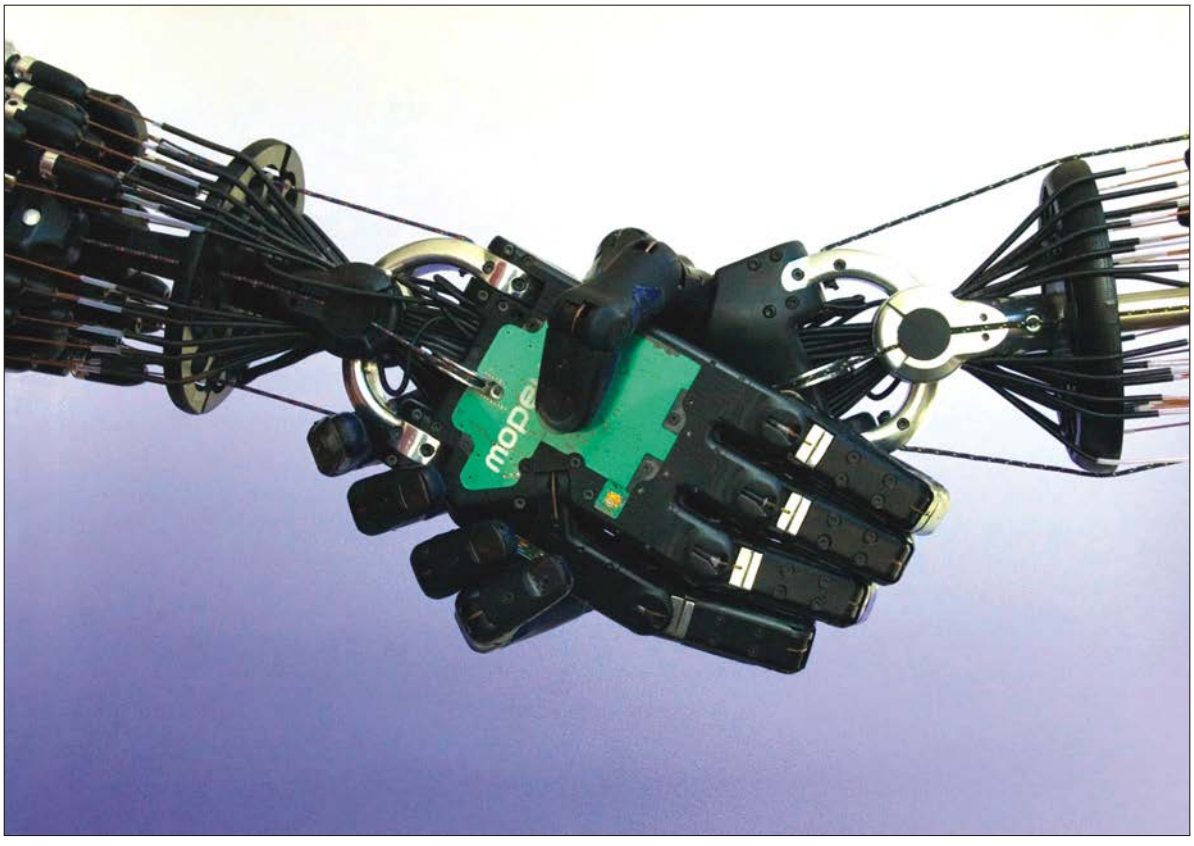

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Artificial Intelligence 8 (1): 10-16, 2015
ISSN 1994-5450 / DOI: 10.3923/jai.2015.10.16

Divide and Merge Classification for High Dimensional Multi-Class Datasets

[1]Sejong Oh and [2]Sangbum Lee
[1]Department of Nanobiomedical Science,
[2]Department of Computer Science, Dankook University, Cheonan, 330-714, South Korea

Corresponding Author: Sejong Oh, Department of Nanobiomedical Science, Dankook University, Cheonan, 330-714, South Korea

ABSTRACT

If a dataset has multiple classes and huge features like microarray data, classification accuracy may be low, even though feature selections are applied to reduce the dimensions of the dataset. Improvement of classification accuracy for the dataset is a challenging task. We propose an efficient classification method based on the "Divide-and-Merge" approach for high dimensional multi-class datasets. In the proposed method, we extracted different feature subsets for each class in an original dataset and generate new datasets. Unknown sample Si is classified into the new datasets and the results are merged for a final decision of the class label.

