Sparse Feature Learning using Ensemble Model for Highly-correlated High-dimensional Data

Ali Braytee*[1], Ali Anaissi[3] and Paul J. Kennedy[2]

[1] School of Biomedical Engineering, University of Technology Sydney, Australia.
[2] School of Software, University of Technology Sydney, Australia.
[3] The University of Sydney, NSW 2006, Australia.
{Ali.Braytee@uts.edu.au,Ali.Anaissi@sydney.edu.au,Paul.Kennedy@uts.edu.au}

Abstract. High-dimensional highly correlated data exist in several domains such as genomics. Many feature selection techniques consider correlated features as redundant and therefore need to be removed. Several studies investigate the interpretation of the correlated features in domains such as genomics, but investigating the classification capabilities of the correlated feature groups is a point of interest in several domains. In this paper, a novel method is proposed by integrating the ensemble feature ranking and co-expression networks to identify the optimal features for classification. The main advantage of the proposed method lies in the fact, that it does not consider the correlated features as redundant. But, it shows the importance of the selected correlated features to improve the performance of classification. A series of experiments on five high dimensional highly correlated datasets with different levels of imbalance ratios show that the proposed method outperformed the state-of-the-art methods.

Keywords: Feature Selection, High-dimensional Data, Feature Correlation.

1 Introduction

In the era of high-throughput technologies, the term "big data" is coined to reflect the amount of the data increasingly being generated in many fields. The available data exceeds the ability of the existing machine learning algorithms to analyse it. The complexities and challenges of data in some fields are reflected in the generated datasets. One of these types of complex structures is high-dimensional data which have a relatively low number samples, known as "the curse of dimensionality" problem or $p >> n$. The problem of the curse of dimensionality has become increasingly common in several domains, especially in biomedicine and genomics applications. Furthermore, the dilemma is exacerbated by the presence of highly correlated features and the imbalanced data problem. In machine learning, many feature selection algorithms have been proposed to select the important features and eliminate the unimportant ones. However, most of these existing algorithms follow an individual feature ranking

approach which discards the existence of the correlated features including SVM-RFE [9], LASSO [17, 7]. Surprisingly, a few feature ranking algorithms based on correlated features are proposed in the literature [5]. Feature selection for predictive models in the presence of high-dimensional and imbalanced data with many highly correlated covariates is a challenging problem that affects many disciplines. Initially, researchers were unaware of the importance of correlated covariates in interpreting predictive models. However, recent studies have been conducted to interpret groups of highly correlated features to identify significant functional modules, to improve classification accuracy, and to reflect on the semantic components of these features.

Recently, we proposed an ensemble SVM (ESVM-RFE) algorithm [2] for individual feature ranking in high-dimensional data. The ESVM-RFE uses the ensemble strategy with the SVM [4] classifier as the base learning model. It uses the binary SVM as a decision boundary to separate two classes, defined by solving a quadratic optimization problem. The decision boundary is specified by a subset of critical training samples named support vectors that lie on the edge. Ensemble techniques have the advantage of handling the problem of the curse of dimensionality and reducing the potential of over-fitting the training data. The ESVM-RFE follows the ensemble and bagging concepts of random forest and adopts a backwards elimination strategy. Also, it handles the problem of imbalanced datasets by constructing roughly balanced bootstrap samples or bootstrap samples biased to the minority class.

In this paper, a novel sparse feature learning algorithm (SFL-ESVM) is proposed to handle the correlated features in high-dimensional data. The SFL-ESVM algorithm consists of three components: first, it generates isolated feature modules based on the network structure of the data. Each module contains the correlated features, and the correlation between the modules is low. Second, our previous study of the ESVM-RFE algorithm [2] is used to select the most important features within each module. Finally, the selected features are aggregated and again, ESVM-RFE is applied to select the optimal features of the modules.

Specifically, **the contributions** of this paper are:

1. Propose an effective ensemble feature ranking method using co-expression networks to select optimal features for classification.
2. Provide comprehensive evaluations of our method on real-world high dimensional imbalanced datasets which show the advantages of our method.

