Final Accepted Version. This paper has been accepted by The 2017 4th International Conference on Systems and Informatics (ICSAI 2017), IEEE press, pp. 1427-1432, Hangzhou, Zhejiang, China, 11-13 November, 2017

Large Margin Distribution Machine Recursive Feature Elimination

Ge Ou, Yan Wang*
College of Computer Science and Technology
Jilin University
Changchun, China
*E-mail: wy6868@jlu.edu.cn

Wei Pang*, George Macleod Coghill
Department of Computing Science
University of Aberdeen
Aberdeen, UK
*E-mail: pang.wei@abdn.ac.uk

Abstract—**In order to eliminate irrelevant features for classification, we propose a novel feature selection algorithm called Large Margin Distribution Machine Recursive Feature Elimination (LDM-RFE). LDM-RFE uses the latest support vector based classification algorithm Large Margin Distribution Machine (LDM) to evaluate all the features of samples, and then generates a ranked feature list during the procedure of Recursive Feature Elimination (RFE). In the experiment section, we report promising results obtained by LDM-RFE in comparison with several common feature selection algorithms on five UCI benchmark datasets.**

Keywords-feature selection; large margin distribution machine; recursive feature elimination; classification

I. INTRODUCTION

In classification, feature selection [1] is a very important technique used to avoid overfitting and reduce computational complexity [2]. There exist many feature selection algorithms used for machine learning [3][4], however, many of them can be used in all kinds of tasks and not specific for classification.

Some feature selection algorithms, such as Principal Components Analysis (PCA) [5], t-test [6], and kullback-Leibler divergence [7], can be used for any machine learning models. But among these algorithms, Support Vector Machine Recursive Feature Elimination (SVM-RFE) [8] is specifically aimed to deal with classification tasks and it has better performance than other commonly used feature selection algorithms in many problems, especially for high-dimension problems. Furthermore, some related feature selection algorithms for classification has been proposed. Su and Hsiao [9] proposed a Multiclass Mahalanobis-Tanguchi system for feature selection and simultaneous multiclassification. Wang [10] studied a feature selection algorithm for big data problems. Liu [11] proposed a framework for multiclass sentiment classification. In addition, the study of classification model has made new progress over the last few years. Zhou and Zhang [12] proposed Large Margin Distribution Machine (LDM) algorithm, which has better classification performance than Support Vector Machine (SVM) [13] in the tested problems. LDM is based on the novel theory of optimizing the margin distribution, and it used the dual coordinate descent (DCD) [14] strategies and the averaged stochastic gradient descent (ASGD) [15] strategies to solve the optimization function.

Considering the above, in this research we propose a novel RFE algorithm for classification based on LDM, which we call Large Margin Distribution Machine Recursive Feature Elimination (LDM-RFE). The proposed LDM-RFE ranks problem features by their contributions to build the LDM model at each iteration and eliminates irrelevant features progressively. Our proposed LDM-RFE is compared with several commonly used feature selection algorithms, such as t-test, PCA, and SVM-RFE. The experimental results indicate that our proposed LDM-RFE leads to better performance than several other algorithms on five UCI [16] benchmark data sets.

II. BACKGROUND

Let $S=\{(x_1,y_1),...,(x_n,y_n)\}$ be a training set of n samples, where $x_i \in R^m$ are the input samples and $y_i = \{-1,+1\}$ is the label set. The objective function in classification problems is $f(x) = w \cdot \phi(x)$, where $w \in R^m$, and ϕ is the mapping function induced by a kernel K, i.e., $K(x_i,x_j) = \phi(x_i) \cdot \phi(x_j)$, which makes the data mapped to the feature space.

