PRIVACY PRESERVING FEATURE SELECTION AND MULTICLASS CLASSIFICATION FOR HORIZONTALLY DISTRIBUTED DATA

Yunmei Lu
Department of Computer Science
33 Gilmer Street SE Atlanta, GA, USA

Mingyuan Yan
University of North Georgia
Dahlonega, GA, USA

Meng Han* and Qingliang Yang
Data-driven Intelligence Research Laboratory
College of Computing and Software Engineering, Kennesaw State University
1100 South Marietta Pkwy, Marietta, GA, USA

Yanqing Zhang
Department of Computer Science
33 Gilmer Street SE Atlanta, GA, USA

(Communicated by Zhipeng Cai)

Abstract. In the last two decades, a lot of scientific fields have experienced a huge growth in data volume and data complexity, which brings data miners lots of opportunities, as well as many challenges. With the advent of the era of big data, applying data mining techniques on assembling data from multiple parties (or sources) has become a leading trend. However, those data mining tasks may divulge individuals' privacy, which leads to the increased concerns in privacy preserving. In this work, a Privacy Preserving feature selection method (PPFS-IFW) and Multiclass Classification method (PPM2C) are proposed. Experiments had been conducted to validate the performance of the proposed approaches. Both PPFS-IFW and PPM2C were tested on six benchmark datasets. The testing results demonstrate PPFS-IFW’s capability in enhancing the classification performance at the level of accuracy by selection informative features. PPFS-IFW can not only preserve private information but also outperform some other state-of-the-art feature selection approaches. Experimental results also show that the proposed PPM2C method is workable and stable. Particularly, it reduces the risk of over-fitting when compared with the regular Support Vector Machine. In the meantime, by employing the Secure Sum Protocol to encrypt data at the bottom layer, users’ privacy is preserved.

2010 Mathematics Subject Classification. Primary: 58F15, 58F17; Secondary: 53C35.
Key words and phrases. Classification, selection, privacy, feature selection.

* Corresponding author: Meng Han.
1. **Introduction.** In the last two decades, a lot of scientific fields have experienced a huge growth in data volume and data complexity, which brings data miners lots of opportunities, as well as many challenges. With the advent of the era of big data, machine learning and data mining approaches have been widely used in a large number of applications to analyze and assemble data. Those approaches have become very important tools in discovering useful information in many domains, such as discovering or analyzing medical data, consumer purchase data and census data. Since applying data mining approaches on aggregated datasets may enable us to generate more reliable prediction models and obtain more useful patterns, applying this technique on assembling data from databases maintained by different sources become pretty popular \[26\]\[6\]\[37\]. Applications focus on medical research, customer service or homeland security could all benefit from the development and usage of data mining techniques \[32, 31\]. For example, the Center for Disease Control (CDC) may want to identify the trend of a disease to better understand its progression via data mining techniques but facing the dilemma that they do not have relevant data. Meanwhile, insurance companies have collected considerable data regarding this issue and could provide assistance.

Data distribution can be roughly divided into two categories: horizontal distribution and vertical distribution. We say data stored at different locations is horizontally distributed if those distributed stored data records share common attributes \[8, 45\]. For example, customer information collected at different bank branches. Data can also be vertically distributed. In this case, data collected at different sites have same records but the attributes collected for each record may be different. Such as bank, insurance company or auto insurance company who collects different information regarding the same customer.

Data mining do play an effective role in providing hidden or more useful information. However, mining on aggregated data might divulge the sensitive information about individuals. It thus leads to increasing concerns about privacy protection during the data mining process, which has a strong intention to prevent different parties from sharing information (especially sensitive information). To deal with this challenge, many researchers focus on the study of Privacy Preserving Data Mining (PPDM) technique. For example, a lot of work has been done in \[47, 17, 50, 2, 41, 44, 30, 11, 52, 21, 22\], which aim to provide a means to address the privacy preserving issue without accessing the actual data values, thus avoid the disclosure of information beyond the final results.

Among the machine learning and data mining approaches, feature selection and classification are two of the most important topics. Feature selection techniques address the issue of dimensionality reduction by selecting an available subset of features via predetermined selecting criteria. Therefore, it plays a vital role in optimizing mining procedure via selecting smaller size features and thus improves the performance of data mining algorithms. Many feature selection approaches have been proposed \[24, 43, 40, 49, 42\], but few of them takes privacy preservation as a concern. Classification targets at providing a method to identify or classify data that belong to unknown groups into different categories through building effective classifiers. During the classification process, data samples known as training set is commonly used to generate and validate the classifiers. Multi-class classification, representing a branch of classification problem, has being a hot topic and research direction in many domains for the past few years and become more and more important in the era of big data. During the classification process, irrelevant and
redundant features are usually removed after feature selection procedure. Thus, a
good feature selection strategy can reduce the computational complexity needed
for classification process. Researchers have proposed a large number of state-the-art
multi-class classification approaches and algorithms based on traditional but
popular classification algorithms, such as Support Vector Machine (SVM) [13][10],
Decision Tree (DT) [39], Naive Bayes (NB) [38] and K-Nearest Neighbor (KNN)[5].

