Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 23 (2013) 5 – 14

Procedia
Computer Science

4th International Conference on Computational Systems-Biology and Bioinformatics, CSBio2013

A comparative study of feature selection and classification methods for gene expression data of glioma

Heba Abusamra*

Computer, Electrical and Mathematical Sciencies and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

Abstract

Microarray gene expression data gained great importance in recent years due to its role in disease diagnoses and prognoses which help to choose the appropriate treatment plan for patients. This technology has shifted a new era in molecular classification. Interpreting gene expression data remains a difficult problem and an active research area due to their native nature of “high dimensional low sample size”. Such problems pose great challenges to existing classification methods. Thus, effective feature selection techniques are often needed in this case to aid to correctly classify different tumor types and consequently lead to a better understanding of genetic signatures as well as improve treatment strategies. This paper aims on a comparative study of state-of-the-art feature selection methods, classification methods, and the combination of them, based on gene expression data. We compared the efficiency of three different classification methods including: support vector machines, k-nearest neighbor and random forest, and eight different feature selection methods, including: information gain, twoing rule, sum minority, max minority, gini index, sum of variances, t-statistics, and one-dimension support vector machine. Five-fold cross validation was used to evaluate the classification performance. Two publicly available gene expression data sets of glioma were used in the experiments. Results revealed the important role of feature selection in classifying gene expression data. By performing feature selection, the classification accuracy can be significantly boosted by using a small number of genes. The relationship of features selected in different feature selection methods is investigated and the most frequent features selected in each fold among all methods for both datasets are evaluated.

Keywords: gene expression; microarray data; feature selection; classification; glioma

* Corresponding author.
E-mail address: heba.abusamra@kaust.edu.sa

1877-0509 © 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of the Program Committee of CSBio2013
doi:10.1016/j.procs.2013.10.003

1. Introduction

Cancer diagnosis nowadays is based on clinical evaluation and physical examination and also refers to medical history. But this diagnosis takes a long time. It might be too late to cure the patient if a tumor is found in its critical stage. Also, it is very important for diagnostic research to develop diagnostic procedures based on inexpensive microarray gene expression data that have adequate number of genes to detect diseases. The classification of gene expression data is challenging due to the enormous number of genes relative to the number of samples. It is common that a large number of genes are not informative for classification because they are either irrelevant or redundant. For that, it is significant to reduce the number of genes in order to get a good accuracy for the classification task. In machine learning community, there are two main approaches to achieve this goal, i.e., dimension reduction and feature selection. The former refers to methods that create new features as the combinations of the original features, such as non-negative matrix factorization[1-3]. The latter refers to methods that select the most relevant features from the original ones[4]. Filter approach and wrapper approach are widely used for feature selection[5]. Filter methods are the ones that select features as a pre-processing step. That is, they select features without considering the classification accuracy. On the other hand, the wrapper methods use the predictive accuracy of an algorithm to evaluate the possible subset of features and select the subset of features that provides the highest accuracy. Although wrapper methods usually have higher accuracy than filter methods, they are much more computationally expensive and have a higher risk of over-fitting than filter methods. In problems with a large number of features, such as the gene expression problems, wrapper methods are often infeasible and filter methods are often adopted due to their computational efficiency[6].