Key words: Classification, feature selection, multi-class dataset, bioinformatics

INTRODUCTION

Various machine learning approaches have been proposed for analyzing biological data to achieve better and more reliable results. One of the most widely used and well-known analysis methods is classification. It belongs to the supervised learning category. In classification analysis, training and testing datasets are used. The classifier learns through a training dataset where class labels are provided along with the samples (instances). Various classification algorithms have been proposed. Decision Trees (Anyanwu and Shiva, 2009), Naïve Bayes (Flach and Lachiche, 2004), Support Vector Machine (SVM) and k-Nearest Neighbor (KNN) are efficient and famous supervised algorithms. Classification techniques can be divided into two categories: Binary and multi-class (Wu *et al*., 2004). In binary classification, the samples of a given dataset are classified into two classes: For example, 'normal' and 'patient'. In contrast, samples are classified into more than two classes in multi-class classification: For example, 'low', 'high' and 'medium'. It is more complicated than binary classification due to the complexity of class separability. Feature selection is the technique that chooses the best subset of features from the dataset by removing redundant and irrelevant features. It makes the classification algorithm fast and efficient. Well-known feature selection algorithms are MRMR (Ding and Peng, 2003), ReliefF (Robnik-Sikonja and Kononenko, 2003), FSDD (Liang *et al*., 2008), FeaLect (Zare, 2010) and CBFS (Seo and Oh, 2012a).

In the previous study, we proposed a Divide-and-Merge Classification (DAMC) approach for high dimensional datasets (Seo and Oh, 2012b). It is just a reference model and can be implemented in various ways. Traditional feature selection and classification algorithms use one dataset to predict the class of unknown samples. In contrast, DAMC uses multiple datasets derived from one

original dataset. The DAMC suggests that the best features subset for each class is different. For example, for prediction class A, feature subset $\{f_1, f_4, f_5\}$ is best, whereas, $\{f_1, f_2, f_6\}$ is best for class B. The DAMC contains two steps for classification:

Step 1: Produce proper feature subsets for each class and make new datasets for classification. In this step, the original dataset is divided into n preliminary datasets, where n is the number of classes in the datasets. Then, we apply the feature selection algorithm on the preliminary datasets to find the best subset of features for each class. Finally, we take n new datasets which contains the same samples and different feature values

Step 2: Classify unknown samples into each new dataset and merge the results for a final decision. Unknown sample US_i is tested on the new datasets and n predicted results are merged for the final decision

Feature selection for Step 1 and the merge scheme for Step 2 are user-defined factors. The DAMC-CD is an application method following the DAMC approach. It is designed for high dimensional datasets which have two classes. It shows the best accuracy compared with previous methods. However, it cannot be used for multi-class datasets. In this, we propose the DAMC-MC method for high dimensional multi-class datasets. It follows the basic idea of the DAMC reference model.

METHODOLOGY

In this section, we describe the steps of DAMC-MC classification. For simplicity, we assume m-dimensional datasets which have three classes, A, B and C. The summary of DAMC-MC is as follows:

Step 1: Prepare preliminary datasets from the original dataset
- Derive 4 preliminary datasets from original dataset

Step 2: Apply feature selection and make new datasets
- Fix the number of features and derive the best subset of features for each preliminary dataset
- Make new datasets using derived feature subsets

Step 3: Calculate support degrees of each class using new datasets
- Each new dataset produces a support degree for each corresponding class

Step 4: Merge support degrees and decide the class of the unknown sample

Now we describe details of four steps.

Step 1

Prepare preliminary datasets from the original dataset: In this step, the original dataset is divided into three preliminary datasets: D_A, D_B and D_C. They correspond to classes A, B and C. They have the same samples and class label of a sample S_i €D_x is 1 where S_i belongs to class X in the original dataset. In the other cases, the class label of a sample S_i €D_x is 0. The preliminary dataset D_{Whole} is the same as the original dataset. It will be used in Step 4. The left side of Fig. 1 presents Step 1.