2   Related Work

Feature selection process is considered as a prerequisite step for many data mining high dimensional datasets including genomic data. It reduces the number of dimensions by selecting a certain number of features or genes which able to explain the differences between patients in regards to the type of the disease [14]. In fact, many benefits are observed from achieving a feature selection include the ability of better understand the data with less informative features, reducing the complexity and computation time of the learning model, removing the noisy

features and others. There are three types of feature selection approaches as defined by Sayes et al [14]: filter, wrapper and embedded approaches. The main difference among them that filter approach is independent of any classification algorithm. However, wrapper and embedded use the classification algorithm in the feature selection process. Wrapper evaluates the goodness of features using the classification algorithm and embedded performs the feature selection during the learning process.

Many feature selection algorithms have been proposed to select the important features and eliminate the unimportant ones. However, most of these existing algorithms follow an individual feature ranking approach which discards the existence of the correlated features. Surprisingly, a few feature ranking algorithms based on correlated features are proposed in the literature. The main two approaches which proposed in literature that consider the correlated features are: the sparse models and feature clustering methods. Sparse models including group LASSO [11] and fused SVM [13] suffer from a correlation bias during the feature weighting process, because they assign the weights based on the group size [18]. Therefore, features which belong to a big group may receive small weights. Furthermore, they are considered as parametric methods which need to set some parameters beforehand, which is not guaranteed to hold in practical applications [6]. Feature clustering methods determine the group features using clustering methods, then select a limited number of features to train the models. It is reported that this approach may remove the correlation bias [12]. But, several issues were found in this approach: firstly, the features are clustered using the standard parametric clustering methods which needs to optimize the numbers of clusters parameter. Secondly, the feature importance scores are unstable due to a single ranking of the features in the proposed model. Finally, the problem of class imbalance problem is not handled by the existing feature selection at the presence of highly correlated features, which may assigns a larger weights for the features which predict the majority class. Recent work, called the fuzzy forests method, has been proposed by [6] which uses recursive feature elimination random forests to select the features from the correlated feature blocks. Fuzzy forests depends on random forest feature selection which has a high computational complexity in terms of running time compared to the feature selection method using the support vector machine [2]. Furthermore, the fuzzy forests method does not take into account the imbalanced data problem which may generate features which are biased towards the majority class. A drawback of supervised clustering methods is that they do not identify the correlated features to improve the classification performance along with the interpretation.

3   The Proposed Method SFL-ESVM

In this section, SFL-ESVM is proposed as a feature module learning framework. Sub-section 3.1 shows how to cluster the co-expression networks to generate the feature modules. Sub-section 3.2 reviews the ESVM-RFE algorithm, and finally, presents the proposed SFL-ESVM algorithm.

3.1 Clustering Co-Expression Networks

A very widely used method to cluster the co-expression biological networks is hierarchical clustering and in particular, the weighted gene co-expression network analysis (WGCNA) algorithm. The WGCNA is initially developed to find the relevant biological modules by detecting a network of highly correlated genes [3]. The gene co-expression network generated by WGCNA can be clustered into groups of highly interconnected nodes.

The WGCNA uses a similarity function such as Pearson correlation to construct a correlated similarity network between the genes. Then, the similarity network is transformed into an adjacency network by taking the absolute value of the similarity network entries and raising it to the power β. This step indicates the strong correlation among genes and rejects the weak ones. Scale-free topology criterion is used to choose the best value of parameter β. Next, the modules are identified by searching for strongly connected genes which is known as high topological overlap. After constructing the topological overlap network for all pairs of genes, the hierarchical clustering algorithm uses this information to identify the modules of correlated genes. The WGCNA has the advantage that it does need to set the number of clusters in advance.

3.2 Review of ESVM-RFE

The ESVM-RFE [2] ranks the features by constructing an ensemble of SVM models in each iteration of SVM-RFE using a random bootstrap subset from the training set. Then, it aggregates all the feature rankings as an ensemble vote. The least important features are eliminated based on multiple votes in each iteration. This process is repeated until a specified number of features is reached.

3.3 The Proposed SFL-ESVM based on the Co-Expression Feature Network

The proposed SFL-ESVM does not consider the correlated features as redundant which must be removed. For example, in microarray gene expression data, genes that have either similar genomic locations or molecular functions are assumed to co-function and are highly correlated [18]. The correlation issue negatively impacts the classical feature selection algorithms which follow an individual feature ranking process.

