakage in each site.

1 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 [11], [13]. Such an intelligence could be exploited to improve revenues, e.g. through best practice sharing [4], market basket analysis [12], 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 [1].

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 it preserves the individual privacy requirement while at the same time the data are still practically useful.

In present study, we assume that two data providers are interested to model a classifier on shared data. For example, two hospitals are interested to use the result of classifiers on patients' records to identify the trends and patterns of diseases. The result on whole dataset brings the benefit for both parties to find the better treatment. However, for privacy concerns, the hospitals are unwilling to disclose the patients' records, unless that a privacy level is satisfied [11]. To this end, we propose a privacy-aware feature selection framework which by secure removing the most *irrelevant* features, from both 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 [5]. 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 efficiency are not necessarily equivalent to the best set of feature in terms of privacy gain. To address this issue, we propose an approach based on trade-off metric between *data utility* (classification accuracy) and *privacy gain*, when data is distributed between two parties.

A privacy-aware feature selection framework has been proposed in [9], which reports the efficiency of feature selection approach when privacy and utility are balanced in data publishing. However, in proposed framework only one data holder publishes securely her own dataset. In present study, we assume that two data holders are interested to share their own datasets to model a classifier on whole data. In the case that two parties are involved, it is required that some secure computation protocols be exploited. In present study, we utilize a *secure weighted average* 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 rest of the paper is structured as follows. In Section 2 the related work on two concepts of privacy preserving and feature selection is presented. Section 3 presents *secure weighted average protocol* exploited in this study. Section 4 describes the problem statement and the proposed framework, detailing the secure computation of privacy and utility scores. Finally Section 5 briefly concludes proposing future research directions.

2 Related Work

Several work have been devoted to secure feature selection in multiparty data analysis framework. As a remarkable study, in [2], 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 [8], 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 [9], 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.

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 it is desired to model a classifier on whole data of two parties.

3 Secure Weighted Average Protocol (WAP)

Alice and *Bob* have (a_1, a_2) and (b_1, b_2), respectively. The two parties are interested to jointly compute $\frac{a_1+b_1}{a_2+b_2}$ in a secure way that each will know the final result without knowing the input of the other party.

In this study, we exploit the secure weighted average protocol proposed in [10] as what follows. Let (G, E, D, M) be an encryption scheme, where G is the function for generating public parameters, E and D are the encryption and decryption functions, and M is the message space, with the following properties:

- The encryption scheme (G, E, D) is semantically secure, i.e. an adversary gains no extra information from inspecting ciphertext.
- For all $m, \alpha \in M$, $m_1 \in E(x)$ means that $m_1^{\alpha} \in E(m\alpha)$, i.e. $E(m)$ denotes the set of ciphertexts can be obtained by the encryption of m.
- There exists a computable function f such that for all messages m_1 and m_2 we have $f(E(m_1), E(m_2)) = E(m_1 + m_2)$.

Keeping the above properties of our required probabilistic encryption system, the protocol of secure weighted average is presented as follows.

Assume that *Alice* sets up a probabilistic encryption scheme (G, E, D, M) where the parameters of G are public [10].

- (Step 1): *Alice* encrypts a_1 and a_2 and sends the encrypted values $a_1^{'} \in E(a_1)$ and $a_2^{'} \in E(a_2)$ to *Bob*.
- (Step 2): *Bob* computes a random message $m \in M$ and encrypts $m \cdot b_1$ and $m \cdot b_2$ to obtain $m_1' \in E(m \cdot b_1)$ and $m_2' \in E(m \cdot b_2)$. Then, *Bob* calculates $\mathcal{M}_1 = f(a_1^{'m}, m_1')$, $\mathcal{M}_2 = f(a_2^{'m}, m_2')$, and sends the result to *Alice*, where f is computed through multiplication.
- (Step 3): From the properties of probabilistic scheme (G, E, D), the following are obtained:
$$\mathcal{M}_1 = E(m \cdot a_1 + m \cdot b_1) \qquad , \qquad \mathcal{M}_2 = E(m \cdot a_2 + m \cdot b_2)$$
Hence, *Alice* is able to compute $m \cdot (a_1 + b_1)$ and $m \cdot (a_2 + b_2)$ and consequently $\frac{a_1+b_1}{a_2+b_2}$. *Alice* sends the final result to *Bob*.