A. Large Margin Distribution Machine

Large Margin Distribution Machine (LDM) [12] [17] aim to optimize the margin distribution, that is, maximize the margin mean and minimize the margin variance at the same time to build the model of classification and improve the classification performance. Let X be a matrix whose element is $\phi(x_i)$, i.e., $X = [\phi(x_1),...,\phi(x_n)]$, $Y = [y_1,...,y_n]^T$ is the label set. Thus, the margin mean has the following form:

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i w \cdot x_i = \frac{1}{n} w \cdot (XY),$$

and the margin variance has the following form:

$$\hat{\mu} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i w \cdot x_i - y_j w \cdot x_j)^2$$

$$= \frac{2}{n^2}(nw^T XX^T w - w^T XYY^T X^T w).$$

Inspired by maximizing the margin mean and minimizing the margin variance simultaneously, the optimization problem of LDM with soft-margin form is as follows:

$$\begin{aligned}\min_{w,\xi} \frac{1}{2}\|w\|^2 + \frac{2\lambda_1}{n^2}(nw^T XX^T w - w^T XYY^T X^T w) \\ -\frac{\lambda_2}{n} w\cdot(XY) + C\sum_{i=1}^{n}\xi_i \\ s.t.\ y_i w\cdot\phi(x_i) \geq 1-\xi_i,\ \ \xi_i \geq 0,\ \ i=1,2,...,n,\end{aligned} \tag{1}$$

where λ_1 and λ_2 are the parameters for a tradeoff between the margin distribution and the model complexity. $\xi = [\xi_1,...,\xi_n]$ are the slack variables which estimate the samples losses. C is a trade-off parameter.

LDM provides two methods to solve (1). When the scale of data sets is medium, LDM uses the dual coordinate descent (DCD) [18] method to solve (1). And when the scale of data is large, LDM uses the average stochastic gradient descent (ASGD) [19] method to solve (1). In the DCD method, it usually selects one variable progressively to minimize while keeping other variables as constants. In the ASGD method, it computes a noisy unbiased estimate of the gradient and randomly samples a subset of the training instances rather than all data, so this method can decrease the computational complexity.

B. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is an effective algorithm for feature selection, which depends on the specific learning model. Guyon *et al.* [20] proposed RFE which is applied in cancer classification by using SVM. RFE employs all features to build a SVM model, and then ranks the contribution of each feature in the SVM model which generates a ranked feature list, and finally eliminates irrelevant features meaning less contribution to the SVM model.

To evaluate the contribution of each feature, RFE uses $\|w\|^2$ as the ranking criterion, where w is a weight vector. Therefore, the irrelevant features are the features whose value of $\|w\|^2$ is small. In other words, the features with bigger values of $\|w\|^2$ are reserved, and the features with smaller value of that are removed.

In SVM [21], w in the feature space can be written as $w = \sum_{i=1}^{n}\alpha_i y_i \phi(x_i)$, where α_i is the Lagrange multiplier, x_i is the i-th instance, and y_i is the label of i-th instance. To obtain the rank of the features, the ranking criterion of the j-th feature can be calculated in the following form:

$$\|w_j\|^2 = \left|\sum_{i=1}^{n}\alpha_i y_i \phi(x_{ij})\right|^2,\ \ j=1,2,...,m, \tag{2}$$

where $\phi(x_{ij})$ means the i-th instance of j-th feature vector.

Then the detailed steps of RFE are as follows:

1) Build the SVM model using all candidate features;

2) Evaluate the contribution of current features to SVM by calculating ranking criterion with current features using (2);

3) Rank current features according to their contributions (a high value of ranking score stands for great contribution) and produce a ranked feature list;

4) Remove a specific number of features at the bottom of the ordered list, and use the rest features as the new candidate features;

5) Go to the step 1), until it satisfied the specific number of features.

III. LARGE MARGIN DISTRIBUTION MACHINE RECURSIVE FEATURE ELINIMATION

In this section, we present LDM-RFE, which uses LDM in the procedure of RFE, and this provides an efficient way for feature selection.

When the number of features of data sets is too large, the computational complexity of a machine learning model will also increase very much. Therefore, to reduce the complexity of a machine learning model brought by a large number of features, feature selection algorithms usually select important or relevant feature subsets from all of features using a ranking criterion.

In fact, general feature selection algorithms may not lead to better performance than other feature selection algorithms which contain machine learning models because these general algorithms do not dependent on specific machine learning models. Therefore, some commonly used algorithms, which rely on specific learning models, have been implemented. RFE is a useful example of the algorithms containing a specific machine learning model, which evaluating the features using the learning performance of a machine learning model [22]. The RFE firstly evaluates the contributions of all features to construct a learning model, and then removes irrelevant or less important features progressively. In our research, we mainly consider classification, thus we use a good classification algorithm in our feature selection algorithm. It has been proved that LDM has been achieved better generalization performance than SVM [17], therefore we decide to implement a more efficient RFE process with LDM for feature selection in classification.