The SFL-ESVM algorithm aims to achieve a feature selection in the presence of the correlated features. It follows a backwards feature elimination method. The flowchart of the SFL-ESVM algorithm is shown in Figure 1. It is divided into two phases: an intra-screening phase and an inter-screening phase. The intra-screening phase is composed of multiple steps: firstly, it constructs a feature co-expression network which captures the correlation between the features. This step can be achieved using WGCNA or any other graph clustering method. Then,
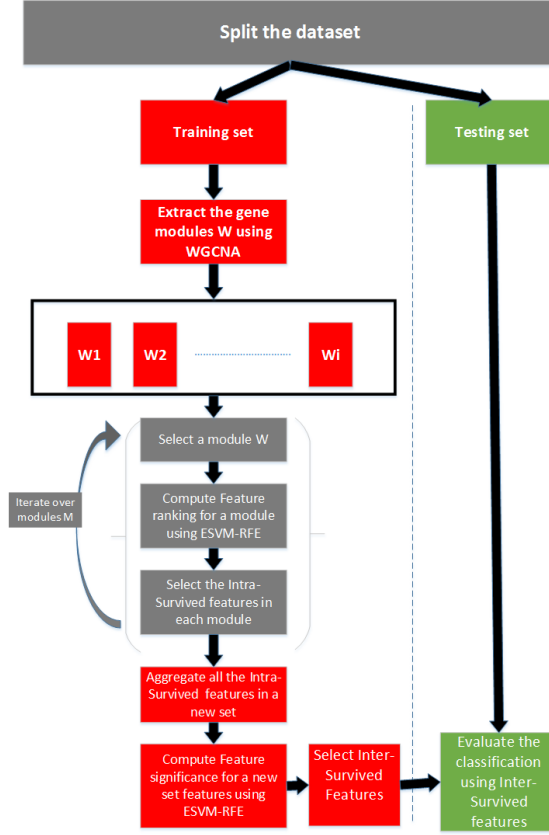
Fig. 1. The flowchart of SFL-ESVM

the modules M_i of the correlated features are extracted from the feature co-expression network using hierarchical clustering. The screening phase is applied on each module M_i to filter out the unimportant features. It is known as intra-screening because it operates on each module independently. For each correlated feature module M_i, ESVM-RFE is used to generate weights for the features. This is described in Algorithm 1. It starts with the entire set of features in the module, and in each iteration, an ensemble SVM is trained by taking bootstrap samples from the training dataset. The feature weights are estimated using the absolute value of coefficients of the support vectors for each SVM model. The estimated feature weights are aggregated from the ensemble of SVM models and ranked in decreasing order to remove the least important features with the small weights. Features are eliminated over multiple iterations on each module until a specific threshold of the number of selected features is reached.

The selected features from each module in the intra-screening phase are aggregated and passed as an input to the inter-screening phase. The inter-screening

phase in the SFL-ESVM algorithm is to capture the interaction among the modules. It aggregates all the surviving features from the previous phase and applies one more ESVM-RFE to select the global surviving features.

The proposed SFL-ESVM algorithm is an appropriate solution to reduce the correlation bias in the presence of imbalanced data. It differs from other feature selection algorithms in the following ways: first, it makes use of the ensemble SVM to reduce the influence of correlation bias, because in each iteration, the ranking decision is generated from multiple SVM models on different bootstraps and it is not related to the module size. Second, feature ranking in each iteration is achieved on equal bootstrap samples to mitigate the effect of the imbalance class problem. Third, bagging ensembles improve the performance in the presence of small sample size. Fourth, it uses WGCNA to estimate the network structure of the data, and consequently, estimate the correlated features. Finally, stability is targeted by achieving multiple perturbations of constructing SVM models in each iteration.

Input: training data X
        Class labels y
parameter : $inter\text{-}d$; // Number of selected features between modules
                $intra\text{-}d$; // Number of selected features from each module
                b ;   // Size of ensemble SVM in each iteration
                E ; // The % of features to eliminate at each iteration
                $bagSize$;   // Balanced bootstrap from training dataset

$modules \leftarrow$ WGCNA(X) ;   // the interconnected features using WGCNA
$l \leftarrow$ length$(modules)$;
for $i \leftarrow 1$ **to** l **do**
  | $M \leftarrow modules_i$;
  | $data \leftarrow X(, M)$;
  | ;   // M is the correlated genes in each module
  | $intra\text{-}features \leftarrow$ ESVMRFE$(data, y, b, E, intra\text{-}d, bagSize)$;
  | $intra\text{-}Set \leftarrow intra\text{-}Set \cup intra\text{-}features$;
end
$selectedData = trainingdata[, intra\text{-}Set]$;
$inter\text{-}features \leftarrow$ ESVMRFE$(selectedData, y, b, E, inter\text{-}d, bagSize)$;
 // inter-features: the surviving features between the modules
 using Algorithm ESVM-RFE
Output: $inter\text{-}features$
   Algorithm 1: Sparse Feature Learning algorithm (SFL-ESVM)