Let us to show the result of this protocol as a function which gets as input two pairs of (a_1, a_2) and (b_1, b_2) and returns *the weighted average* of these pairs through the above protocol. We denote this function by *WAP* ($(a_1, a_2), (b_1, b_2)$).

4 Problem Statement and Proposed Framework

Let consider that two data holders are interested to share their own data to model a classifier. 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 categorical 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 $z_i = (v_{i1}, v_{i2}, \ldots, v_{it}, C_i)$, where the first t components correspond to the features describing the record z_i, i.e. $v_{ij} \in A_j$, for $1 \leq j \leq t$, and the last component presents the class label of z_i, i.e $C_i \in \mathcal{C}$. For privacy issues, the data holders accept to share their own dataset only if a minimum amount of privacy is guaranteed.

To obtain the optimal subset of features, say $\{A_1^*, A_2^*, \ldots, A_p^*\} \subseteq \mathcal{A}$, first two parties set together the minimum amount of *privacy gain* denoted by θ, desired to be preserved through removing a subset of features. Then, *privacy* and *utility* scores for each feature, on whole data, are computed with the use of *secure weighted average protocol*. Finally, after secure computation of feature utility and privacy scores, the trade-off score for each feature is computed to obtain the optimum subset of features. The feature with the minimum privacy-utility trade-off score is the first candidate to be removed. If after removing the first feature, the privacy requirements of both 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, the data holders are required to refine the restriction of privacy gain threshold.

4.1 Secure Utility Score Computation

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 [5]. 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 [7]. *Feature ranking* as a well-known feature selection approach, creates a scoring function, say $\upsilon(A)$, computed from the impact of feature A in discriminating class labels. Generally, a high score is indicative of a valuable feature. In present study, we apply a well-known feature ranking technique based on *Mutual Information* [5] to score the features based on their *utility*.

Formally, let $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ be the class labels of a dataset D, and $A = \{v_1, v_2, \ldots, v_{|A|}\}$ be the set of values of the feature A. Then, the *Shannon entropy* of variable $\mathcal{C}$ is defined as follows:

$$H(\mathcal{C}) = -\sum_{C_i \in \mathcal{C}} p(C_i) \log(p(C_i))$$

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

$$H(\mathcal{C}|A) = -\sum_{v_k \in A} \sum_{C_i \in \mathcal{C}} p(v_k \cdot A, C_i) \log(p(C_i|v_k \cdot A))$$

where $p(v_k \cdot A, C_i)$ represents the number of elements respecting k'th value of A and having the class label C_i divided by the whole number of records in D. The decrease in uncertainty of the output $\mathcal{C}$ observing feature A, denoted by function $\upsilon(A)$, is computed as $\upsilon(A) = H(\mathcal{C}) - H(\mathcal{C}|A)$, where $H(\mathcal{C})$ and $H(\mathcal{C}|A)$ represent the *Shannon entropy* and *conditional entropy*, respectively. We call $\upsilon(A)$ the *utility score* of feature A. The feature with the lower score is the one desired to be first removed, since it has the minimum relevance in identifying the class labels of the records. Algorithm 1 details the process of secure *utility score* computation between two parties.

Algorithm 1: *utility.score()* : Secure Utility Score Computation

```
1  initialization;
2  for $1 \le j \le t$ do
3      for $1 \le k \le |A_j|$ do
4          for $1 \le i \le m$ do
5              Alice: $a_1 \leftarrow$ number of records in $D_a$ respecting $v_k \cdot A_j$ and $C_i$
6              Alice: $a_2 \leftarrow$ total number of elements in $D_a$
7              Alice: $a'_1 \leftarrow$ number of records in $D_a$ with class label $C_i$ that respect
               $v_k \cdot A_j$
8              Alice: $a'_2 \leftarrow$ number of records in $D_a$ with class label $C_i$
9              Bob: $b_1 \leftarrow$ number of records in $D_b$ respecting $v_k \cdot A_j$ and $C_i$
10             Bob: $b_2 \leftarrow$ total number of elements in $D_b$
11             Bob: $b'_1 \leftarrow$ number of records in $D_b$ with class label $C_i$ that respect
               $v_k \cdot A_j$
12             Bob: $b'_2 \leftarrow$ number of records in $D_b$ with class label $C_i$
13             $p(v_k \cdot A_j, C_i) \leftarrow WAP((a_1, a_2), (b_1, b_2))$
14             $p(C_i|v_k \cdot A_j) \leftarrow WAP((a'_1, a'_2), (b'_1, b'_2))$
15             $\upsilon(A_j) \leftarrow \upsilon(A_j) - p(v_k \cdot A_j, C_i) \log(p(C_i|v_k \cdot A_j))$
16         end
17         return $\upsilon(A_j)$
18     end
19 end
```