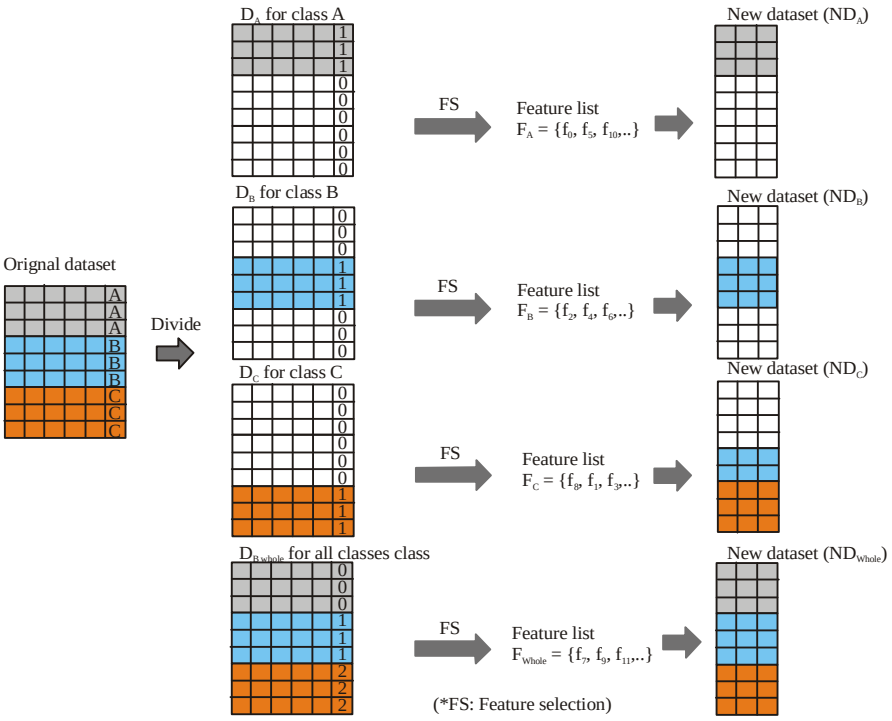

Fig. 1: Preparation of preliminary datasets and making new datasets

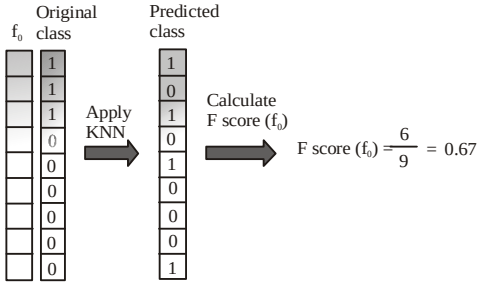

Fig. 2: Idea of Fscore

Step 2

Apply feature selection and make new datasets: Many feature selection algorithms can be used for this step. We apply a new feature selection scheme which was developed using KNN. All feature selection algorithms have evaluation functions to measure the quality of each feature in a dataset. We used Fscore as an evaluation function. We evaluated feature f_0 in preliminary dataset D_A. Figure 2 shows the idea of Fscore. To calculate Fscore, we applied KNN and found the predicted class for each feature value in f_0. We then compared the 'original class' and 'predicted class'. Fscore (f_0) is a rate of matched instances; it is calculated by Eq. 1:

$$\text{Fscore}(f_0) = \frac{\text{No. of matched instances}}{\text{No. of total instances}} \quad (1)$$