4   Experiments

In this section, the experimental evaluations on high-dimensional, highly correlated datasets are reported. This section analyses and compares the classification performance of the proposed SFL-ESVM against the state-of-the-art algorithms

Function *ESVMRFE (data, class, b, E, d, bagSize)*
   Input: $surviveIndexes = seq(1 : ncol(data))$
         $n = nrow(data)$

   for $d \leftarrow 1$ **to** $length(surviveIndexes)$ **do**
      $m = length(surviveIndexes)$;
      $survive = m - m \times E$;   `// survive: number of features in the current iteration`
      $ensRes = matrix(n, b)$; `// ensRes: feature's weight of each SVM model`
      for $i \leftarrow 1$ **to** b **do**
         $bag \leftarrow bootstrap(data, bagSize)$;
         $bagClass \leftarrow bootstrap(class, bagSize)$;
         $model \leftarrow svm(bag[, survivingIndexes], bagClass)$;
         $weightVector \leftarrow transpose(model\$coefs) \% * \% model\$SV$; `// Compute the weight vector`
         $featureWeight \leftarrow weightVector * weightVector$;   `// Compute ranking criteria`
         $ensRes \leftarrow merge(ensRes, featureWeight)$;   `// Accumulate feature's weight`
      end
      $totalWeight = rowSum(ensRes)$; `// Aggregate feature's weight`
      $sortedWeight \leftarrow sort(totalWeight)$; `// Sort the total feature's weight by decreasing order`
      $sortedIndexes \leftarrow index(sortedWeight)$;
      $surviveIndexes \leftarrow surviveIndexes[sortedIndexes[1 : survive]]$; `// Eliminate features with smallest weight`
   end
   Output: $selectedData = data[, surviveIndexes]$

Algorithm 2: ESVM-RFE for feature learning

---

namely SVM-RFE [9], Fuzzy Forests [6], and Hybrid L1/2 L2 regularization (HLR) [10]. SVM-RFE is evaluated as a baseline method, Fuzzy Forests as an ensemble feature ranking algorithm for correlated features, and HLR from the point of view of a sparse model for dimensionality reduction in the presence of correlated features. It is important to note that the main purpose of these experiments is to evaluate the potential of the proposed SFL-ESVM algorithm to improve the classification performance in the presence of a large number of correlated features.

Without loss of generality, linear SVM is used as a classifier to evaluate the performance of the selected features from the compared algorithms. The performance is measured by the widely used metric AUC under the receiver operating characteristic (ROC) analysis. The optimal tuning parameters of the SFL-ESVM, Fuzzy Forests, HLR and SVM-RFE approaches were identified by five-fold cross-validation on the training set. The datasets are divided at random such that approximately 75% is used as a training set and 25% as a test set. The datasets are z-score normalized.

4.1 Datasets

The experiments are conducted on one dataset collected from The Children's Hospital at Westmead, and four public datasets. The details of these datasets are summarised in Table 1. The common characteristics of these datasets are highly dimensional, highly correlated, have a small number of samples and some of them are imbalanced. A stratified random sampling function (stratified) in R is applied on the evaluated datasets to split the data into a training and testing set, with a quarter of the dataset considered as a testing set and the reminder as a training set.

Table 1. Datasets

Dataset	#Attributes	#Instances	Source
Childhood Leukaemia	22277	60	TB-CHW
DLBCL-FSCC	7129	77	[15]
Prostate cancer	6033	102	[16]
ALL/AML	7129	73	[8]
Breast cancer	8141	295	[19]

4.2 Results and Discussion

The goal of this section is to evaluate the performance of the selected features from the compared algorithms on the real-world datasets. In the following experiments, for a fair comparison of all algorithms, the AUC accuracy is estimated using the .632+ bootstrap method [1] with 100 bootstrap samples. For each bootstrap sample, AUC accuracy is obtained on the test dataset.