Theorem 1. *Algorithm 1 reveals nothing to the other party, except* feature utility *scores on whole data.*

Proof. The only communication between two parties occur at lines 13 and 14, which is a call to secure weighted average computation protocol, proven to be secure in [10]. □

4.2 Secure Privacy Score Computation

Data privacy is quantified as the degree of uncertainty that the original data can be inferred from the sanitized one [3]. Generally, reducing the information which a dataset carries will increase privacy gain. Hence, removing a feature from a dataset can be considered as a tool which increases this uncertainty.

We compute the *privacy score* [11] resulting from removing feature A_j from original dataset D and obtaining the sanitized dataset $D - \{A_j\}$, denoted by $\rho(D, D - \{A_j\})$, as follows:

$$\begin{aligned}\rho(D, D - \{A_j\}) = &-(\textstyle\sum_{s=1}^{t} \sum_{k=1}^{|A_s|} (p(v_k \cdot A_s) \cdot \log(p(v_k \cdot A_s)) \\ &- \textstyle\sum_{s=1, s \neq j}^{t} \sum_{k=1}^{|A_s|} (p(v_k \cdot A_s) \cdot \log(p(v_k \cdot A_s)))\end{aligned}$$

where $|A_s|$ is the number of values for the s'th attribute, $p(v_k \cdot A_s)$ denotes the number of records that respects k'th value of s'th attribute divided by the number of records. For the sake of simplicity, when the privacy gain score is computed for one specific attribute, say A_j, we denote the privacy score of feature A_j as $\rho(A_j)$ instead of $\rho(D, D - \{A_j\})$. Algorithm 2 details the process of secure privacy score computation between two parties.

Algorithm 2: *privacy.score()*: Secure Privacy Score Computation

```
1  initialization;
2  for $1 \leq j \leq t$ do
3      for $1 \leq j' \leq t, j' \neq j$ do
4          for $1 \leq k \leq |A_j|$ do
5              for $1 \leq k' \leq |A_{j'}|$ do
6                  Alice: $a_1 \leftarrow$ number of records in $D_a$ respecting $v_k \cdot A_j$
7                  Alice: $a_2 \leftarrow$ total number of elements in $D_a$
8                  Alice: $a' \leftarrow$ number of records in $D_a - \{A_j\}$ which respect $v_{k'} \cdot A_{j'}$
9                  Bob: $b_1 \leftarrow$ number of records in $D_b$ respecting $v_k \cdot A_j$
10                 Bob: $b_2 \leftarrow$ total number of elements in $D_b$
11                 Bob: $b' \leftarrow$ number of records in $D_b - \{A_j\}$ which respect $v_{k'} \cdot A_{j'}$
12                 $p(v_k \cdot A_j) \leftarrow WAP((a_1, a_2), (b_1, b_2))$
13                 $p'(v_{k'} \cdot A_{j'}) \leftarrow WAP((a'_1, a_2), (b'_1, b_2))$
14                 $\rho(A_j) \leftarrow$
                   $-(p(v_k \cdot A_j) \log(p(v_k \cdot A_j)) - p'(v_{k'} \cdot A_{j'}) \log(p'(v_{k'} \cdot A_{j'})))$
15             end
16         end
17     end
18     return $\rho(A_j)$
19 end
```

Theorem 2. *Algorithm 2 reveals nothing to other party except the privacy scores of features.*

Proof. The only communication between *Alice* and *Bob* occur at lines 12 and 13, which is a call to secure weighted average protocol, proven to be secure in [10]. □

4.3 Privacy-Utility Feature Selection

A simple expression matching trade-off score properties, which gives the same weight to feature privacy gain and feature utility,could be defined as $\tau(\upsilon(A), \rho(A)) = \frac{1}{2}(\upsilon(A) + \rho(A))$, where $\rho(A)$ and $\upsilon(A)$ are the privacy and utility scores of feature A. From the proposed metric, the feature which gets the minimum score is the best candidate to be removed. Algorithm 3 details the process of secure privacy-utility score computation.