In this work, we are interested in the investigation of Privacy Preserving Feature
Selection and Multi-Class Classification problems for horizontally distributed data.
Based on one of our previous published paper [33], we propose a method of Privacy
Preserving Feature Selection via Integrating Filter and Wrapper approaches for
Horizontally Distributed Data, named PPFS-IFW and a privacy preserving multi-
class classification algorithm named PPM2C. Experiments had been conducted
to validate the performance of the proposed approaches. Both PPFS-IFW and
PPM2C were tested on six benchmark datasets. The testing results demonstrate
PPFS-IFW’s capability in enhancing the classification performance at the level of
accuracy by selection informative features. PPFS-IFW can not only preserve pri-
vate information but also outperform some other state-of-the-art feature selection
approaches. Experimental results also show that the proposed PPM2C method is
workable and stable. Particularly, It reduces the risk of over-fitting when compared
with the regular Support Vector Machine. In the meantime, by employing the Se-
cure Sum Protocol to encrypt data at the bottom layer, users’ privacy is preserved.

In the following sections, the details of PPFS-IFW will be introduced in section
2. PPM2C will be introduced in section 3, followed by the experimental results and
discussion presented in section 4. A short conclusion is included in section 5.

2. Privacy preserving feature selection via integrating filter.

2.1. Background. Feature selection methods can be grouped into two categories
according to their searching directions: forward selection and backward selection
[35]. Forwarding selection usually starts searching relevant features from an empty
subset and adds one or more at each step until a stop criterion is met. On con-
trast, the backward selection methods usually start searching from the entire feature
space and eliminate one or more at each step, until a predetermined stop criteria is
reached.

According to different selecting strategies and procedures, feature selection meth-
ods can also be formulated into three main categories: filter, wrapper and embedded
approaches [24]. The filter methods usually take account of the statistical properties
of features and rank them according to predefined criteria on relevant information.
This step is always done prior to classification and is completely independent of
data mining algorithms. The selected feature subset will be applied to the clas-
sification or cluster algorithms. Since the selection procedure is independent of
the mining algorithms, the effects of the subset of features on the performance of
the mining algorithms will also be avoided. Since filter methods are always sepa-
rare from mining algorithms, they are fast. Just as the name implies, the wrapper
methods often wrapped the feature selection step into the process of data mining
algorithms. Compared with the filter methods, wrapper methods have the advan-
tages that it takes the performance of data mining algorithms into account. Thus
a better classification model will be constructed. However, it needs to repeatedly
train, test the data, and build classification model at each step when a subset of
features are selected. The computational complexity thus will be increased sharply.
The third kind of feature selection approaches is embedded methods, which perform feature selection in the process of constructing data mining model by adding or modifying the optimizing process of classification as discussed in [36],[9]. Many feature selection approaches have been proposed [24],[43],[40],[49],[42],[28],[29] as data are integrated into a central location, while the privacy concerns of sharing data among distributed parties bring many challenges to feature selection.

To solve the problem of privacy concern during feature selection, we propose a Privacy Preserving Feature Selection method via Integrating Filter and Wrapper approaches (named PPFS-IFW in brief) for Horizontally Distributed Data. PPFS-IFW is based on our previously published privacy preserving framework, which is appointed by Privacy Aware Non-linear SVM for Multi-source Big Data (PAN-SVM)[33].

2.2. Methods. In this section, we discuss in detail that the methods used in feature selection.

2.2.1. Filter Measurements. Filter methods often employ several independent measurements to evaluate features and filter relevant, informative features before the classification process. It is usually much faster than wrapper methods. Especially for high dimensional datasets, it is necessary to remove unnecessary features first to speed up the selection procedure. In the current work, three popular measurements are used as the filtering criteria, which are Fisher Score [16], Welch-t-test [46] and between versus within class scatter ratio [15].

- Fisher Score.

\[
F(j) = \left| \frac{\mu_j^+ - \mu_j^-}{(\sigma_j^+)^2 + (\sigma_j^-)^2} \right|
\]  

- Welch-t-test.

\[
|t_j| = \frac{\mu_j^+ - \mu_j^-}{\sqrt{\frac{(\sigma_j^+)^2}{n^+} + \frac{(\sigma_j^-)^2}{n^-}}}
\]  

- Between versus Within Class Scatter Ratio.

\[
S_w = \sum_{i=1}^{c} \sum_{j=1}^{n_c} (x_j - \mu_i) * (x_j - \mu_i)^T
\]

\[
S_b = \sum_{i=1}^{c} (\mu_i - \mu) * (\mu_i - \mu)^T
\]

\[
S = \frac{S_b}{S_w}
\]

For fisher score and welch-t-test, \( j \) represents the \( j^{th} \) feature, \( \mu_j^+ \) and \( \mu_j^- \), \( \sigma_j^+ \) and \( \sigma_j^- \) denote the means and standard deviations of the \( j^{th} \) feature in the corresponding positive class and negative class. \( n^+ \) and \( n^- \) denote the sample numbers in class + and class -, respectively.