Figure 2 shows the AUC evaluated on the test dataset across a different number of features. The figures present the results of up to 100 features because the evaluated datasets contain a small number of samples which needs a small number of features to avoid over-fitting. As shown in Figure 2, the proposed SFL-ESVM algorithm outperforms the state-of-the-art feature selection methods in most feature sets in all datasets. The AUC classification performance is further investigated based on the best number of selected features. As shown in Table 2, several statistical measures are included, namely minimum, maximum, first quartile, third quartile, median and mean on 100 bootstrap samples on the test data. For example, as shown in Table 2 for the DLBCL-FSCC dataset, the best AUC is achieved for the compared algorithms: SFL-ESVM, Fuzzy Forest, HLR and SVM-RFE is 50, 74, 74, and 78 features respectively. This clearly shows that the proposed algorithm achieves better results than the compared algorithms using different statistical measures. Furthermore, the proposed algorithm SFL-ESVM obtained the best accuracy results compared to the others with a small number of features in most datasets, which leads to less computational complexity during the training process. It can also be observed that the classification results of the SFL-ESVM algorithm tend to be stable after increasing the selected features above approximately 50 features in the evaluated

datasets. This indicates the stability and capability of SFL-ESVM to select a lower percentage of features and realise good accuracy results.

On the other hand, the experimental results indicate that the feature selection methods that handle correlated features such as the proposed SFL-ESVM, Fuzzy forest, and HLR, perform better than SVM-RFE which does not consider the correlated features and achieves individual feature ranking. Therefore, it demonstrates the importance of handling correlated features in high-dimensional datasets to improve the performance of the classifiers. Finally, a statistical t-test is also conducted between the vector results of the proposed algorithm against state-of-the-art methods under the null hypothesis that AUC on vectors of the used method is not significantly different to SFL-ESVM. The p-value is lower than 0.05 which rejects the null hypothesis.

Table 2. The quartile and mean values of AUC accuracies of the compared algorithms on the evaluated datasets at the best number of features.

Dataset	Method	Min	1st Qu	Median	Mean	3rd Qu	Max	Best Features
Childhood leukaemia	SVM-RFE	0.250	0.400	0.500	0.509	0.600	0.800	64
	Fuzzy Forest	0.250	0.387	0.450	0.448	0.500	0.700	48
	HLR	0.081	0.331	0.381	0.370	0.431	0.681	30
	SFL-ESVM	**0.500**	**0.600**	**0.700**	**0.692**	**0.750**	**0.850**	24
DLBCL -FSCC	SVM-RFE	0.585	0.741	0.811	0.797	0.848	0.904	78
	Fuzzy Forest	0.647	0.743	0.810	0.800	0.854	0.897	74
	HLR	0.679	0.834	0.904	0.891	0.942	0.997	74
	SFL-ESVM	**0.833**	**0.900**	**0.966**	**0.939**	**1.00**	**1.00**	50
Prostate	SVM-RFE	0.562	0.601	0.639	0.639	0.678	0.716	64
	Fuzzy Forest	0.744	0.783	0.821	0.829	0.860	0.898	68
	HLR	0.690	0.730	0.730	0.745	0.769	0.807	74
	SFL-ESVM	**0.807**	**0.884**	**0.884**	**0.886**	**0.923**	**0.923**	34
Breast	SVM-RFE	0.505	0.659	0.710	0.701	0.738	0.818	24
	Fuzzy Forest	0.471	0.544	0.624	0.618	0.669	0.760	30
	HLR	0.462	0.570	0.633	0.618	0.678	0.741	70
	SFL-ESVM	**0.556**	**0.677**	**0.727**	**0.716**	**0.772**	0.818	30
ALL/AML	SVM-RFE	0.683	0.754	0.754	0.784	0.826	0.826	32
	Fuzzy Forest	0.804	0.834	0.876	0.859	0.876	0.905	50
	HLR	0.800	0.800	0.841	0.827	0.841	0.871	68
	SFL-ESVM	**0.815**	**0.928**	**0.958**	**0.953**	**1.00**	**1.00**	46