Many machine learning methods have been introduced into microarray classification to attempt to learn the gene expression data pattern that can distinguish between different classes of samples in recent years. The work done by Pirooznia *et al.*[7] evaluated and compared the efficiency of different classification methods, including support vector machine (SVM), neural network, Bayesian classification, decision tree (J84, ID3) and random forest methods. Also, a number of clustering methods including K-means, density-based clustering, and expectation maximization clustering were applied to eight different binary (two class) microarray datasets. Further, the efficiency of the feature selection methods including support vector machine recursive feature elimination (SVM-RFE), Chi-squared, and correlation based feature selection were compared. Ten-fold cross validation was used to calculate the accuracy of the classifiers. First the classification methods were applied to all datasets without performing any feature selection. In most datasets SVM and neural networks performed better than other classification methods. Then the effect of feature selection methods was examined on the different classification methods. Various number of genes were tested (500, 200, 100, and 50) and the top 50 genes were selected because it gave a good accuracy, consumed less processing time, and required less memory configurations comparing to others. Almost in all cases, the accuracy performance of classifiers was improved after applying feature selections methods to the datasets. In all cases SVM-RFE performed very well when it was applied with SVM classification methods. Liu, Li and Wong[8] presented a comparative study of five feature selection methods using two datasets (Leukemia and Ovarian cancer). The feature selection methods are: entropy-based, Chi-squared, t-statistics, correlation-based, and signal-to-noise statistic. The top 20 genes that have the highest score in chi-squared, t-statistics, and signal-to-noise were selected, and all the features recommended by correlation-based were selected. For entropy, features having an entropy value less than 0.1 were selected if existed, or the 20 features with the lowest entropy values were selected otherwise. The effectiveness of these features was evaluated using k-nearest neighbour (KNN), C4.5, naïve Bayes, and SVM classifiers. SVM reported the least error rate among the other classification methods when applied to the datasets without feature selection. When applying feature selection on the datasets, the accuracy performance of the four classifiers was greatly improved in most cases. The work by Li *et al.*[9] studied and compared the results of multiclass classification using many feature selection and classification methods on nine multiclass gene expression datasets. The "RankGene" software[10] was used to select the informative and related genes on the training set with eight methods supported in this software: information gain, twoing rule, sum minority, max minority, Gini index, sum of variances, one-dimensional SVM, and t-statistics. The top 150 ranked genes in every dataset were selected. The multiclass classifiers that were used to evaluate the selected genes were: SVM, KNN, naive Bayes and decision tree.

In the experiments the original partition of the datasets into training and test sets was used whenever information about the data split was available. Otherwise four-fold cross validation was applied. They concluded that the SVM had the best performance in all the datasets. The KNN classifier gave reasonably good performance on most of the datasets which means it is not problem-dependent. Other interesting discussions of their report were that it was difficult to choose the best feature selection method, and the way that feature selection and classification methods interacted seemed very complicated. Due to the separation of the gene selection part from the classification part, there is no learning mechanism to learn how those component interact with each other.

Although a number of comparative studies have been done on feature selection and classification methods on gene expression data, they were all conducted on different gene expression data sets. Furthermore, there was no detailed analysis of the genes identified from those studies and how to possibly combine the strength of them. The main contributions of this paper are based on three folds. First, to our knowledge, this is the first comparative study of machine learning methods on gene expression data of glioma. Second, we found out two gene expression data sets that shared exactly the same genes for the same disease. Thus we can study the stability of different methods. Third, instead of selecting genes only, we analyzed those genes and proposed a way to combine genes found from different data sets together to get a more accurate and stable set of features.

2. Method

The methodology is simply divided into two phases: (i) the feature (genes) selection that will be used for training and testing and (ii) evaluation of the effectiveness of feature selection methods using different classifiers. Eight different feature selection methods are considered in our study: information gain, twoing rule, sum minority, max minority, gini index, sum of variances, one-dimensional SVM, and t-statistics. The former six methods have been widely used either in machine learning (e.g., information gain and gini index) or in statistical learning theory (e.g., twoing rule, max minority, and sum of variances). The latter two determine the effectiveness of a feature by evaluating the strength of class predictability when the prediction is made by splitting the full range of expression of a given gene into two regions, the high region and the low region. The split point is chosen to optimize the corresponding measure. Feature selection was performed on the training set only. Several number of genes were tested using the eight feature selection methods (20, 50, 100, 150, 200, and 250) and top 20 genes that have the highest score were selected because it performed well, consumed less time, and required less memory configurations comparing to others.

After selecting the top 20 genes, we applied three different classification methods SVM, KNN, and random forest, and obtained the accuracy on the test set. The classification results are then used to compare the effectiveness of various feature selection methods.

SVMs: the ability of SVMs to deal with high dimensional data such as gene expression made this method the first choice for classification. LIBSVM package for SVM implementation in MATLAB is used in the experiments. The four different kernels (linear, polynomial, radial, and sigmoid) are tested. In each case two parameters (C, γ) are changed to different values, and the pair of (C, γ) with the best cross-validation accuracy is picked. Trying exponentially growing sequences of C is a practical method to identify good parameters. Here, we tried the values of parameters as $C = 2^5, 2^4, \ldots, 2^{-6}$ and $\gamma = 0.0001, \gamma \times 2, \ldots, 1$.

KNN: The simplicity is its main advantage. Euclidian distance is used as the distance metric, and to break the tie, K is set to be an odd number from 1 to 11, and the K that reports the best cross validation accuracy is selected.