For scatter ratio, \( S_w \) and \( S_b \) denote the within class and between class scatter matrix, respectively. \( S \) is the within versus between class scatter ratio. It is a vector with \( n \) (number of features) elements. \( c \) equals to the number of class. Particularly, \( c \) equals to 2 for binary classification. \( x_j \) represents the \( j_{th} \) record in the \( j^{th} \) class, \( n_c \) denotes the number of samples in the \( i_{th} \) class, \( \mu_i \) denotes the mean of \( i_{th} \) class, and \( \mu \) denotes overall mean in these \( c \) classes.
2.2.2. **PAN-SVM Classifier.** When applied to a classification problem, wrapper methods used for selecting features are closely related to the classifiers. In the current work, classification accuracy is used as the wrapper method, by which features will be evaluated at each iteration, and PAN-SVM [33] is used as the binary classifier. PAN-SVM contains three layers, where each layer performs corresponding functions. The bottom layer protects individual’s data privacy, where sampled data from multiple parties are encrypted via the Secure Sum Protocol [12] and sent to the miner. At this layer, data are sampled by $k$-means clustering method and the center data are used as landmarks [51][14]. At the medium layer, the landmark points are used to approximate kernel matrix via Nystrom technique [51][14] to reduce the complex computation burden and then the matrix will be decomposed via eigenvalue decomposition method. After the procedure of kernel matrix approximation and decomposition, the non-linear separable SVM will be converted into a linear separable SVM. The linear SVM will be optimized and speeded up by linear search and cutting plane techniques [18] at the top layer. Although the classification accuracy sacrifices slightly when compared with regular SVM, like LIBSVM [54], the individual private information is preserved. Furthermore, the training process is speeded up when compared with other distributed classification methods. Details about PAN-SVM could be found in [33].

2.2.3. **Workflow.** The proposed privacy preserving feature selection algorithm PPFS-IFW integrates the filter and wrapper methods. The work flow can be summarized as follows:

**Step 1.** Calculate the three measurements as stated in equation (1), (2) and (3).

**Step 2.** Choose features that selected by all of the three measurements and remove unselected ones. Features will be selected if they meet the predetermined thresholds based on the three measurements. The chosen feature set is referred as voted feature set in the following steps.

**Step 3.** Ranking features in the voted set (Wrapper method). The Ranking process can be summarized as five sub-steps. First, computing the overall classification accuracy with the selected features, denoted as $overall_{acc}$. Second, computing the classification accuracy for each classifier (PAN-SVM) that without feature $i (i = 1...k)$, denoted as $ith_{acc}$. For data with $k$ features, there will be $k$ classifiers. Third, ranking features based on their accuracy. Fourth, Removing a feature in this way. Suppose $ith_{acc}$ is the highest accuracy among the $k$ accuracies calculated in step 2), if $ith_{acc} > overall_{acc}$, then remove feature $i$, since it increases the classification accuracy by eliminating it. Otherwise, if there is no feature increases accuracy, a local maxima accuracy of $ith_{acc}$ is obtained, then set $overall_{acc} = ith_{acc}$ and remove feature $i$, since it gives the classifier the highest negative affection. Last, Repeat the ranking process.

**Step 4.** Return a ranked feature list.

In step 2, the user can define a threshold indicating the approximate number of features needed to be kept or eliminated. For example, for the Welch-t-test measurement, features that have $p$-value larger 0.01 will be deleted. Therefore, a large number of features can be eliminated at this step. The highest negative affection used in step 3 can be illustrated through the following example. For example, if removing feature $i$, the classifier could obtain 98% classification accuracy, while removing feature $j$, the accuracy is 90%, we say feature $i$ has a higher negative
affection to classifier than feature \( j \), so feature \( i \) will be eliminated at this step. 5-fold cross validation is used at each iteration for every classifier. PPFS-IFW returns a ranked list of features that are selected based on the three filter methods.

3. Privacy preserving multiclass classification.

3.1. Methods. In this section, the details of PPM2C method is presented.

3.1.1. Multiclass Support Vector Machines. Currently, the methods for solving multi-classification problem can be formulated into two categories. The first category aims at directly solving the multi-classification problem by extending existed classifiers and taking all collected data under consideration in one optimization problem. The other category, on the contrary, focusing on indirectly solving the problem by converting it into several binary classification problems via constructing and combining multiple binary classifiers, which are usually SVM classifiers. SVM is a well-known sophisticated classification method, which was originally designed for binary classification problem. Now it has been widely used in many other domains, including multi-class classification problem. Furthermore, solving multiclass classification problem indirectly are commonly be formulated into two cases: One-Versus-All (OVA) method \([7]\) and One-Versus-One (OVO) method \([27]\), which are briefly discussed in the following two subsections.