A further investigation is made of the selected features using the proposed SFL-ESVM from the ALL/AML dataset to see if the proposed algorithm can define separated clusters based on ALL and AML class outcomes. To do this, Singular Value Decomposition (SVD) is applied to the original ALL/AML training set using all the features and project the testing samples using the first three principal components. Then, the testing samples are visualised with different shapes for the ALL and AML samples. A similar process is applied on the training set with the top 46 features selected by the proposed SFL-ESVM feature selection algorithm. Without loss of generality, the top 46 features are used in these figures. As shown in Figure 3, it is clear that the clusters of the ALL
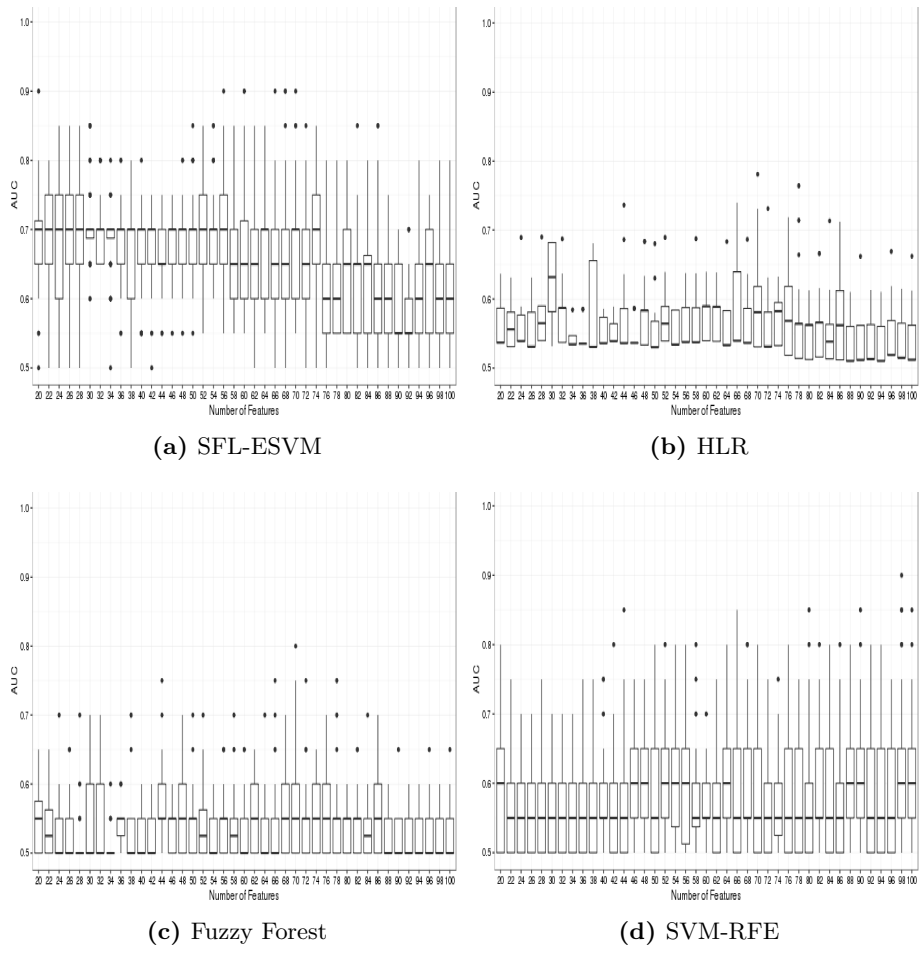
(a) SFL-ESVM

(b) HLR

(c) Fuzzy Forest

(d) SVM-RFE

Fig. 2. Classification performance comparison between algorithms evaluated on Childhood Leukaemia dataset using the 0.632+ bootstrap method with 100 bootstrap samples across a different number of features

and AML classes in Figure 3b are well separated compared to the clusters in Figure 3a, which overlap. This example confirms that efficacy of the proposed algorithm to select the optimal features in the presence of complex datasets which importantly, are able to explain the differences between different classes.
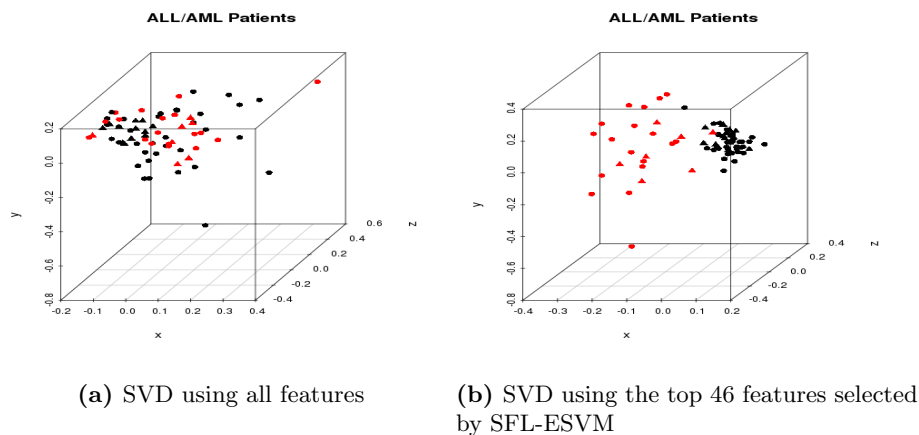