In order to select the relevant features which are important for building a classifier, our LDM-RFE directly uses the weight vector of LDM to rank the contribution of each feature, and finally it generates a ranked feature list. Due to the generalization performance of LDM, our LDM-RFE also has a good ability to achieve good learning performance in classification.

LDM solves (1) by the DCD method, so the weight vector w of the $f(x)$ have the following form:

$$w = \sum_{i=1}^{n}\alpha_i y_i \phi(x_i),$$

where α_i is the parameter in LDM.

We usually consider the nonlinear cases in classification. In other words, all data can be mapped to a feature space in classification and we do not know the specific form of $\phi(x_i)$. However, we can acquire the form of kernel function, thus we use $\|w\|^2$ as the ranking criterion to evaluate the contribution of features. Then, $\|w\|^2$ can be calculated as follows:

$$\|w\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \phi(x_i)\phi(x_j)$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j)$$

where α_i and α_j are the parameters in LDM.

According to LDM, parameter α can be solved by the DCD method. Then we define a score function $S(j)$ to calculate a concrete value for each feature after building a LDM model, and this score function can evaluate the contributions and importance of individual features to a LDM model with the ranking criterion $\|w\|^2$. To estimate the contribution of the j-th feature, the score function $S(j)$ for the j-th feature has the following form:

$$S(j) = \alpha^T K\alpha - \alpha^T K(-j)\alpha. \quad (3)$$

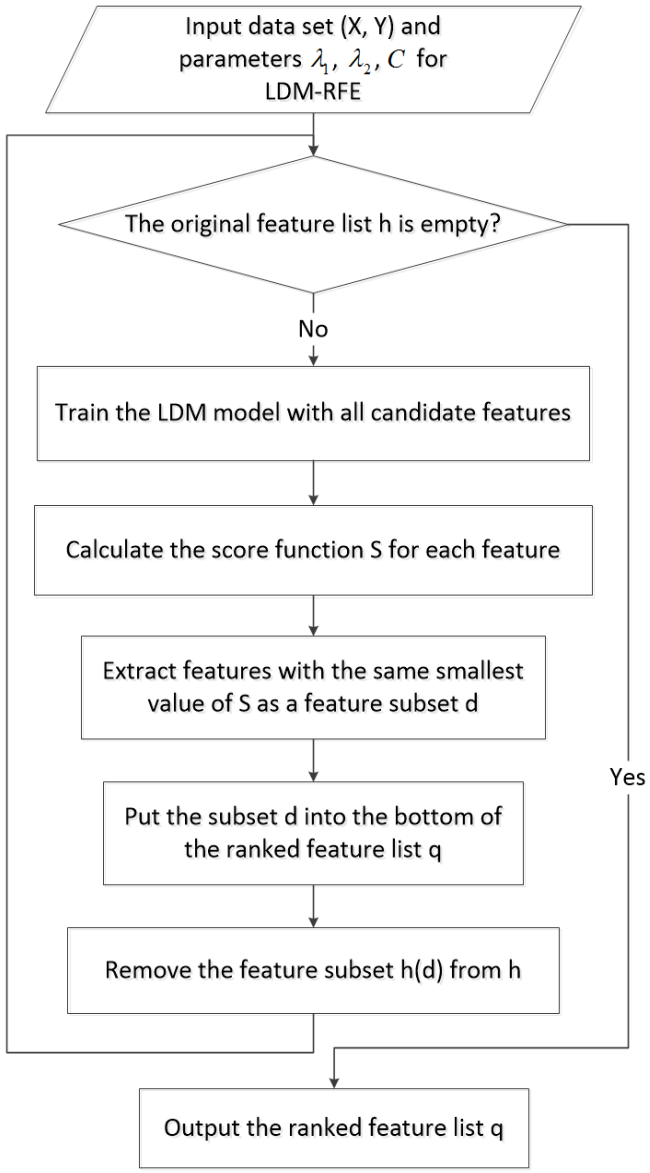
In the above, $K = K(x,x)$ stands for the kernel matrix. Then $K(-j) = \phi(x)\cdot\phi(x) - \phi(x(j,:))\cdot\phi(x(j,:))$, and $\phi(x(j,:))$ is the j-th row vector of $\phi(x)$. $K(-j)$ means the matrix K without the j-th feature from $\phi(x)$.