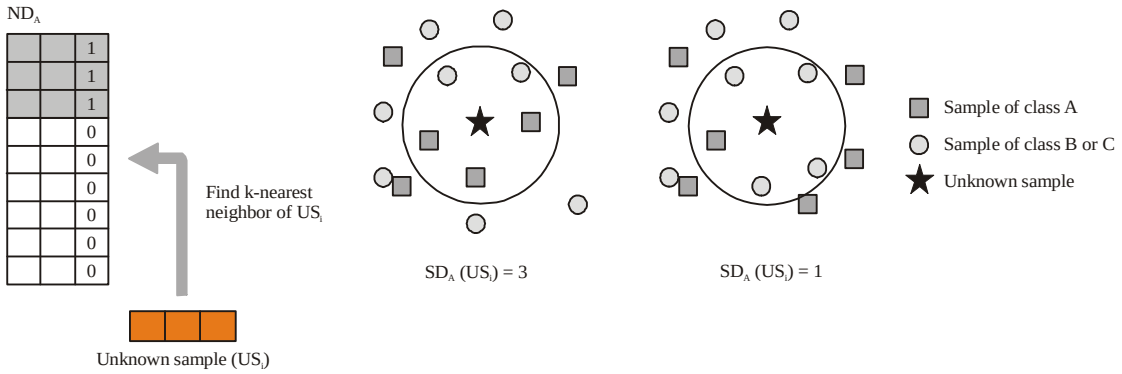

Fig. 3: Calculating support degree of class A for unknown sample USi

If Fscore is 1, it means every instance is clearly classified when we use feature f_0. In other words, f_0 has a strong classification power. After Fscore is calculated for all features, those with high Fscores are selected. When we take the best features from preliminary datasets, we can make new datasets using the selected feature subset. New datasets ND_A, ND_B, ND_C and ND_{Whole} contain the same samples (instances) but their feature values are different. The number of selected features and k of the KNN algorithm are user defined parameters. The right side of Fig. 1 presents Step 2.

Step3

Calculate support degrees of each class using new datasets: By our assumption, unknown sample US_i is predicted to be classified into one of three classes: A, B or C. Support degrees of A, B and C are calculated from different new datasets ND_A, ND_B and ND_C. If the support degree of class B is largest, then class B becomes the final predicted class of US_i. Support degrees from ND_{Whole} will be used in Step 4 only if the largest support degree from ND_A, ND_B and ND_C is not unique. Figure 3 describes how to calculate the support value of class A for unknown sample US_i, denoted by $SD_A(US_i)$. The $SD_A(US_i)$ is calculated using ND_A. We find k-nearest neighbor samples and count samples labeled as 1. Label 1 means the samples belong to class A. In the example in the right side of Fig. 3, we find 5 nearest neighbors and we take $SD_A(US_i)$ as 3 and 1. After calculating $SD_A(US_i)$, we also calculate $SD_B(US_i)$ and $SD_C(US_i)$.

Step 4

Merge support degrees and decide the class of the unknown sample: The merge process is simple. We compare $SD_A(US_i)$, $SD_B(US_i)$ and $SD_C(US_i)$. If $SD_X(US_i)$ is unique and the largest value, then X is a predicted class for unknown sample US_i. If $SD_X(US_i)$ is not unique, the predicted class is made using support values from the ND_{whole} dataset. Figure 4 shows the summary of the merge process.

RESULTS AND DISCUSSION

To evaluate the classification accuracy of DAMC-MC, we chose three well known and excellent classifications: KNN (Athitsos *et al*., 2005), SVM (Furey *et al*., 2000) and Alpha (Seo and Oh, 2013). We chose 8 microarray multi-class datasets, which are shown in Table 1. As we mentioned, feature selection is required as preprocessing for classification of a high dimensional dataset. We tested