3.1.2. One Versus All. For a \( k \)-class classification problem, the OVA method constructs \( k \) SVM classifiers. The \( h \)th SVM is trained by taking all of the samples in the \( h \)th class with a positive label \((+1)\), and all samples in the remaining classes with a negative label \((-1)\). Thus, given a multiclass classification problem, the \( h \)th SVM solves the problem described in equation (4):

\[
\min_{w^h, b^h, \xi^h} \frac{1}{2}(w^h)^T w^h + C \sum_{i=1}^{m} \xi_i \\
(w^h)^T \Phi(x_i) + b^h \geq 1 - \xi_i, \text{ if } y_i = h, \\
(w^h)^T \Phi(x_i) + b^h \leq -1 + \xi_i, \text{ if } y_i \neq h \\
\xi_i \geq 0, i = 1, ..., m
\]

Where \( x_i \in \mathbb{R}^n \) \((i = 1, 2, ..., m)\) is a sample with \( n \) number of attributes, \( y_i \in \{1, 2, ..., k\} \) is the class label of \( x_i \), and \( \Phi \) is a kernel matrix. The Radial Basis Function (RBF) kernel function \( \Phi = \exp(-\|x - x'\|^2) \) is used here. \( C \) is a penalty parameter. To train the \( h \)th SVM is equivalent to find the maximal separate hyperplane by maximizing the term \( 2/\|w^h\| \). Once solving equation (4), there are \( k \) decision functions will be obtained and can be used to predict unknown samples. There will be \( k \) output values for \( k \) classifiers, a strategy of “Winner takes all” will be used to decide the label of sample \( x \). If the predicted value for \( x \) is positive, \( x \) will be classified into the class with a positive label in the current classifier; otherwise, \( x \) will be put in the class which has the largest absolute approximate value based on the decision function. For example, if the predicted value for \( x \) is \(-2.9\), then \( x \) belongs to class 3; if it is \(-2.1\), then \( x \) belongs to class 2.

The advantage of OVA scheme is that only \( k \) binary SVM classifiers have to be trained for a \( k \)-class classification problem, which speeds up the entire training process. However, the one versus all method might result in making the training data dramatically unbalanced.
3.1.3. One Versus One. On the other hand, for a $k$-class classification problem, OVO method will constructs binary classifiers to separate each one from the other, such as class 1 vs class 2, class 1 vs class 3, class 2 vs class 3, class $k-1$ vs class $k$ and so on. There are different approaches that can be used to test unknown data after all the SVM classifiers are built. [19] proposed a “Max Wins” strategy based on a sign function. The decision function for class $i$ and class $j$ works in the way that if it determines $x$ in the $i^{th}$ class, then the vote for the $i^{th}$ class will be increased by one; Otherwise, the vote for $j^{th}$ class will be increased by one. In the end, the predicted class for $x$ will be the one with the largest voting value. Although, OVO approach might have to train more binary classifiers than that of the OVA strategy, it is usually much faster. That is because generating several more SVM classifiers with smaller size is much faster than building fewer classifiers with larger size when solving the quadratic programming optimization problems exist in SVM. However, a potential limitation of the “Max wins” is that it might not be a good strategy in the case that two classes have identical votes. Since efficiency is an important concern, in this work, the OVA schema is used for PPM2C.

3.1.4. Workflow. The workflow of PPM2C can be briefly presented in Fig. 1. PPM2C is also based on [33]. PAN-SVM proposed a privacy preserving framework for binary classification based on horizontally located data. The data is encrypted via Secure Sum Protocol proposed in [12], a Secure Multiparty Computation (SMC) technique presented in [12] is used. Furthermore, the computation of the kernel matrix in SVM is reduced significantly via Nystrom kernel matrix approximation method discussed in [14] [48] and eigenvalue decomposition approach. In PPM2C, PAN-SVM is used to construct multiple binary classifiers for the purpose of preserving privacy during the data mining process.

4. Experimental results and discussion. The performance will be assessed by the measurement of classification accuracy, which is formulated in equation (5).

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]
Table 1. Details of Datasets used in Evaluation of PPFS-IFW

| Dataset                  | num. samples | num. features | C     | γ      |
|--------------------------|--------------|---------------|-------|--------|
| Diabetes (DIA)           | 768          | 8             | 512.0 | 0.0078125 |
| Ionosphere               | 351          | 34            | 8.0   | 0.5    |
| Colon                    | 62           | 2000          | 32.0  | 0.0078125 |
| Leukemia                 | 72           | 7129          | 128.0 | 0.0001221 |
| Lymphoma (DLBCL)         | 47           | 4026          | 2.0   | 0.0078125 |
| Breast Cancer (WBC)      | 569          | 30            | 128.0 | 8.0    |

Where TP represents the number of True Positive, TN states the number of True Negative, FP denotes the number of False Positive and FN indicates the number of False Negative.

