Schedae Informaticae Vol. 24 (2015): 53–62
doi: 10.4467/20838476SI.15.005.3027

Feature Selection and Classification Pairwise Combinations for High-dimensional Tumour Biomedical Datasets

Agnieszka Wosiak, Agata Dziomdziora
Institute of Information Technology
Łódź University of Technology
ul. Wólczańska 215, 90-924 Łódź, Poland
e-mail: *agnieszka.wosiak@p.lodz.pl*

Abstract. This paper concerns classification of high-dimensional yet small sample size biomedical data and feature selection aimed at reducing dimensionality of the microarray data. The research presents a comparison of pairwise combinations of six classification strategies, including decision trees, logistic model trees, Bayes network, Naïve Bayes, k-nearest neighbours and sequential minimal optimization algorithm for training support vector machines, as well as seven attribute selection methods: Correlation-based Feature Selection, chi-squared, information gain, gain ratio, symmetrical uncertainty, ReliefF and SVM-RFE (Support Vector Machine-Recursive Feature Elimination). In this paper, SVM-RFE feature selection technique combined with SMO classifier has demonstrated its potential ability to accurately and efficiently classify both binary and multi-class high-dimensional sets of tumour specimens.

Keywords: feature selection, classification, high-dimensional tumour biomedical datasets.

1. Introduction

Feature selection and classification methods have become of particular interest in the field of bioinformatics as the high-dimensional nature of biomedical data makes the process of characterizing samples a challenging task. Nowadays, biomedical data are usually high-dimensional, frequently containing thousands of features yet much fewer samples. There is no denying that more information gives a better chance for knowledge discovery. Due to the curse of dimensionality, researchers face serious issues connected with the fact that the number of genes (features) far exceeds the size of the dataset examples. Thus, dimensionality reduction appears to be crucial for the effective classification of tumour samples. This can be achieved by applying either feature extraction or feature selection methods before construction of the classification model. When compared to feature extraction, gene selection does not alter the original representation of genes. Hence, it may not only enhance the performance of tumour classification by removal of both redundant and irrelevant genes, but also determine the informative gene subsets, capable of serving as tumour biomarkers and potential drug targets. As already mentioned, details concerning tumour development are still obscure. In order to potentially address this problem, feature selection seems to be able to provide deep insight into the underlying molecular mechanism of tumour development.

The aim of this paper is to create a comparison of pairwise combinations of feature selection methods and classification techniques applied to the problem of binary and multi-class cancer classification. Even though both binary and multi-class sample classification have been studied extensively over the past few years [14, 15, 17], no exact solution has been discovered. Nowadays, still there is no perfect combination of feature selection and classification methods as applied to high-dimensional yet small sample size microarray data. This research does not only constitute an independent contribution to the relevant literature, but also strive for finding a successful way to perform efficient feature selection enhancing accurate classification of tumour specimens.

The rest of the paper is organized as follows. In Section 2 we describe the methodology of our research. Section 3 is dedicated to the experiments conducted on sample data and the results. Finally, in Section 4, the concluding remarks are discussed.

2. Feature selection and classification methods in terms of cancer diagnosis

High-throughput technologies used in the fields of genomics and proteomics provide the opportunity to examine a large number of biological samples simultaneously. This leads to high amounts of multivariate data corresponding to different biological aspects such as various variants of a disease, different stages of a disease, as well as different survival rates and responses to treatment agents. Having a possibility to look at the expression levels of thousands of genes and proteins is, of course, beneficial, yet only to some extent. The problem occurs when there are only few samples available, increasing the risk of overfitting the data, especially leading to unsatisfactory classification in terms of new data points. What is more, using a classifier that is too complex for the quantity of available data can only promote overfitting. In particular, high complexity can be induced by too many features present in the examined data. Therefore, feature selection is exceptionally crucial in the case of biological research where sample size is usually limited [2, 15].

The proposed methodology of choosing best pairwise combinations of feature selection and classification methods consists of four steps:

- data pre-processing, which results in the initial dataset;
- feature selection, which enables the choice of the set of attributes crucial for the automated diagnosis;
- classification process based on the attributes derived from the previous step;
- verification by assessing appropriate comparison criteria.