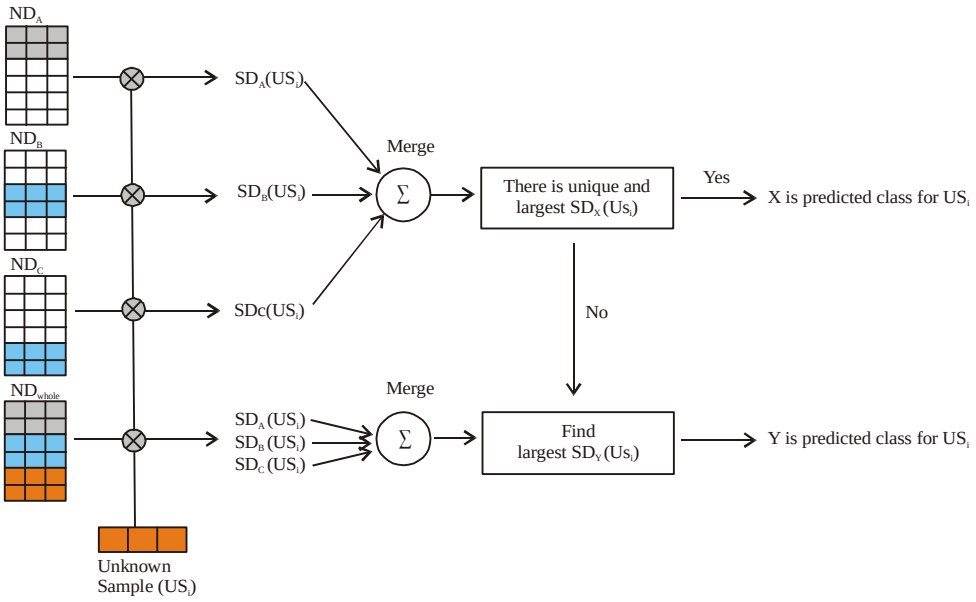

Fig. 4: Calculating support degree and merge process

Table 1: Summary of benchmark datasets

Data set name	No. of classes	No. of features	No. of instances	References
Smoke	3	11999	75	Berrar *et al.* (2009)
DLBCL	3	661	141	Hoshida *et al.* (2007)
Phoneme	5	256	4509	Hastie *et al.* (2009)
Multi tissues	4	1573	103	Hoshida (2010) and Su *et al.* (2002)
Mfeat	10	649	2000	Frank and Asuncion (2010)
Spectrometer	10	101	531	Frank and Asuncion (2010)
Lung	16	492	402	Hoshida *et al.* (2007)
Isolet	26	617	7797	Frank and Asuncion (2010)

RFS, FSDD, ReliefF and CBFS algorithms for classification algorithms, except DAMC-MC. The DAMC-MC has an embedded feature selection mechanism and does not need another feature selection algorithm. The RFS (Lee *et al.*, 2013) is based on the R-value (Oh, 2011), which is a measure used to capture the congestion area among classes in a feature. The basis of the FSDD algorithm is to identify the features that result in good class separability among classes and to ensure that samples in the same classes are as close as possible. ReliefF is regarded as one of the most successful features of selection algorithms. The basic idea of ReliefF is to iteratively estimate feature weights according to their ability to discriminate between neighboring instances. The CBFS is a very efficient feature selection algorithm based on the CScore measure. The CScore calculates the degree of samples located in the correct class region. The CBFS showed less computation time and good classification accuracy. For the SVM test, we used the LIBSVM tool (Chang and Lin, 2011) with a linear kernel. User defined values of FSDD were Beta = 3 and K = 3. In the case of ReliefF, we used K = 7. The proposed DAMC-MC uses K = 5. We used a well-known validation method, k-fold cross validation (Bengio and Grandvalet, 2004), where k = 5 to avoid the over-fitting problem of classification work. Table 2 shows the experimental parameters for each classification algorithm when they achieve the best accuracies.

Table 2: Parameters for each classifier when they get the best accuracy

Parameters	KNN	SVM	Alpha	MDMC
Smoke	RFS/15/k = 9	CBFS/20/Linear	CBFS/20	5
DLBCL	CBFS/35/ k = 5	CBFS/35/ Linear	CBFS/40	11
Multi tissues	CBFS/35/k = 5	CBFS/25/Linear	CBFS/40	18
Mfeat	RFS/5/k = 1	RFS/40/Linear	CBFS/40	25
Spectrometer	RFS/25/k = 5	RFS/25/Linear	CBFS/30	20
Lung	RFS/40/k = 1	RFS/35/Linear	CBFS/40	37
Isolet	RFS/40/k = 7	RFS/40/Linear	RFS/40	40
Phoneme	FSDD/40/k = 13	FSDD/40/Linear	CBFS/40	47

Variables for KNN include feature selection algorithm, number of selected features and number of nearest neighbor; for SVM they include feature selection algorithm, number of selected features and selected kernel; for Alpha they include feature selection algorithm and number of selected features; in DAMC-MC, the number of selected features are included

Table 3: Experimental result comparing classification accuracy

Parameters	KNN	SVM	Alpha	DAMC-MC
Smoke	0.800	0.706	0.738	0.846
DLBCL	0.933	0.872	0.896	0.955
Multi tissues	0.937	0.932	0.937	0.975
Mfeat	0.927	0.506	0.764	0.933
Spectrometer	0.880	0.847	0.637	0.881
Lung	0.856	0.850	0.711	0.894
Isolet	0.768	0.613	0.275	0.890
Phoneme	0.915	0.922	0.781	0.921

Best classification accuracy for 8 benchmark datasets

Table 3 shows the comparison result of all the classification algorithms including proposed DAMC-MC. DAMC-MC shows the best classification accuracy for 8 benchmark datasets.