According to (3), we can see that the first part of $S(j)$ means the contribution of all features to a LDM model; the second part calculates the impact without the j-th feature from $\phi(x)$. It can be seen that (3) evaluates the contribution of the j-th feature by calculating the difference of the two parts above. Because the first part of (3) does not change, we can only calculate the value of the second part of (3) which can obtain the same ranked feature list. Then, we get a simpler form of score function as follows:

$$S(j) = -\alpha^T K(-j)\alpha. \quad (4)$$

A higher value of $S(j)$ from (4) indicates the j-th feature is more important to the classifier, which means j-th feature will have high probability to rank the top in the ranked feature list. In addition, we eliminate all features simultaneously when they have the same value from (4).

The contribution of each candidate feature of original feature list for building the LDM model will be calculated using (4) at iteration of each. Because the whole procedure of LDM-RFE is recursive, which means that each subset of features should be evaluated by the ranking criterion $\|w\|^2$ until the original feature list is empty. It is one of the advantages of RFE process as it mainly considers the influence of the relationships between individual features instead of that of a simple feature.



To show the recursive process, we present the detailed steps of LDM-RFE in Fig. 1.

Figure 1. The procedure of LDM-RFE.

To better understand our proposed algorithm, we provide the whole procedure of LDM-RFE and present the detailed calculation process about the ranking criterion $\|w\|^2$ in Algorithm 1.

The parameters of Algorithm 1 have the following meaning: parameter h means the original feature list of all features of a data set; parameter q stands for the final ranked feature list after feature selection process; parameter d stands for the subset of features which have the same smallest value of $S(j)$ at iteration of each.

Algorithm 1 LDM-RFE

Input: Data set X, Y, λ_1, λ_2, C;

Output: q ;

Initialization: $h = [1,2,...,m]$; $d = []$;

while $h \neq []$ do

$X' \leftarrow X(h,:)$;

$len \leftarrow length(h)$;

$\alpha \leftarrow LDM(X', Y, \lambda_1, \lambda_2, C)$;

$j = 1$;

while $j \leq len$ **do**

$K = \phi(X') \cdot \phi(X')$;

$K(j) = \phi(x(j,:)) \cdot \phi(x(j,:))$;

$K(-j) = K - K(j)$;

$S(j) = -\alpha^T K(-j)\alpha$;

$j++$;

end for

$d \leftarrow set(\arg\min(S))$;

$q = [q, h(d)]$;

$h \leftarrow h(-d)$;

end while

IV. EXPERIMENTS

In this section, we compare our LDM-RFE with other feature selection algorithms on five UCI benchmark data sets, and we present the experimental results to demonstrate whether our LDM-RFE has better performance.

A. Experimental Setup

In our experiment, we select all experimental data sets from UCI database, and the attribute of each data set is presented in Table I. Then we choose PCA, t-test, and SVM-RFE to compare with LDM-RFE. The selected features by these four algorithms are evaluated on SVM algorithm. To evaluate the performance of these feature selection algorithms, we decide to use accuracy of classification as the evaluation metric.

To reduce the influence of the difference led by the different ranges of features, we firstly normalize all features of the data sets into [0,1] before the feature selection process. Then, we divide each data set into the training set and the test set, that is, 2/3 of the data set is used as the training set for the feature selection process and the rest 1/3 of the data set is used as the test set for the evaluation process. For PCA, the cumulative contribution is set to 0.97. For SVM-RFE and LDM-RFE, we use RBF kernel function. For SVM-RFE, parameters C and δ of kernel function are both needed: parameter C is fixed at 0.1, and parameter δ is fixed at 10^{-4}. For LDM-RFE, parameters C , λ_1 , λ_2 , and δ of kernel function are needed: parameter C is set to 0.1, parameter δ is fixed to 10^{-4} , and parameter λ_1 is set to 1, and parameter λ_2 is set to 1.