Data pre-processing includes two main steps: firstly excluding housekeeping genes and then normalization. In general, housekeeping genes take part in basic cell maintenance, and hence are supposed to maintain constant expression levels in all cells and conditions [3]. This indicates that housekeeping genes may provide serious redundancy and noise into the classification once these are chosen by attribute selection methods. Therefore, these genes are to be removed from the datasets before applying feature selection techniques. Identification of these genes facilitates exposure of the underlying cellular infrastructure and increases understanding of functional characteristics, evolutionary as well as various structural genomic and epigenetic features [1, 3]. However Affymetrix (American manufacturer of DNA microarrays) housekeeping genes IDs are marked in datasets by the prefix 'AFFX-' and therefore can be identified automatically. Tissue specific (TS) genes are those genes which are only or chiefly expressed in a certain tissue or an organ, and thus in charge of particular functions and development. Tissue specific genes provide relevant information in terms of classification so that these are to be maintained and subjected to experiments. Besides, the values in the datasets are normalized (standardized) so that every gene expression value is characterized by mean of zero and unit variance.

The main purpose of feature selection is to extract the smallest feature subset using a defined generalization error, or finding the best feature subset with k features which provides the minimum generalization error. There is a number of additional benefits connected with feature selection. Firstly, it improves the generalization performance concerning the model created using the entire set of features. Secondly, it offers a substantially more robust generalization and a faster response with test data. Moreover, feature selection enables researchers to gain a deeper insight into the underlying processes that generated the data [9].

In this paper, seven feature selection methods are used: Correlation-based Feature Selection (CFS), Chi-squared, Information Gain, Gain Ratio, Symmetrical Uncertainty, ReliefF and SVM-RFE. All of these feature selection methods except for SVM-RFE (support vector machine method and recursive feature elimination introduced by Guyon et al. [4]) belong to filter algorithms. They are well-suited for high-dimensional datasets, in terms of accuracy, time as well as memory efficiency [11], [13].

For the classification purpose we will consider six approaches: J48, logistic model trees (LMT), Bayes network (BayesNet), Naïve Bayes (NaiveBayes), k-nearest neighbours (IBk) and sequential minimal optimization algorithm for training support vector machines (SMO). These classifiers are used in order to compare feature selection methods discussed in this paper.

Table 1. Datasets description.

Dataset name	No. of samples	Initial no. of features	No. of features after pre-processing	No. of classes
ALL/AML	72	7129	7070	2
CNS	60	7129	7070	2
Colon	62	2000	1988	2
Lung	181	12600	12533	2
Lymphoma	96	4026	4026	11
GCM	192	16063	16004	14

In order to assess the performance of various pairwise combinations of feature selection and classification methods, following comparison criteria have been used: accuracy, sensitivity, specificity, FP rate, precision, root mean square error as well as the number of features used while performing a given test.

3. Experimental analysis and results

There were six different either binary or multi-class cancer microarray gene expression datasets used in the research: Colon Cancer Dataset (binary), Lung Cancer Dataset (binary), ALL/AML Dataset (binary), Lymphoma Dataset (multi-class), GCM Dataset (multi-class) and CNS Dataset (binary). The summary of all the sets of biomedical data is given in Table 1. Additionally, Table 1 includes the number of genes before and after the pre-processing stage (i.e. removal of housekeeping genes).

The final results of pairwise combinations of feature selection techniques and classification methods are presented in Table 3. The best classification results without conducting any feature selection method are shown in Table 2.

In the case of classifications conducted without the usage of any feature selection method, SMO significantly outperformed other classifiers in terms of classification accuracy. In the case of both binary leukaemia and multi-class lymphoma datasets, SMO achieves up to the 100% accuracy. This is likely to be an optimistically biased classification, chiefly due to the application of the same data for both model development and model validation. A considerably more promising result was obtained in the case of colon cancer dataset (94% accuracy) and CNS dataset (95% accuracy). There is no denying that the multi-class classification provided the worst results, attaining 67% accuracy for SMO classifier.

As far as the results pertaining to the InfoGain-filtered Correlation-based Feature Selection method are concerned, again too optimistic, biased classification accuracies were observed (100% classification accuracy). The most statistically meaningful results were obtained in the case of CNS cancer dataset (98% accuracy using SMO model), with the reduced number of features from 7070 to only 38 genes. In the case of multi-class GCM dataset, the most promising results (81% of instances classified correctly) were attained for the combination of SMO classifier and InfoGain-filtered

Table 2. Best classification results without feature selection.