Random forest (RF): gains many advantages of decision trees as well as achieves better results through the usage of bagging on samples, random subsets of variables, and a majority voting scheme. The random forest package by Leo Breiman and Adele Cutler is used in the experiments. Each experiment is repeated ten times then the average accuracy, sensitivity, and specificity of the ten trails is reported.

3. Results

3.1. Data Sets

Two publicly available gene expression data sets of glioma have been used in the experiments. The datasets were labeled according to the names of the authors of the respective studies: 'Freije'[11] and 'Phillips'[12]. Both data sets are from Gene Expression Omnibus (GEO)[13], and were originally generated using the affymetrix arrays (U133A platform). The Freije data set contains 85 expression profiles for seven patients, including a total of 11 repeat samples. Here, we excluded the repeated samples to end up with 74 samples of grade III (24 samples) and grade IV (50 samples) with 12,287 features or genes. The Phillips data set shares the same genes as the Freije data set, but with 98 samples of grade III (23 samples) and grade IV (75 samples).

Before starting the experiments, data pre-processing is an important step due to the noisy nature of the data generated by microarray technology. Both data sets have to be normalized to decrease the expression measurements variation. Min-max normalization is used to normalize the data sets. The gene expression values are scaled such that the smallest value for each gene becomes zero and the largest value becomes one.

3.2. Experimental Results

The classification methods were first applied to both datasets without performing any feature selection. Results of the 5-fold cross validation are shown in Fig. 1. In both data sets, SVM performed better than other classification methods due to its suitability for high dimensional data. SVM classification had the best accuracy in the Freije data set with an accuracy of 91.89% and in the Phillips data set with an accuracy of 86.73%. For the Freije data, linear, radial, and sigmoid kernels of SVM gave a higher accuracy than polynomial, while linear gave lower accuracy than other kernels in the Phillips. KNN achieved the best performance when K=3, i.e., an accuracy of 90.54% and 84.69% for the Freije and Phillips data sets, respectively. Random forest reports the minimum accuracy 87.88% in the Freije data, and in the Phillips data set has almost the same accuracy as 3-NN but with worse specificity 49.13%.

The next experiment, we applied the classification methods to both datasets after selecting the top twenty genes of each feature selection methods.

Results on the Freije data set

The performance of SVM and 3-NN vary in different feature selection methods. With only twenty features selected by information gain, gini index, sum of minority methods, or 1-dimentional SVM, SVM maintained the same accuracy 91.89% as that by using all features. Using sum of variances and max minority achieved the minimum accuracy 89.18%. Gini index and 1-dimentional SVM achieved the best accuracy in 3-NN 90.54%. For random forest, the accuracy performance was improved after applying feature selection mostly in all cases. The highest accuracy is achieved 91.22% by sum of variances, while the minimum accuracy 87.57% was the same as that by using all features. Fig. 2 summarizes the performance of feature selection and classification methods.

Results on the Phillips data set

For the Phillips data set, almost in all cases, the accuracy performance of SVM was improved after applying feature selections. With only twenty features, SVM had the highest accuracy 88.77% with t-statistics and gini index. Almost for the other feature selection methods it maintained the same accuracy as that by using all features. The performance of random forest had not improved much after performing feature selection. It maintained the same accuracy 84.14% as using all features with 1-dimetional SVM and sum of minority, while the accuracy decreased by maximum 2% with other methods. The same situation happened for 3-NN except it achieved the minimum accuracy 76.53% with the information gain method, 8% less than the best accuracy 84.28% achieved by using t-statistics. Fig. 3 summarizes the performance of classification methods after performing feature selection on the Phillips data set.

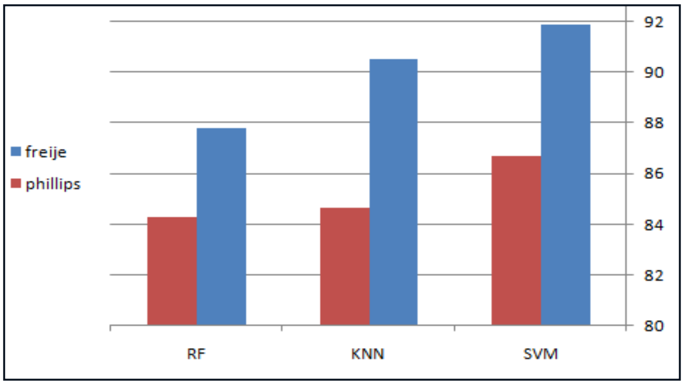


Fig. 1. Accuracy of the classification methods on the 5-fold cross validation without performing feature selection.