4.1. Evaluation of PPFS-IFW. The Cross Validation (CV) method is often used to assess the performance of classifier according to the classification due to lack of data that could be used as separate testing samples. During the cross validation process, data will be randomly split into $k$ (k-fold) subsets. At each training round, $k - 1$ subsets are used as the training data, and the remaining one subset is used as the testing set. However, as pointed by [49], the feature selection results may vary due to even a single difference in the training set, especially for small datasets. Many feature selection methods are done with all samples, and the cross validation step was only done during the classification process, which makes the feature selection external to the cross validation procedures, and leads to information leakage in the feature selection step. It calls this kind of error made by cross validation as CV1 error. [49] demonstrates the existing bias via simulation data and suggested another error evaluation method, named CV2. Under CV2 scenario, a separate dataset will be used as test samples and excluded from the training set before the feature selection process.

4.1.1. Datasets. PPFS-IFW was assessed and tested on real datasets including three microarray datasets with different number of features. The Diabetes and Ionosphere data are downloaded from LIBSVM repository [54], the Wisconsin Breast Cancer data (WBC) are downloaded from University of California, Irvine (UCI) Machine Learning Repository [3], and the microarray datasets of Leukemia, Lymphoma and colon [23] are download from [53]. The details about the datasets are presented in Table 1. $C$ is the penalty parameter for SVM and $γ$ is a free parameter for Radial Basis Function kernel (RBF) used in SVM. They are generated by 10-fold cross validation.

4.1.2. Effectiveness of PPFS-IFW. To check whether the proposed algorithm is workable or not and its effectiveness, PPFS-IFW is tested on six datasets under CV1 and CV2 scenario, respectively. The classification accuracies achieved by PAN-SVM before and after the feature selection of PPF-IFW are shown in Figure 2 and Figure 3, respectively. In both figures, the blue bars denote the overall classification accuracy with all features (without feature selection step), and the red bars represent classification accuracy after executing the feature selection step of PPFS-IFW. From Figure 2 and Figure 3, we can observe that the classification accuracies of PAN-SVM are significantly improved by PPFS-IFW for most of the data, especially for the three microarray dataset of Colon, DLBCL and Leukemia, no matter under
CV1 or CV2 testing schema. These results illustrate that the proposed algorithm PPFS-IFW is not only workable but also effective.

4.1.3. Performance Comparison under CV1 and CV2. Further comparison between the classification accuracies made by PAN-SVM is conducted to check whether there is any difference under CV1 and CV2 testing schema. The detail results are shown as in Figure 4 and Table 2. It can be seen from Figure 4 that before the proposed
After feature selection, the classification accuracies are improved significantly for all of the datasets, no matter under which testing situation, but there is no obvious pattern found. The results described in Table 2 show a wider picture and illustrate that the accuracy could be improved higher under CV1 test condition. It makes sense because there exists information leakage during the feature selection step under CV1 test schema. The fourth and fifth column in Table 2 show the selected number of features under CV1 and CV2. From these two columns we can observe that although accuracy can be improved slightly higher under CV1, the number of selected features under CV2 is usually smaller than that under CV1.

The experimental results show that PPFS-IFW can significantly improve the classification performance by selecting informative features under both CV1 and CV2 testing situation. The improvement of classification accuracy has no obvious pattern for different datasets when tested under CV1 and CV2. Overall, higher
4.1.4. Comparison with Other Methods. We also compare the proposed method in the current work with the other state-of-the-art methods, such as Fisher-SVM, FSV, RFE-SVM and KP-SVM [34][25] on the three common datasets of DIA, WBC and Colon, and the result data of these methods are selected from [34]. From Table 3, we can observe that the proposed method of PPFS-IFW (referred as ours in the table) outperforms the other methods for datasets DIA and WBC, which own few features and defeated by Fisher-SVM, RFE-SVM and KP-SVM on Colon data with high dimensionality. The proposed feature selection algorithm PPFS-IFW is based on PAN-SVM, whose kernel matrix is approximated and decomposed to alleviate the complex computation and communication cost. Therefore, it is reasonable that the classification accuracy sacrifices for some data.

PPFS-IFW outperforms the other four methods on two datasets that have few features but defeated on the microarray dataset of Colon. Slight sacrifice on accuracy is acceptable for PPFS-IFW. Since it is based on PAN-SVM, which employs the Nyström technique to approximate the kernel matrix in order to speed up the computation and communication for distributed mining. The most important difference from the other methods, PPFS-IFW can preserve individual privacy while significantly improve the classifier accuracy.

4.2. Evaluation of PPM2C.

4.2.1. Datasets. To test the performance of the proposed PPM2C, experiments had been conducted. All the following results shows the average value obtained from multiple tests. To be specific, PPM2C was tested on 6 datasets with the consideration of different number of classes, sample sizes and features. DNA, Vowel and Letter datasets were downloaded from LIBSVM repository [54] and Lung cancer dataset was download from the University of California, Irvine (UCI) Machine Learning Repository [4]. The microarray dataset of Leukemia data [23] was downloaded from [53]. The details regarding datasets are summarized in Table 4.
4.2.2. Reliability of PPM2C. In PPM2C, PAN-SVM was employed to construct the multiple SVM classifiers. The performance of PPM2C was accessed by the classification accuracy, as formulated in equation (5). The experimental results of PPM2C were compared with that in LIBSVM [54], which use a regular binary SVM classifier to solve the multi-class classification problem under the OVA schema. The results were obtained from the average accuracy of ten rounds, where each round contains a 5-fold cross validation results. In other word, the experimental results are the average of 50 tests. For convenience, we use PrivacySVM and RegularSVM to represent PAN-SVM and LIBSVM binary classifiers, respectively.