Algorithm 3: *privacy.utility()*: Secure Privacy-Utility Score Computation

```
   Data: Alice and Bob have statistical information of $\{A_1, A_2, \ldots, A_k\}$
   Result: The feature respecting minimum privacy-utility score
 1 initialization;
 2 $A^* = A_1$
 3 for $1 \leq j \leq t$ do
 4     $\upsilon(A_j) \leftarrow utility.score(A_j)$
 5     $\rho(A_j) \leftarrow privacy.score(A_j)$
 6     $\tau(A_j) = \frac{1}{2}(\upsilon(A_j) + \rho(A_j))$
 7     if $\tau(A^*) > \tau(A_j)$ then
 8         $A^* \leftarrow A_j$
 9     end
10 end
11 return $A^*$
```

Algorithm 3 is secure, since the only communications between *Alice* and *Bob* occur at lines 4 and 5, and they have proven to be secure in Theorems 1 and 2.

After executing Algorithm 3, both *Alice* and *Bob* compute the *privacy gain* on their original dataset, and sanitized dataset resulted from removing *irrelevant* features, say $\{A_{j1}, A_{j2}, \ldots, A_{jl}\}$, as the following:

$$\rho(D, D - \{A_{j1}, A_{j2}, \ldots, A_{jl}\}) = -(\textstyle\sum_{s=1}^{t} \sum_{k=1}^{|A_s|} (p(v_k \cdot A_s) \cdot \log(p(v_k \cdot A_s)) \\ - \textstyle\sum_{s=1, s \notin \{j1, \ldots, jl\}}^{t} \sum_{k=1}^{|A_s|} (p(v_k \cdot A_s) \cdot \log(p(v_k \cdot A_s)))$$

If in both sides the *privacy gain* $\rho(D, D - \{A_{j1}, A_{j2}, \ldots, A_{jl}\})$ is higher than θ, *Alice* and *Bob* publish the sanitized datasets; Otherwise, the next feature satisfying the minimum privacy-utility score is found through Algorithm 3. The process continues till both parties reach the required privacy threshold.

5 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 secure weighted average protocol, securely removes the set of irrelevant features to shape a tool for modeling a classifier on shared data.

In the future directions, we plan to generalize the proposed approach to a framework respecting different trade-off metrics of different privacy and utility metrics. Moreover, we plan to exploit the proposed approach on real benchmark datasets to evaluate a real case-study.

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. Artoisenet, C., Roland, M., Closon, M.: Health networks: actors, professional relationships, and controversies. In: Collaborative Patient Centred eHealth. vol. 141. IOSPress (2013)
2. Banerjee, M., Chakravarty, S.: Privacy preserving feature selection for distributed data using virtual dimension. In: Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM. pp. 2281–2284 (2011)
3. Bertino, E., Lin, D., Jiang, W.: A survey of quantification of privacy preserving data mining algorithms. In: Privacy-Preserving Data Mining, vol. 34, pp. 183–205. Springer US (2008)
4. Bogan, C.E., English, M.J.: Benchmarking for best practices : winning through innovative adaptation. New York : McGraw-Hill (1994)
5. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. 40(1), 16–28 (Jan 2014)
6. Faiella, M.F., Marra, A.L., Martinelli, F., Francesco, Saracino, A., Sheikhalishahi, M.: A distributed framework for collaborative and dynamic analysis of android malware. In: 25th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP), St. Petersburg, Russia (2017)
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning 3, 1157–1182 (2003)
8. Jafer, Y., Matwin, S., Sokolova, M.: Task oriented privacy preserving data publishing using feature selection. In: Advances in Artificial Intelligence - 27th Canadian Conference on Artificial Intelligence. pp. 143–154 (2014)
9. Jafer, Y., Matwin, S., Sokolova, M.: A framework for a privacy-aware feature selection evaluation measure. In: 13th Annual Conference on Privacy, Security and Trust, PST 2015, Izmir, Turkey, July 21-23, 2015. pp. 62–69 (2015)
10. Jha, S., Kruger, L., McDaniel, P.: Privacy Preserving Clustering, pp. 397–417. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
11. Martinelli, F., Saracino, A., Sheikhalishahi, M.: 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)
12. Oliveira, S.R.M., Zaïane, O.R.: Privacy preserving frequent itemset mining. In: Proceedings of the IEEE International Conference on Privacy, Security and Data Mining - Volume 14. pp. 43–54. CRPIT ’14 (2002)
13. Sheikhalishahi, M., Mejri, M., Tawbi, N., Martinelli, F.: 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. pp. 161–178 (2016)