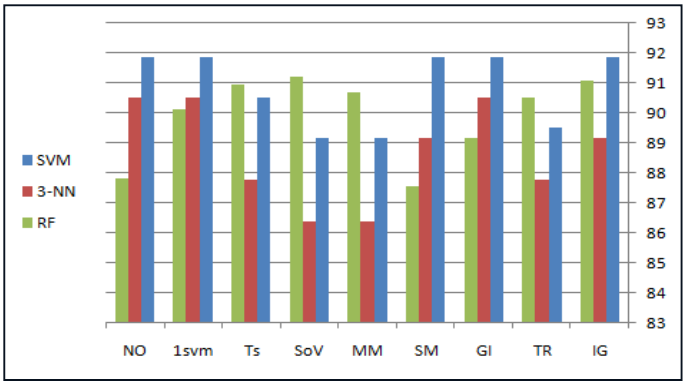


Fig. 2. Accuracy of the 5-fold cross validation of feature selection and classification methods on the Freije data set. No (without feature selection), Ts (t-statistics), SoV (sum of variances), MM (max minority), SM (sum of minority), GI (gini index), TR (twoing rule), and IG (information gain).

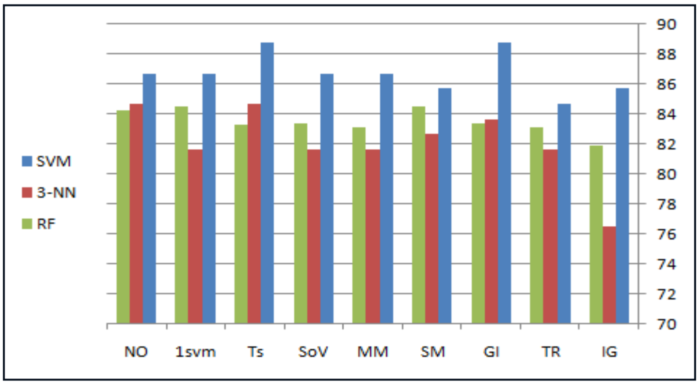


Fig. 3. Accuracy of the 5-fold cross validation of feature selection and classification methods on the Phillips data set.

It is difficult to select the best feature selection method. There does not seem to exist a clear winner. For that we were interested to deeply study the relationship of features selected in each method. Consensus methods have been shown to outperform individual methods in a number of bioinformatics problems[14]. Therefore, we selected the top twenty features that are most frequent among all feature selection methods. The methodology was as the following: as there are 5-fold cross validation, in each fold we applied the feature selection methods on the training set and selected the top twenty features. The union of the top twenty features selected in all folds for each method was recorded. Then we computed the frequency of each feature among all the feature selection methods. We first applied this experiment on the Freije data set, and we found seven features appeared in all methods while seventeen features appeared in seven methods. The correlation coefficient between the seventeen features was high. Thus thirteen of seventeen features were randomly selected. Results of 5-fold cross validation for the twenty common features on the Freije and Phillips data sets are shown in Table 1.

Table 1. Results of the 5-fold cross validation using the twenty most frequent selected from the Freije data set.

	Frieje Dataset		
	Accurecy	**Sensitivity**	**Specificity**
SVM			
linear	**94.59%**	**96.07%**	**91.30%**
polynomial	**93.24%**	**96.07%**	**86.95%**
radial	**94.59%**	**96.07%**	**91.30%**
sigmoid	**94.59%**	**96.07%**	**91.30%**
k-NN			
K=1	**90.54%**	**94.11%**	**82.60%**
K=3	**94.59%**	**96.07%**	**91.30%**
K=5	**93.24%**	**96.07%**	**86.95%**
K=7	**93.24%**	**96.07%**	**86.95%**
K=9	**93.24%**	**96.07%**	**86.95%**
K=11	**93.24%**	**96.07%**	**86.95%**
Random Forest	**90.95%**	**95.49%**	**80.87%**