CONCLUSION

We introduce an efficient classification method called DAMC-MC for high dimensional multi-class datasets. The proposed method is one of application following DAMC model. Proposed method contains both feature selection and classification scheme. In previous classification work, the relationship between feature set and classes of dataset is not considered. As a matter of fact, some feature subsets more accurately predict specific class than other classes. It is true for both binary class and multi-class dataset. The idea of KNN is used for feature evaluation and classification in DAMC-MC. If we change internal mechanism of DAMC-MC, we may expect improvement of classification accuracy. There is no superior classification method for all datasets because each dataset has special characteristics of data distribution and we do not know master principle for separating class and class. Dataset-specific classification method is valuable as a research topic. The DAMC-MC may also be modified for specific dataset. Both source program and used datasets are available at our website http://biosw.dankook.ac.kr/damc-mc.

ACKNOWLEDGMENTS

This study was supported by the National Research Foundation of Korea Grant funded by the Korean Government NRF-2012S1A2A1A01028576. The Ariundelger Gantulga, Minseok Seo gave their time and efforts to organize and establish this study.

REFERENCES

Anyanwu, M.N. and S.G. Shiva, 2009. Comparative analysis of serial decision tree classification algorithms. Int. J. Comput. Sci. Security, 3: 230-240.

Athitsos, V., J. Alon and S. Sclaroff, 2005. Efficient nearest neighbor classification using a cascade of approximate similarity measures. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 1, June 20-25, 2005, USA., pp: 486-493.

Bengio, Y. and Y. Grandvalet, 2004. No unbiased estimator of the variance of k-fold cross-validation. J. Mach. Learn. Res., 5: 1089-1105.

Berrar, D.P., W. Dubitzky and M. Granzow, 2009. A Practical Approach to Microarray Data Analysis. Springer, New York, ISBN: 9781441912268, Pages: 368.

Chang, C.C. and C.J. Lin, 2011. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol., 2: 1-39.

Ding, C. and H. Peng, 2003. Minimum redundancy feature selection from microarray gene expression data. Proceedings of the IEEE Computer Society Conference on Bioinformatics, August 2003, Stanford, CA, USA., pp: 523-528.

Flach, P.A. and N. Lachiche, 2004. Naive Bayesian classification of structured data. Mach. Learn., 57: 233-269.

Frank, A. and A. Asuncion, 2010. UCI machine learning repository Irvine. University of California, School of Information and Computer Science, USA. http://archive.ics.uci.edu/ml/.

Furey, T., N. Cristianini, N. Duffy, D. Bednarski, M. Schummer and D. Haussler, 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16: 906-914.

Hastie, T., R. Tibshirani and J. Friedman, 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer Science and Business Media, New York, ISBN: 9780387848587, Pages: 767.

Hoshida, Y., J.P. Brunet, P. Tamayo, T.R. Golub and J.P. Mesirov, 2007. Subclass mapping: Identifying common subtypes in independent disease data sets. PloS One, Vol. 2. 10.1371/journal.pone.0001195.g005

Hoshida, Y., 2010. Nearest template prediction: A single-sample-based flexible class prediction with confidence assessment. PloS One, Vol. 5. 10.1371/journal.pone.0015543.g006

Lee, J., N. Batnyam and S. Oh, 2013. RFS: Efficient feature selection method based on R-value. Comput. Biol. Med., 43: 91-99.

Liang, J., S. Yang and A. Winstanley, 2008. Invariant optimal feature selection: A distance discriminant and feature ranking based solution. Pattern Recognition, 41: 1429-1439.

Oh, S., 2011. A new dataset evaluation method based on category overlap. Comput. Biol. Med., 41: 115-122.

Robnik-Sikonja, M. and I. Kononenko, 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learn., 53: 23-69.

Seo, M. and S. Oh, 2012a. CBFS: High performance feature selection algorithm based on feature clearness. PloS One, Vol. 7. 10.1371/journal.pone.0040419

Seo, M. and S. Oh, 2012b. Derivation of an artificial gene to improve classification accuracy upon gene selection. Comput. Biol. Chem., 36: 1-12.

Seo, M. and S. Oh, 2013. A novel divide-and-merge classification for high dimensional datasets. Comput. Biol. Chem., 42: 23-34.

Su, A.I., M.P. Cooke, K.A. Ching, Y. Hakak and J.R. Walker *et al*., 2002. Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci., 99: 4465-4470.

Wu, T.F., C.J. Lin and R.C. Weng, 2004. Probability estimates for multi-class classification by pairwise coupling. J. Mach. Learn. Res., 5: 975-1005.

Zare, H., 2010. FeaLect: Feature selection by computing statistical scores. The R Journal, 2010. http://cran.rakanu.com