Dataset	Classif. method	No. of features	Comparison criteria	

ALL/AML	SMO	7070	ACC = 100.000
			SPEC = 1.000
			RMSE = 0.000
				SENS = 1.000
				FP rate = 0.000

CNS	SMO	7070	ACC = 95.000
			SPEC = 0.929
			RMSE = 0.224
				SENS = 0.950
				FP rate = 0.071

Colon	SMO	1988	ACC = 93.548
			SPEC = 0.924
			RMSE = 0.254
				SENS = 0.935
				FP rate = 0.076

Lung	LMT	12533	ACC = 96.059
			SPEC = 0.945
			RMSE = 0.121
				SENS = 0.961
				FP rate = 0.055

Lymphoma	SMO	4026	ACC = 94.792
			SPEC = 0.987
			RMSE = 0.266
				SENS = 0.948
				FP rate = 0.013

GCM	SMO	16004	ACC = 67.361
			SPEC = 0.981
			RMSE = 0.245
				SENS = 0.674
				FP rate = 0.019

CFS method. Additionally, SMO classifier appeared to provide the best average classification accuracy when combined with CFS. In the case of various combinations of classifiers and chi-squared filter attribute selection method, SMO appeared to provide the best classification accuracies amongst all the other pairs of solutions. Besides in general too optimistic, biased accuracies obtained for ALL/AML dataset, the best results were observed in the case of lung cancer binary set of data (97% accuracy).

The classification accuracies obtained in the case of pairwise combinations of classification models and InfoGain feature selection method (applied individually, not as a pre-processing step) appeared to be worse than in the case of a hybrid approach (InfoGain/CFS feature selection). The results occurred to be statistically worse particularly in the case of GCM dataset consisting of multiple classes of cancer. Unfortunately, nearly 60% classification accuracy cannot be regarded as a meaningful and satisfactory outcome (in the case of InfoGain/CFS-SMO hybrid method as much as 81% was observed). The GainRatio feature selection method appeared to provide the most satisfactory classification accuracies when combined with Sequential Minimal Optimization (SMO) algorithm.

Symmetrical uncertainty-based filter method provided the best classification outcomes when combined with SMO classifier. The GCM multi-class cancer dataset attained up to 58% classification accuracy with respect to LMT classifier. The CNS dataset was classified satisfactorily using SMO technique.

ReliefF feature selection method provided satisfactory results for the majority of both binary and multi-class tumour sets of data. In the case of GCM dataset, again the classification accuracy appeared to be worse than expected (up to 58% cancer instances classified correctly using LMT classifier). The most reliable results were obtained for the lung cancer dataset (99% classification accuracy using Naïve

Table 3. Best classification results for datasets with feature selection.

Dataset	Classif.	FS method	No. of features	Comparison criteria	
ALL/AML	CFS	SMO	34	ACC = 100.000 SPEC = 1.000 RMSE = 0.000	SENS = 1.000 FP rate = 0.000
ALL/AML	Chi-squared	SMO	150	ACC = 100.000 SPEC = 1.000 RMSE = 0.000	SENS = 1.000 FP rate = 0.000
CNS	InfoGain	SMO	150	ACC = 95.000 SPEC = 0.929 RMSE = 0.224	SENS = 0.950 FP rate = 0.071
CNS	SymmUncer	SMO	150	ACC = 95.000 SPEC = 0.929 RMSE = 0.224	SENS = 0.950 FP rate = 0.071
Colon	SVM-RFE	SMO	150	ACC = 95.161 SPEC = 0.932 RMSE = 0.220	SENS = 0.952 FP rate = 0.068
Lung	Chi-squared	SMO	150	ACC = 97.044 SPEC = 0.946 RMSE = 0.318	SENS = 0.970 FP rate = 0.054
Lung	ReliefF	Naive Bayes	150	ACC = 98.551 SPEC = 0.967 RMSE = 0.076	SENS = 0.986 FP rate = 0.033
Lymphoma	CFS	Naive Bayes	152	ACC = 100.000 SPEC = 1.000 RMSE = 0.000	SENS = 1.000 FP rate = 0.000
Lymphoma	Chi-squared	Naive Bayes	150	ACC = 93.939 SPEC = 0.981 RMSE = 0.062	SENS = 0.939 FP rate = 0.019
Lymphoma	GainRatio	LMT	150	ACC = 78.125 SPEC = 0.950 RMSE = 0.183	SENS = 0.781 FP rate = 0.050
GCM	CFS	SMO	42	ACC= 81.250 SPEC = 0.989 RMSE = 0.243	SENS = 0.813 FP rate = 0.011
GCM	InfoGain	SMO	150	ACC= 59.722 SPEC = 0.976 RMSE = 0.246	SENS = 0.597 FP rate = 0.024
GCM	SVM-RFE	SMO	150	ACC = 73.611 SPEC = 0.984 RMSE = 0.244	SENS = 0.736 FP rate = 0.016