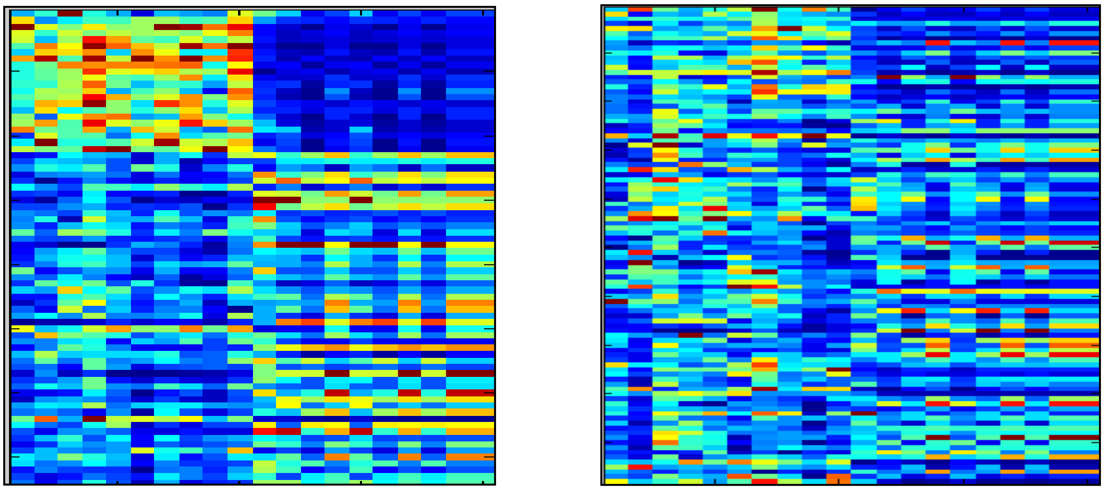
	Phillips Dataset		
	Accurecy	**Sensitivity**	**Specificity**
SVM			
Linear	**80.61%**	**94.66%**	**34.78%**
Polynomial	**76.53%**	**100%**	**0**
Radial	**81.63%**	**94.66%**	**39.13%**
Sigmoid	**84.69%**	**94.66%**	**52.17%**
k-NN			
K=1	**76.53%**	**80.11%**	**65.21%**
K=3	**79.59%**	**85.33%**	**60.86%**
K=5	**82.65%**	**90.66%**	**56.52%**
K=7	**81.63%**	**90.66%**	**52.17%**
K=9	**81.63%**	**90.66%**	**52.17%**
K=11	**80.61%**	**89.33%**	**52.17%**
Random Forest	**79.39%**	**91.20%**	**40.87%**

For the three classification methods, the accuracy performance was improved using these twenty features on the Freije data set. SVM and 3-NN reported the best accuracy 94.59%. Random forest achieved 90.95% with very high sensitivity and specificity. To validate these features, we used the Phillips data set as a validation set. However, the results were not promising. For sigmoid-SVM, it achieved the highest accuracy of 84.69% while polynomial-SVM achieved the minimum accuracy 76.53%. The highest accuracy for KNN when K=5 was 82.65%, and 79.39% for random forest with very high sensitivity. However, the specificity is rather low for the three classification methods. The heat-maps for these features on both data sets are shown in Fig 4.

Here, we can see a cut between the two classes. Ten features are highly expressed in one class, and the other ten are lowly expressed in the same class. But for the Phillips data set we cannot clearly see that. This explains why we had high accuracy on the Freije data set, but not on the Phillips data set.

We repeated the same experiment but this time the most frequent features between the eight feature selection methods were selected from the Phillips data set. The same methodology was followed. The twenty most frequency features were selected. Here we found four features were in common among all methods. These features achieved better accuracy in both data sets than the features selected from the Freije data set. Results are shown in Table 2. SVM reported the best accuracy 91% in both data sets with high sensitivity and specificity. In the Phillips data set, 7-NN and random forest reported good accuracy 88%. In the Freije data set, 9-NN reported accuracy of 87.8% with good sensitivity and specificity 94% and 73%, respectively. Random forest reported 84% accuracy, 91% sensitivity, and 66% specificity.

The heat-maps of these features on both datasets are shown in Fig 5. Here, we can see a cut between two classes in the Phillips dataset. The effect of these features on the Freije data set is better than the effect of the features selected from Freije data on the Phillips data in the previous experiment.



(a) (b)

Fig. 4. Heat-maps of the 20 most frequent features selected from the Freije data set on. Rows represent samples and columns represent features. (a) Freije data; (b) Phillips data.

Table 2. Results of the 5-fold cross validation using the twenty most frequent features selected from the Phillips data set.