(a) SVD using all features

(b) SVD using the top 46 features selected by SFL-ESVM

Fig. 3. SVD on ALL/AML dataset to show the clusters of ALL and AML patients. Black=ALL, red=AML, circle=training samples, and triangle=testing samples

5   Conclusion

This paper proposed a novel algorithm to select the best features in the presence of highly correlated features that improves the classification performance. The proposed SFL-ESVM does not consider the correlated features as redundant, rather it selects the top correlated features from each feature module using the ESVM-RFE algorithm. Then, it aggregates all features from different modules and again applies ESVM-RFE to rank the combined features. The proposed algorithm can improve the classification accuracy in the presence of very complex datasets. These datasets contain high-dimensional, highly correlated features and a low number of samples. Extensive experiments are conducted on different datasets. Our results show the high-performing quality of the proposed method on benchmark datasets which outperformed the-state-of-the-art methods.

References

1. Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. PNAS 99(10), 6562 – 6566 (2002)

2. Anaissi, A., Goyal, M., Catchpoole, D.R., Braytee, A., Kennedy, P.J.: Ensemble Feature Learning of Genomic Data using Support Vector Machine. PLOS ONE 11(6), 1 – 17 (2016)
3. Bin, Z., Steve, H.: A General Framework for Weighted Gene Co-expression Network Analysis. Statistical Applications in Genetics and Molecular Biology 4(1), 11 – 28 (2005)
4. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A Training Algorithm for Optimal Margin Classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. pp. 144 – 152. ACM, Pittsburgh (1992)
5. Braytee, A., Liu, W., Kennedy, P.J.: Supervised Context-aware Non-negative Matrix Factorization to Handle High-dimensional High-correlated Imbalanced Biomedical Data. In: 2017 International Joint Conference on Neural Networks (IJCNN). pp. 4512 – 4519. IEEE, Anchorage (2017)
6. Conn, D., Ngun, T., Li, G., Ramirez, C.: Fuzzy Forests: Extending Random Forests for Correlated, High-dimensional Data. UCLA Biostatistics Working Paper Series (2015)
7. Cui, C., Wang, D.: High Dimensional Data Regression using Lasso Model and Neural Networks with Random Weights. Information Sciences 372, 505 – 517 (2016)
8. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286(5439), 531 – 537 (1999)
9. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46(1), 389 – 422 (2002)
10. Huang, H.H., Liu, X.Y., Liang, Y.: Feature Selection and Cancer Classification via Sparse Logistic Regression with the Hybrid L1/2+ 2 Regularization. PLOS ONE 11(5), 1 – 15 (2016)
11. Meier, L., Van De Geer, S., Bühlmann, P.: The Group Lasso for Logistic Regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(1), 53 – 71 (2008)
12. Park, M.Y., Hastie, T., Tibshirani, R.: Averaged Gene Expressions for Regression. Biostatistics 8(2), 212 – 227 (2007)
13. Rapaport, F., Barillot, E., Vert, J.P.: Classification of arrayCGH Data using Fused SVM. Bioinformatics 24(13), i375 – i382 (2008)
14. Saeys, Y., Inza, I., Larrañaga, P.: A Review of Feature Selection Techniques in Bioinformatics. Bioinformatics 23(19), 2507 – 2517 (2007)
15. Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., et al.: Diffuse Large B-cell Lymphoma Outcome Prediction by Gene-expression Profiling and Supervised Machine Learning. Nature Medicine 8(1), 68 – 74 (2002)
16. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., et al.: Gene Expression Correlates of Clinical Prostate Cancer Behavior. Cancer Cell 1(2), 203 – 209 (2002)
17. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58(1), 267 – 288 (1996)
18. Tolosi, L., Lengauer, T.: Classification with Correlated Features: Unreliability of Feature Ranking and Solutions. Bioinformatics 27(14), 1986 – 1994 (2011)
19. Van De Vijver, M.J., He, Y.D., Van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., et al.: A Gene-expression Signature as a Predictor of Survival in Breast Cancer. New England Journal of Medicine 347(25), 1999 – 2009 (2002)