Table 4. Comparison of no. of features and accuracy with and without FS.

Dataset	No. of features without FS	No. of features after FS	Features reduction [%]	ACC without FS	ACC after FS	ACC diff [%]
ALL/AML	7070	34	99.52	100.00	100.00	0.00
CNS	7070	150	97.88	95.00	95.00	0.00
Colon	1988	150	92.45	93.55	95.16	+1.61
Lung	12533	150	98.80	96.06	98.55	+0.967
Lymphoma	4026	150	96.27	94.79	93.94	−0.85
GCM	16004	42	99.74	67.36	81.25	+13.89

Bayes classifier), while the most optimistic ones (100% classification accuracy) for the lymphoma and ALL/AML datasets. ReliefF-SMO combination appeared to be successful with respect to the colon cancer dataset (89% accuracy).

The combination of Information Gain/SVM-RFE/SMO hybrid classifier can be regarded as the most accurate one, providing up to 74% accuracy on the most problematic multi-class GCM dataset. As far as the remaining sets of data are concerned, over 90% classification accuracy was obtained in each of the considered cases. Besides, taking into consideration the lymphoma dataset, Naïve Bayes classifier provided the most optimistic results (100% classification accuracy), suggesting that the application of the same data for model development and model validation results in the undesirable classification bias.

The comparison of numbers of features and accuracies for all datasets with and without feature selection methods is presented in Table 4. It is noticeable that by performing feature selection on high-dimensional tumour datasets we can significantly reduce the attribute space with slight loss of accuracy. The best pairwise combinations produced the reduction of features equalled more than 92% and the accuracy decrease less than 1%. In the cases of colon, lung and GCM datasets the best classification results were even better after feature selection applied than before for the whole dataset, which may be observed in the cases, where the search for the best feature set is still an active research topic [5].

4. Conclusions and future work

Classification of high-dimensional biomedical datasets is regarded as a challenging task, requiring extremely high accuracy and as short computational time as possible. There is no denying that the enormous dimensionality of the microarray expression data is a serious concern during gene selection. All of the already reported results concerning attribute selection methods in terms of microarray data suggest that multi-class classification issues are typically more difficult than the binary ones. Therefore research on finding the most appropriate methods for a multi-class classification are conducted and often succeed in new approaches like in [10] by Podolak et al. Moreover classification datasets often have an unequal class distribution among their examples. This problem is known as imbalanced classification. To overcome the problems produced by noisy and borderline examples in imbalanced datasets, new methods are introduced in Sáez et al. [12].

By comparing all the possible pairwise combinations of classification algorithms and feature selection methods, it was demonstrated that the hybrid strategy resulted in the most satisfactory outcomes and confirmed other up-to-date researches on multiple classifier systems led by Woźniak et al. [16] and Kumar et al. [6]. In order to specifically tailor the hybrid approach so that the high classification accuracy is to be obtained regardless of the set of input data, one has to take into account a variety of aspects. Moreover, the question remains whether it is possible to discover the optimal number of genes to be selected by ranking methods. These, and all the other circumstances, contribute to the difficulties related to finding the optimal and universal feature selection and classification method, specifically tailored to handle biomedical datasets.