	Frieje Dataset		
	Accurecy	Sensitivity	Specificity
SVM			
linear	**90.54%**	**96.07%**	**78.26%**
polynomial	**89.18%**	**96.07%**	**73.91%**
radial	**91.89%**	**96.07%**	**82.60%**
sigmoid	**91.89%**	**94.11%**	**86.95%**
k-NN			
K=1	**77.02%**	**86.27%**	**56.52%**
K=3	**83.78%**	**86.27%**	**78.26%**
K=5	**82.43%**	**88.23%**	**69.56%**
K=7	**86.48%**	**94.11%**	**69.56%**
K=9	**87.83%**	**94.11%**	**73.91%**
K=11	**86.48%**	**96.07%**	**65.21%**
Random Forest	**84.05%**	**91.96%**	**66.52%**

	Phillips Dataset		
	Accurecy	Sensitivity	Specificity
SVM			
linear	**90.81%**	**93.33%**	**82.60%**
polynomial	**90.81%**	**96.07%**	**73.91%**
radial	**90.81%**	**94.66%**	**78.26%**
sigmoid	**91.83%**	**94.66%**	**82.60%**
k-NN			
K=1	**84.69%**	**92.11%**	**60.86%**
K=3	**87.75%**	**92.11%**	**73.91%**
K=5	**86.73%**	**89.33%**	**78.26%**
K=7	**88.77%**	**90.66%**	**82.60%**
K=9	**88.77%**	**92.24%**	**78.26%**
K=11	**87.75%**	**90.66%**	**78.26%**
Random Forest	**88.16%**	**91.46%**	**77.39%**

The next experiment is important to investigate the difference between the two groups of most frequent features selected from the Freije and Phillips data sets. We compared the two groups of features, and there was no feature in common between two groups found. The correlation coefficient of features in the two groups was calculated to find out if there is a relationship between them. We found the seven features in Freije and the four features in Phillips that appear in all feature selection methods are correlated. We then evaluated the performance of the combination of the two sets of twenty features selected from both data sets. The best accuracy was achieved in both data sets with high sensitivity and specificity. Results of the 5-fold cross validation are shown in Table 3 for both data sets. SVM reported the highest accuracies in Freije and Phillips of 94.59% and 90.81%, respectively. KNN achieved its best performance when K is set to 7 in both datasets. Random forest reported an accuracy of 90.54% in the Freije data set and 87.96% in the Phillips data set. The good performance of these features is reflected in heat-maps as well. The heat-maps for these features on both datasets are shown in Fig 6.