The reliability of PPM2C was firstly tested to measure whether the proposed method is workable and reliable or not. PPM2C will be considered as reliable and workable if it can achieve approximately the same classification accuracies as using a regular SVM, like LIBSVM. The experimental results were shown in Fig. 5. From the results we can tell that the classification performance achieved by PPM2C is as good as LIBSVM for the five out of six datasets (Leukemia_{3c}, Leukemia_{4c}, vowel, lung_{3c} and letter). Since PAN-SVM which is used in PPM2C use landmarks (selected samples) [14][51] to approximate the kernel matrix, the classification accuracies were sacrificed slightly, which is normal. From the lone run, this sacrifice is worth. Because PAN-SVM framework [33] reduce the communication and computation costs sharply for distributed classification. Especially, it further protects individual information data. The reason that the classification accuracy for the DNA dataset is higher than that of LIBSVM probably because the data in this dataset is sparse. Therefore, we could conclude that PPM2C is workable and reliable.

4.2.3. Stability and Over-fitting. As in many other literatures, the classification accuracy is usually accessed by cross validation (CV). During the cross validation process, data will be randomly split into $k$ ($k$-fold) subsets. At each training round,
$k-1$ subsets are used as the training data, and the remaining 1 subset will be used as the testing set. In the other word, all the samples will be involved in the training and testing process, which probably lead to information leak during the process. Similarly, performance was tested under both CV1 and CV2 errors. To be specific, 1/5 of the total samples are randomly selected to be used as a separate testing set and will not be involved in the training process under CV2 scenario. While under CV1, all samples will be used either in the training process or the testing process via 5-fold cross validation.

Fig. 6 and 7 shows the experimental results tested on PAN-SVM and LIBSVM under CV1 and CV2 scenario, respectively. As shown in Fig. 6 and 7, the predicting accuracies are increased by PAN-SVM for datasets Leukemia3c, Leukemia4c and Lung cancer, but only slight improvements are obtained (less than 1.68%). Therefore, we say that the improvement is not significant. In other word, by employing PAN-SVM, PPM2C is stable no matter under which testing situations. On the opposite, the predicting performance of LIBSVM classifier decreased (about 5.19%) under the CV2 situation, which means that CV1 makes LIBSVM achieve higher classification accuracy, especially on small datasets, such as Leukemia3c, Leukemia4c and Lung cancer. The phenomenon of decreased classification accuracy under CV2 scenario indicates that LIBSVM has the problem of over-fitting, especially on small datasets, while PAN-SVM is relative stable and reduces such a risk. To further demonstrate the stability of PPM2C, more tests were constructed. Since the core of PPM2C: PAN-SVM depends on landmarks for approximating kernel matrix, so the tests were done regarding different percentages (25%, 30%, 35%, 40%, 45%, 50% and 55%) of landmarks under CV1 and CV2 situation, respectively.

The experiments were tested on Leukemia3c, Leukemia4c, DNA and Lung cancer datasets and the experimental results are shown in Fig. 8 and 9. The curves in Fig. 8 and 9 show that no matter under CV1 case or CV2 case, the classification
accuracy has no obvious change for different landmarks (from 0.13% to 1.16% under CV1, and from 0.34% to 1.59% under CV2). This evidence demonstrates that PPM2C’s capability to classify data and the classification accuracy keeps stable under different landmarks (sample sizes).

5. Conclusion. In this work, a privacy preserving feature selection method (PPFS-IFW) and a Multi-Class Classification (PPM2C) method are proposed based on our previous proposed privacy preserving classification framework PAN-SVM.
PPFS-IFW inherits the privacy preserving property of PAN-SVM and integrates three filter methods. Features are firstly evaluated based on three predefined measurements, and then are voted and selected according to the calculated scores. The voted features will further be ranked according to their classification accuracy. At last, PPFS-IFW will return a feature ranking list, with the most important feature on the top. PPM2C converts the multi-class classification problem into multiple binary classifiers, which are PAN-SVM classifiers here. The privacy preserving step works just like PAN-SVM, data are encrypted via the Secure Sum Protocol at the bottom layer, and sampled landmarks are used to approximate kernel matrix, which has to be computed during SVM training process. PPM2C inherits the privacy preserving and effectiveness properties of PAN-SVM, but can be used to solve multi-class classification problem. The performance of PPFS-IFW and PPM2C were tested on six benchmark datasets at two situations, refereed as CV1 and CV2. These experimental results demonstrate that PPFS-IFW and PPM2C outperforms the others.