It was successfully proved that the SMO classifier outperforms other classification methods in the majority of cases, regardless of the input dataset used for the purpose of training the model. As far as the attribute selection is concerned, the SVM-RFE approach appeared to be perfectly suited for classification using SMO method. In order to reduce the computational complexity of the classifier, each of the biomedical datasets used in this paper were pre-processed by removing the housekeeping genes, normalization and more importantly, by using the information gain-based filter, significantly reducing the number of genes subjected to the classification task. Basically, the genes filtered using information gain feature selector proved to be considerably more informative for the SMO classifier. SVM-RFE algorithm was already demonstrated to be the most reliable and efficient feature selection method when compared to others, as well as the SVM-RFE algorithm combined with SMO classification was considered as the most beneficial choice for constructing the learning model in the studies of both Li et al. [7, 8].

Future studies ought to involve other algorithms and strategies as well, especially those ones specifically tailored to deal with the most challenging multi-class cancer classification tasks. In order to find the optimal solutions, other combinations of various classifiers and attribute selectors should be investigated in depth.

5. References

[1] Chang C.-W., Cheng W.-C., Chen C.-R., Shu W.-Y., Tsai M.-L., et al., *Identification of Human Housekeeping Genes and Tissue-Selective Genes by Microarray Meta-Analysis.* PLoS ONE, 2011, 6(7): e22859, doi:10.1371/journal.pone.0022859.

[2] Dougherty E.R., Hua J., Sima C., *Performance of Feature Selection Methods.* Curr. Genomics. 2009, 10, pp. 365–374.

[3] Eisenberg E., Levanon E.Y., *Human housekeeping genes, revisited.* Trends in Genetics, October 2013, 29(10), pp. 569–574, doi:10.1016/j.tig.2013.05.010.

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

[5] Janecek A., Gansterer W., Demel W., Ecker G., *On the relationship between feature selection and classification accuracy.* Journal of Machine Learning and Research, 2008, 4, pp. 90–105.

[6] Kumar A.P., Valsala P., *Feature Selection for high Dimensional DNA Microarray data using hybrid approaches.* Bioinformation, 2013, 9(16), pp. 824–828.

[7] Li X., Lu H., Wang M., *A Hybrid Gene Selection Method for Multi-category Tumor Classification using Microarray Data.* Int. J. Bioautomation, 2013, 17(4), pp. 249–258.

[8] Li X., Peng S., Zhan X., Zhang J., Xu Y., *Comparison of feature selection methods for multiclass cancer classification based on microarray data.* Proceedings of the 4th International Conference on Biomedical Engineering and Informatics (BMEI), 2011, 3, pp. 1692–1696.

[9] Liu G., Kong L., Gopalakrishnan V., *A Partitioning Based Adaptive Method for Robust Removal of Irrelevant Features from High-dimensional Biomedical Datasets.* AMIA Summits on Translational Science Proceedings, 2012, pp. 52–61.

[10] Podolak I. T., Roman A., *CORES: fusion of supervised and unsupervised training methods for a multi-class classification problem.* Pattern Analysis and Applications, 2011, 14, pp. 395–413.

[11] Saeys Y., Inaki I., Larranaga P., *A review of feature selection techniques in bioinformatics.* Bioinformatics, 2007, 23(19), pp. 2507–2517.

[12] Sáez J.A., Luengo J., Stefanowski J., Herrera F., *SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a resampling method with filtering.* Information Sciences, 10 January 2015, 291, pp. 184–203, http://dx.doi.org/10.1016/j.ins.2014.08.051.

[13] Trevino V., Falciani F., Barrera-Saldana H.A., *DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research.* Molecular Medicine, 2007, 13(9–10), pp. 527–541.

[14] Wang X., Gotoh O., *A Robust Gene Selection Method for Microarray-based Cancer Classification.* Cancer Informatics, 2010, 9, pp. 15–30.

[15] Wang Y., Tetko I.V., Hall M.A., Frank E., Facius A., Mayer K.F., *Gene selection from microarray data for cancer classification–a machine learning approach.* Comput. Biol. Chem., 2005, 29, pp. 37–46.

[16] Woźniak M., Graa M., Corchado E., *A survey of multiple classifier systems as hybrid systems.* Information Fusion, 2014, 16, pp. 3–17.

[17] Zhang H., Wang H., Dai Z., Chen M.S., Yuan Z., *Improving accuracy for cancer classification with a new algorithm for genes selection.* BMC Bioinformatics, 2012, 13 (298), pp. 1.