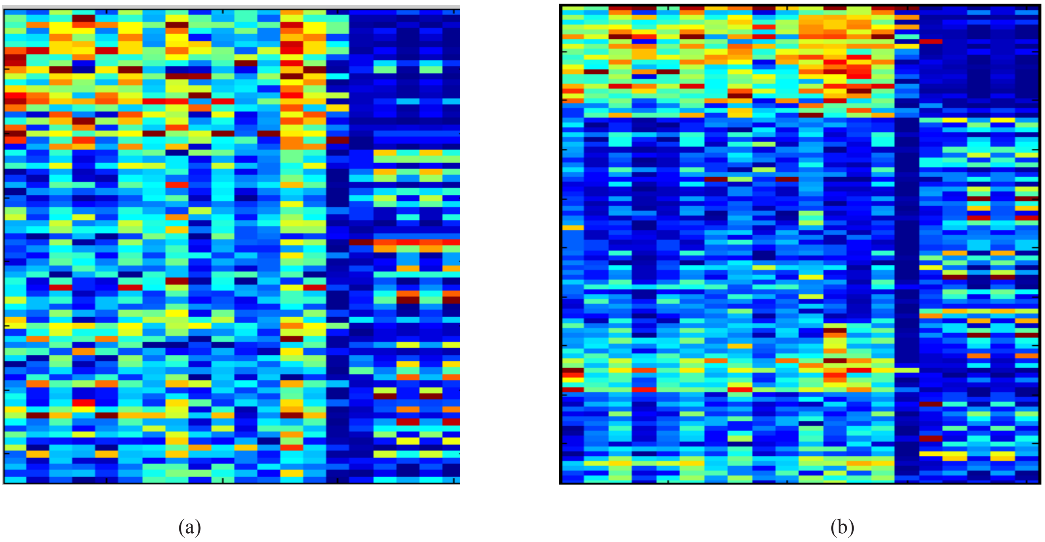


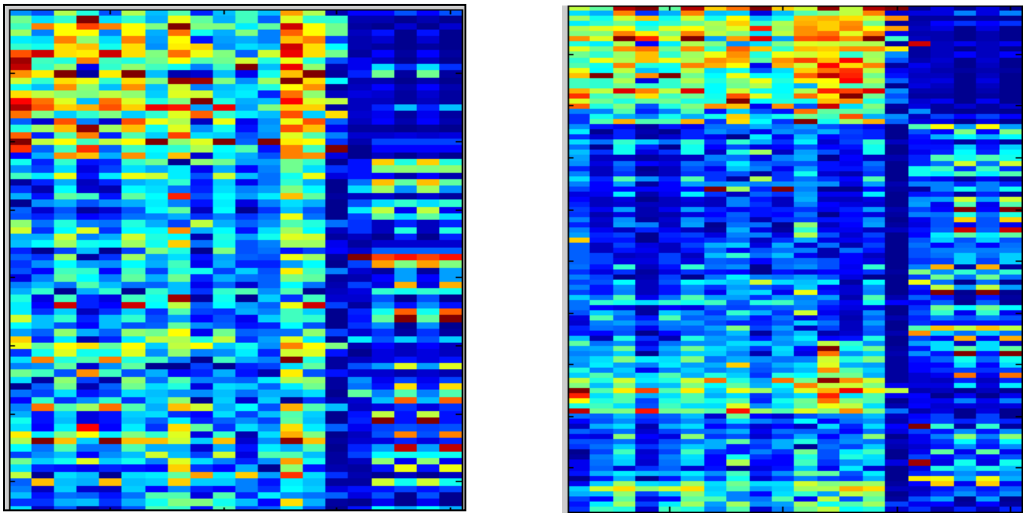
Fig. 5. Heat-maps of the 20 most frequent features selected from the Phillips data set on (a) Freije data; (b) Phillips data.

Table 3. Results of the 5-fold cross validation using the 40 features in the Freije and Phillips data.

	Frieje Dataset		
	Accurecy	**Sensitivity**	**Specificity**
SVM			
linear	**94.59%**	**96.07%**	**91.30%**
polynomial	**91.89%**	**96.07%**	**82.60%**
radial	**94.59%**	**96.07%**	**91.30%**
sigmoid	**94.59%**	**96.07%**	**91.30%**
k-NN			
K=1	**90.54%**	**94.11%**	**82.08%**
K=3	**90.54%**	**96.07%**	**78.26%**
K=5	**90.54%**	**96.07%**	**78.26%**
K=7	**91.89%**	**96.07%**	**82.60%**
K=9	**91.89%**	**96.07%**	**82.60%**
K=11	**90.54%**	**96.07%**	**78.26%**
Random Forest	**90.54%**	**94.51%**	**81.73%**

	Phillips Dataset		
	Accurecy	**Sensitivity**	**Specificity**
SVM			
linear	**90.81%**	**94.66%**	**78.26%**
polynomial	**84.69%**	**92.28%**	**60.86%**
radial	**90.81%**	**94.66%**	**78.26%**
sigmoid	**90.81%**	**94.66%**	**78.26%**
k-NN			
K=1	**82.65%**	**88.45%**	**65.21%**
K=3	**84.69%**	**88.24%**	**73.91%**
K=5	**85.71%**	**88.24%**	**78.26%**
K=7	**86.73%**	**89.33%**	**78.26%**
K=9	**85.71%**	**88.24%**	**78.26%**
K=11	**84.69%**	**86.66%**	**78.26%**
Random Forest	**87.96%**	**91.46%**	**76.52%**

From these forty genes, we found some genes that are related to glioma such as microtubuleassociated protein tau (MAPT)[15,16], bone morphogenic protein receptor, type II (BMPR2)[17], glutamate dehydrogenase 1 (GLUD1)[18], syndecan 1 (SDC1)[19] and programmed cell death 4 (PDCD4).



(a)

(b)

Figure 6. Heat-map of the 40 features selected from both the Freije and Phillips data sets on (a) Freije data (b) Phillips data.

4. Conclusion and future work

We presented a comparative study of state-of-the-art feature selection methods and classification methods based on gene expression data. The efficiency of three different classification methods including: SVM, KNN and random forest, and eight different feature selection methods, including: information gain, twoing rule, sum minority, max minority, gini index, sum of variances, t-statistics, and one-dimension support vector machine was compared. These methods were applied to two publicly available gene expression data sets of glioma. Five-fold cross validation was used to evaluate the classification performance.