TABLE I. THE CHARACTERISTICS OF BENCHMARK DATA SETS

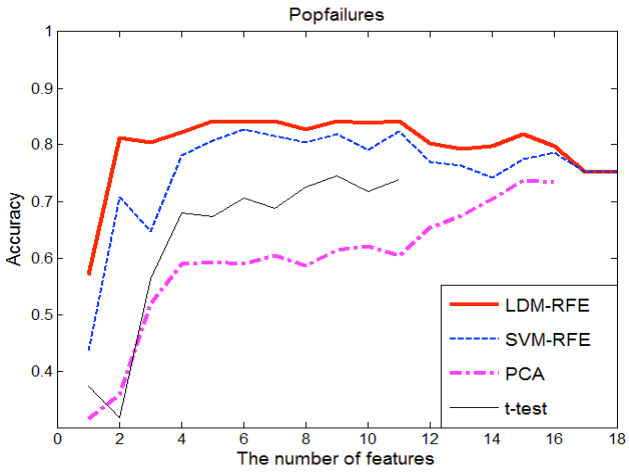
Data Sets	Instances	Features
Popfailure	540	18
Ionosphere	351	34
Sonar	208	60
Quality	287	62
Hillvalley	606	100

After feature selection, four ranked feature lists can be obtained by four algorithms. We finally evaluate these four ranked feature lists on SVM and present the result of accuracy for SVM on Fig. 2~6. All experiments are implemented with MATLAB R2014a on a PC, which has a 2.50GHz CPU and 8GB memory.

B. Results and Discussion

Figs. 2~6 give the result of accuracy for SVM with four feature selection algorithms including PCA, t-test, SVM-RFE, and LDM-RFE on five benchmark data sets, and finally generates four ranked feature lists on each data set. In Figs. 2~6, the value of x -axis stands for the first n features in the feature list, and the value of y -axis means the accuracy for SVM algorithm. In addition, the ranked feature list generated by LDM-RFE is indicated by the red bold line; the list generated by SVM-RFE is indicated by the blue dashed line; the list obtained by PCA is indicated by the purple dash dot line; the list obtained by t-test is indicated by the black solid line.

As one can see that our LDM-RFE presents much better results than other feature selection algorithms, because the red line generated by LDM-RFE is higher than other lines. We can also see from Figs. 2~6 that LDM-RFE always achieves better performance than SVM-RFE. With the number of features increasing to a certain extent, the accuracy of SVM no longer has significantly changes, and it even decreases in some cases



(Fig. 3 and Fig. 5), which demonstrate the essentiality of feature selection. We thus can select the appropriate number of features for next classification problems according to Figs. 2~6.

Figure 2. The accuracy on Popfailure data set for SVM

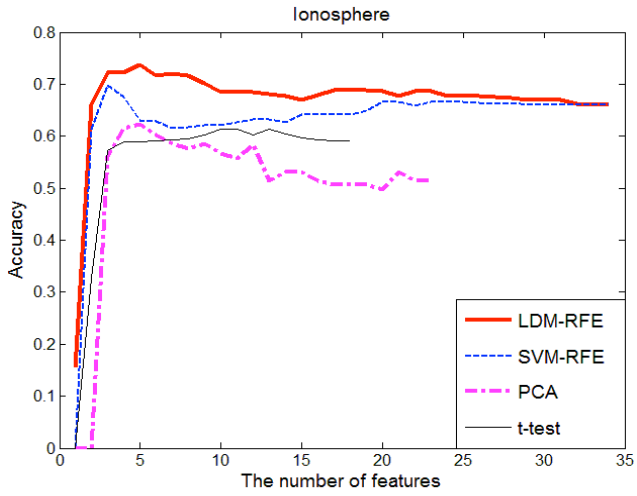


Figure 3. The accuracy on Ionosphere data set for SVM.