REFERENCES

[1] C. Ambroise and G. J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, Proceedings of the National Academy of Sciences, 99 (2002), 6562–6566.
[2] V. G. Ashok, K. Navuluri, A. Alhafdi and R. Mukkamala, Dataless data mining: Association rules-based distributed privacy-preserving data mining, in Information Technology-New Generations (ITNG), 2015 12th International Conference on, IEEE, 2015, 615–620.
[3] K. Bache and M. Lichman, Uci machine learning repository, http://archive.ics.uci.edu/ml, 2013.
[4] K. Bache and M. Lichman, Uci machine learning repository, http://archive.ics.uci.edu/ml, 2013.
[5] S. D. Bay, Combining nearest neighbor classifiers through multiple feature subsets. in ICML, 98 (1998), 37–45.
[6] M. Benchehache and M.-T. Kechadi, Distributed clustering algorithm for spatial data mining, in *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2015 2nd IEEE International Conference on*. IEEE, 2015, 60–65.

[7] L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun, U. A. Muller, E. Sackinger, P. Simard et al., Comparison of classifier methods: a case study in handwritten digit recognition, in *Pattern Recognition, 1994. Vol. 2-Conferece B: Computer Vision & Image Processing, Proceedings of the 12th IAPR International Conference on*, vol. 2. IEEE, 1994, 77–82.

[8] Z. Cai, R. Goebel, M. R. Salvatipour, Y. Shi, L. Xu and G. Lin, Selecting genes with dissimilar discrimination strength for sample class prediction, in *Proceedings Of The 5th Asia-Pacific Bioinformatics Conference*, World Scientific, 2007, 81–90.

[9] P. S. Bradley and O. L. Mangasarian, Feature selection via concave minimization and support vector machines, in *ICML*, 98 (1998), 82–90.

[10] C. J. Burges, A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, 2 (1998), 121–167.

[11] Z. Cai, T. Zhang and X.-F. Wan, A computational framework for influenza antigenic cartography, *PLoS Computational Biology*, 6 (2010), e1000949.

[12] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin and M. Y. Zhu, Tools for privacy preserving distributed data mining, *ACM Sigkdd Explorations Newsletter*, 4 (2002), 28–34.

[13] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning*, 20 (1995), 273–297.

[14] P. Drineas and M. W. Mahoney, On the nyström method for approximating a gram matrix for improved kernel-based learning, *Journal of Machine Learning Research*, 10 (2009), 2157–2192.

[15] S. Dudoit, Y. H. Yang, M. J. Callow and T. P. Speed, Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments, *Statistica Sinica*, 12 (2002), 111–139.

[16] R. A. Fisher, The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7 (1936), 179–188.

[17] Z. Cai and X. Zheng, A private and efficient mechanism for data uploading in smart cyber-physical systems, *IEEE Transactions on Network Science and Engineering*, 2018, 1–1.

[18] V. Franc, S. Sonnenburg et al., Optimized cutting plane algorithm for large-scale risk minimization, *Journal of Machine Learning Research*, 10 (2009), 2157–2192.

[19] J. Friedman, *Another Approach to Polychotomous Classification*, Technical report, Department of Statistics, Stanford University, Tech. Rep., 1996.

[20] C. Furlanello, M. Serafini, S. Merler, and G. Jurman, Entropy-based gene ranking without selection bias for the predictive classification of microarray data, *BMC Bioinformatics*, 4 (2003), 54.

[21] M. Han, J. Li, Ji and Z. Cai, Q. Han, Privacy preserved influence maximization in gps-enabled cyber-physical and online social networks, in *2016 IEEE International Conferences on Social Computing and Networking (SocialCom)*, 2016, 284–292.

[22] H. Albinali, M. Han, J. Wang, H. Gao, Y. Li, The roles of social network mavens, in *2016 12th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*, 2016, 1–8.

[23] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri et al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *science*, 286 (1999), 531–537.

[24] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *Journal of Machine Learning Research*, 3 (2003), 1157–1182.

[25] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using support vector machines, *Machine Learning*, 46 (2002), 389–422.

[26] I. Kholod, M. Kuprianov and I. Petukhov, Distributed data mining based on actors for internet of things, in *Embedded Computing (MECO), 2016 5th Mediterranean Conference on*, IEEE, 2016, 480–484.

[27] S. Knerr, L. Personnaz and G. Dreyfus, Single-layer learning revisited: A stepwise procedure for building and training a neural network, in *Neurocomputing*, Springer, 68 (1990), 41–50.

[28] L. Liu, M. Han, Y. Zhou, Y. Wang, LSTM Recurrent Neural Networks for Influenza Trends Prediction, in *International Symposium on Bioinformatics Research and Applications*, 2018, 259–264.

[29] Y. Lu, M. Yan, M. Han, Q. Yang, Y. Zhang, Privacy Preserving Multiclass Classification for Horizontally Distributed Data, in *Proceedings of the 19th Annual SIG Conference on Information Technology Education*, 2018, 165–165.
[30] Y. Lindell and B. Pinkas, Privacy preserving data mining, *Journal of Cryptology*, 15 (2002), 177–206.