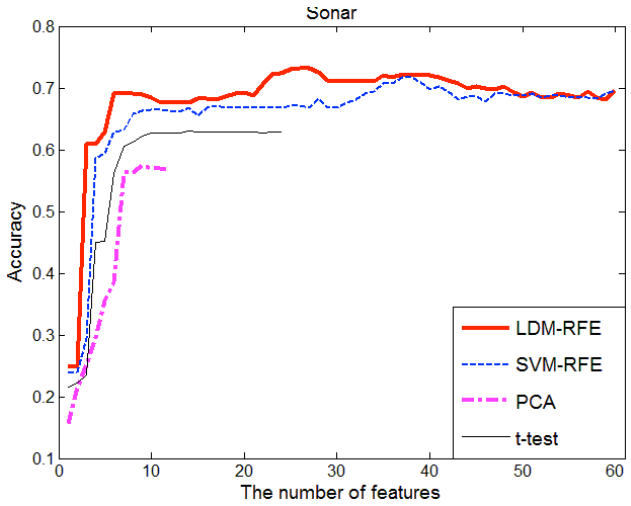


Figure 4. The accuracy on Sonar data set for SVM.

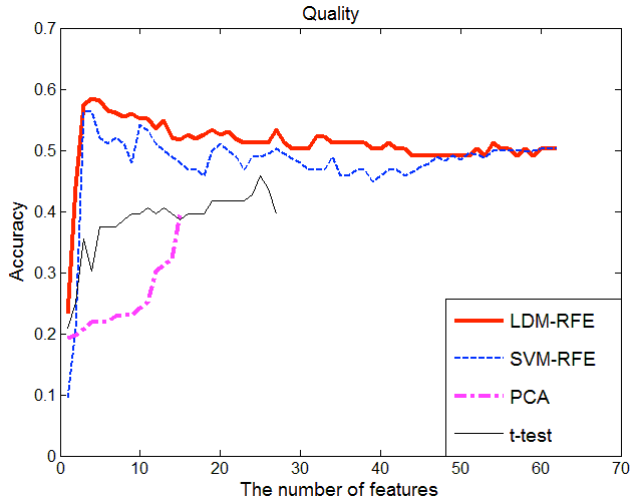
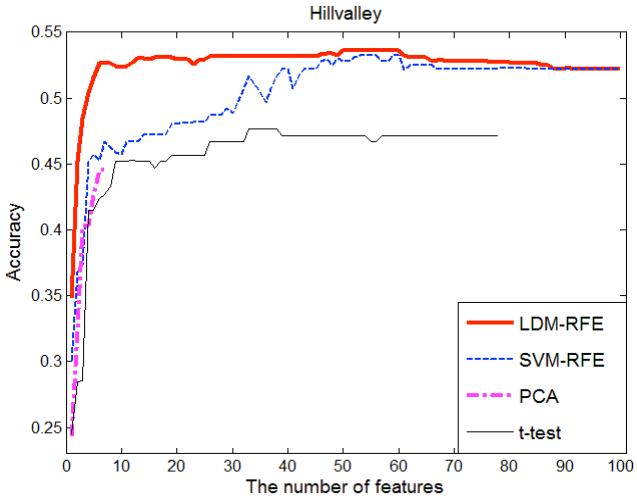


Figure 5. The accuracy on Quality data set for SVM.

Figure 6. The accuracy on Hillvalley data set for SVM.

In fact, PCA can be regarded as the coordinate transformation, which can avoid the curse of dimensionality and decrease the complexity of the learning model. However, it



does not achieve better performance than other algorithms because it changes the original characteristics of data. T-test mainly eliminates the irrelevant features considering the internal attributes or characteristics of the data sets, and it is used for many machine learning tasks for initial feature selection. When the dimension of the data sets are too large, it will need other more efficient algorithms to perform further feature selection. The better learning performance of LDM contributes to the performance of LDM-RFE, and the promising results of our experiments also demonstrate our LDM-RFE is highly competitive to SVM-RFE.

V. CONCLUSIONS

In this research, we propose a novel feature selection algorithm called LDM-RFE. It employs LDM to evaluate the contribution of all features to the classification model, and it eliminates irrelevant features progressively. Our LDM-RFE provides an efficient way for feature selection. The promising experimental results indicate that our LDM-RFE achieves better performance than several feature selection algorithms on five UCI benchmark data sets.

In the near future, we will apply LDM-RFE to other research fields or big data problems, such as bioinformatics.