[31] M. Han, J. Wang, M. Yan, C. Ai, Z. Duan, Z. Hong, Near-complete privacy protection: cognitive optimal strategy in location-based services, in *Procedia Computer Science*, 129 (2018), 298–304.

[32] A. Joshi, M. Han, Y. Wang, A survey on security and privacy issues of blockchain technology, in *Mathematical Foundations of Computing*, 1 (2018), 121–147.

[33] Y. Lu, P. Phoungphol and Y. Zhang, Privacy aware non-linear support vector machine for multi-source big data, in *Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2014 IEEE 13th International Conference on, IEEE, 2014, 783–789.

[34] S. Maldonado, R. Weber and J. Basak, Simultaneous feature selection and classification using kernel-penalized support vector machines, *Information Sciences*, 181 (2011), 115–128.

[35] J. Miao and L. Niu, A survey on feature selection, *Procedia Computer Science*, 91 (2016), 919–926.

[36] J. Miranda, R. Montoya and R. Weber, Linear penalization support vector machines for feature selection, in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2005, 188–192.

[37] K. Parmar, D. Vagheela and P. Sharma, Performance prediction of students using distributed data mining, in *Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2015 International Conference on, IEEE, 2015, 1–5.

[38] I. Rish, An empirical study of the naive bayes classifier, in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3 (2001), 41–46.

[39] S. L. Salzberg, C4. 5: Programs for machine learning by j. ross quinlan. morgan kauflmann publishers, inc., 1993, *Machine Learning*, 16 (1994), 235–240.

[40] A. Sharma, S. Imoto and S. Miyano, A top-r feature selection algorithm for microarray gene expression data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TUBB)*, 9 (2012), 754–764.

[41] Y. Shen, H. Shao and Y. Li, Research on the personalized privacy preserving distributed data mining, in *Future Information Technology and Management Engineering, 2009. FITME’09. Second International Conference on*. IEEE, 2009, 436–439.

[42] C.-A. Tsai, C.-H. Huang, C.-W. Chang and C.-H. Chen, Recursive feature selection with significant variables of support vectors, *Computational and Mathematical Methods in Medicine*, 2012 (2012), Art. ID 712542, 12 pp.

[43] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio and V. Vapnik, Feature selection for svms, in *Advances in Neural Information Processing Systems*, 2001, 668–674.

[44] Z. Xu and X. Yi, Classification of privacy-preserving distributed data mining protocols, in *Digital Information Management (ICDIM)*, 2011 Sixth International Conference on, IEEE, 2011, 337–342.

[45] K. Yang, Z. Cai, J. Li and G. Lin, A stable gene selection in microarray data analysis, *BMC bioinformatics*, 7 (2006), p228.

[46] J. Ye and T. Xiong, Computational and theoretical analysis of null space and orthogonal linear discriminant analysis, *Journal of Machine Learning Research*, 7 (2006), 1183–1204.

[47] L. Ying-hua, Y. Bing-rui, C. Dan-yang and M. Nan, State-of-the-art in distributed privacy preserving data mining, in *Communication Software and Networks (ICCSN)*, 2011 IEEE 3rd International Conference on, IEEE, 2011, 545–549.

[48] K. Zhang, L. Lan, Z. Wang and F. Moerchen, Scaling up kernel svm on limited resources: A low-rank linearization approach, in *Artificial Intelligence and Statistics*, 2012, 1425–1434.

[49] X. Zhang, X. Lu, Q. Shi, X.-q. Xu, E. L. Hon-chiu, L. N. Harris, J. D. Iglehart, A. Miron, J. S. Liu and W. H. Wong, Recursive svm feature selection and sample classification for mass-spectrometry and microarray data, *BMC Bioinformatics*, 7 (2006), p197.

[50] F. Zhang, C. Rong, G. Zhao, J. Wu and X. Wu, Privacy-preserving two-party distributed association rules mining on horizontally partitioned data, in *Cloud Computing and Big Data (CloudCom-Asia)*, 2013 International Conference on, IEEE, 2013, 633–640.

[51] K. Zhang, I. W. Tsang and J. T. Kwok, Improved nyström low-rank approximation and error analysis, in *Proceedings of the 25th International Conference on Machine Learning*. ACM, 2008, 1232–1239.

[52] X. Zheng, Z. Cai, and Y. Li, Data linkage in smart internet of things systems: A consideration from a privacy perspective, *IEEE Communications Magazine*, 56 (2018), 55–61.
[53] Z. Zhu, Y.-S. Ong and M. Dash. Markov blanket-embedded genetic algorithm for gene selection, Pattern Recognition, 40 (2007), 3236–3248.

[54] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/, 2013.

Received August 2018; revised October 2018.

E-mail address: heatherluym@gmail.com
E-mail address: mingyuan.yan@ung.edu
E-mail address: mhan9@kennesaw.edu
E-mail address: qyang5@students.kennesaw.edu
E-mail address: yzhang@gsu.edu