ACKNOWLEDGMENT

We gratefully thank Dr Teng Zhang and Prof Zhi-Hua Zhou for providing the source code of "LDM" source code and their kind technical assistance. This work is supported by the National Natural Science Foundation of China (Nos. 61472159, 61572227) and Development Project of Jilin Province of China (Nos. 20160204022GX, 2017C033). This work is also partially supported by the 2015 Scottish Crucible Award funded by the Royal Society of Edinburgh and the 2016

PECE bursary provided by the Scottish Informatics & Computer Science Alliance (SICSA).

REFERENCES

[1] Y. H. Peng, Z. Q. Wu, and J. M. Jiang, "A novel feature selection approach for biomedical data classification," *Journal of Biomedical Informatics,* vol. 43, pp. 15-23, 2010.

[2] L. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2002.

[3] D. M. Witten and R. Tibshirani, "A Framework for Feature Selection in Clustering," *Journal of the American Statistical Association,* vol. 105, pp. 713-726, 2012.

[4] J. A. Ting, A. D'Souza, S. Vijayakumar, and S. Schaal, "Efficient learning and feature selection in high-dimensional regression," *Neural Computation,* vol. 22, pp. 831-86, 2010.

[5] R. R. T. Castro, M. Magini, S. Pedrosa, A. R. K. Sales, and A. C. L. Nóbrega, "Principal components analysis to evaluate ventilatory variability: comparison of athletes and sedentary men," *Medical & Biological Engineering & Computing,* vol. 49, pp. 305-311, 2011.

[6] R. Chaves, J. Ramírez, J. M. Górriz, M. López, D. Salas-Gonzalez, I. Álvarez*, et al.*, "SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting," *Neuroscience Letters,* vol. 461, pp. 293-297, 2009.

[7] A. K. Seghouane, "A Kullback-Leibler divergence approach to blind image restoration," IEEE Transactions on Image Processing, vol. 20, pp. 2078-2083, 2011.

[8] X. Zhou and D. P. Tuck, "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data," *Bioinformatics,* vol. 23, pp. 1106-1114, 2007.

[9] C. T. Su and Y. H. Hsiao, "Multiclass MTS for Simultaneous Feature Selection and Classification," *IEEE Transactions Knowledge & Data Engineering,*vol. 21, pp. 192-205, 2009.

[10] L. P. Wang, Y. L. Wang, and C. Qing, "Feature selection methods for big data bioinformatics: a survey from the search perspective," *Methods,* vol. 111, pp. 21-31, 2016.

[11] Y. Liu, J. W. Bi, and Z. P. Fan, "Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms," *Expert Systems with Applications*, vol. 80, pp. 323-339, 2017.

[12] T. Zhang and Z. H. Zhou, "Large margin distribution machine," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, Ban□, Alberta, Canada, 2013, pp. 313-322.

[13] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[14] C. J. Hsieh, K. W. Chang, C. J. Lin, S. S. Keerthi, and S. Sundararajan, "A Dual Coordinate Descent Method for Large-scale Linear SVM," *Icml,* vol. 9, pp. 1369-1398, 2008.

[15] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging." *Siam Journal on Control & Optimization,* vol. 30, pp. 838-855, 1992.

[16] M. Lichman, "{UCI} Machine Learning Repository," http://archive.ics.uci.edu/ml, 2013

[17] Z. H. Zhou, "Large Margin Distribution Learning," in *Proceedings of the 6th Aritificial Neural Network in Pattern Recognition*, Montreal, Canada, 2014, pp. 1-11.

[18] G. X. Yuan, C. H. Ho, and C. J. Lin, "Recent advances of large-scale linear classification," Proceedings of the IEEE, vol. 100, pp. 2584–2603, 2012.

[19] F. Bach, "Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression," *Journal of Machine Learning Research,* vol. 15, pp. 595-627, 2014.

[20] L. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning,* vol. 46, pp. 389-422, 2002.

[21] A. Iosifidis and M. Gabbouj, "Multi-class Support Vector Machine classifiers using intrinsic and penalty graphs," *Pattern Recognition,* vol. 55, pp. 231-246, 2016.

[22] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence,* vol. 97, pp. 273-324, 1997.