In the future, we plan to study the effect of doing feature selection and classification altogether. One possible way is to use wrapper methods. Another possible way is to combine the feature selection in the classification learning process. For instance, we can add a feature selection term in the objective function of SVMs, so that the optimization problem can do feature selection and classification simultaneously. Another direction of future research is to combine the information from different data sets together. It is a commonly seen scenario that there are a number of biological data sets, such as the two used in this thesis, that share the same features but are collected by different groups under different experimental conditions. Thus, they may have different underlying distributions. Yet they share highly relevant information. Each data set may be small and not sufficient to learn a good classifier. In such cases, transfer learning is a possible way to borrow information between the data sets. For instance, if we combine the two data sets used in this thesis together, we will end up with 172 samples (74 from Freije and 98 from Phillips). Direct feature selection and classification on this combined data set does not result in good accuracy, we plan to apply transfer learning techniques on this combined data set in the future.

References

1. Wang JY, Bensmail H, and Gao X. Multiple graph regularized nonnegative matrix factorization. *Pattern Recognition* 2013;**46**(10):2840-2847.
2. Wang JY, Wang X, and Gao X. Non-negative matrix factorization by maximizing correntropy for cancer clustering. *BMC Bioinformatics* 2013;**14**:107.

3. Wang JY, Almasri I, and Gao X. Adaptive graph regularized nonnegative matrix factorization via feature selection. The 21st International Conference on Pattern Recognition (ICPR2012). Tsukuba, Japan. November 2012.
4. Wang JY, Bensmail H, and Gao X. Joint learning and weighting of visual vocabulary for bag-of-feature based tissue classification. *Pattern Recognition* 2013 http://dx.doi.org/10.1016/j.patcog.2013.05.001.
5. Kohavi R. and John G. Wrappers for feature subset selection. *Artificial Intelligence* 1997; **97**(1-2):273-324.
6. Xing E, Jordan M, and Karp R. Feature selection for high dimensional genomic microarray data. In Proceedings of the 18th International Conference on Machine Learning. 2001. p. 601-608.
7. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study on different machine learning methods on microarray gene expression data. *BMC Genomics* 2008;**9** Suppl 1:S13.
8. Liu H, Li J, and Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics* 2002;**13**:51–60.
9. Li T, Zhang C, and Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 2004;**20**(15):2429-2437.
10. Su Y, Murali T, Pavlovic V, Schaffer M, and Kasif S. RankGene: identification of diagnostic genes based on expression data. *Bioinformatics* 2003;**19**(12):1578-1579.
11. Freije W, Castro-Vargas F, and Fang Z. Gene expression profiling of gliomas strongly predicts survival. *Cancer Research* 2004;**64**:10-6503.
12. Phillips H, Kharbanda S, and Chen R. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 2006;**9**:73-157.
13. Barrett T, Troup D, Wilhite S, Ledoux P, and Rudnev D. NCBI GEO: archive for high throughput functional genomic data. *Nucleic Acids Research* 2009;**37**:885.
14. Gao X, Bu D, Xu J, and Li M. Improving consensus contact prediction via server correlation reduction. *BMC Structural Biology* 2009;**9**(1):28.
15. Stoothoff W, Jones P, Spires-Jones T, Joyner D, Chhabra E, Bercury K, Fan Z, Xie H, Bacskai B, Edd J, Irimia D, and Hyman B. *J Neurochem*. Differential effect of three-repeat and four-repeat tau on mitochondrial axonal transport. 2009;**11**: 417-427.
16. Shafaati M, Solomon A, Kivipelto M, Björkhem I, Leoni L. Levels of ApoE in cerebrospinal fluid are correlated with Tau and 24S-hydroxycholesterol in patients with cognitive disorders. *Neurosci Lett* 2007;**425**:78-82.
17. Rosenzweig B, Imamura T, Okadome T, Cox G, Yamashita H, Dijke P, Heldin C, and Miyazono K. Cloning and characterization of a human type II receptor for bone morphogenetic proteins. *Proceedings of the National Academy of Science of the United States of America* 1995;7632-7636.
18. Yang C, Sudderth J, Dang T, Bachoo R, McDonald J. DeBerardinis RJ Glioblastoma cells require glutamate dehydrogenase to survive impairments of glucose metabolism or Akt signaling. *Cancer Research* 2009;**69**:7986–7993.
19. Watanabe A, Mabuchi T, Satoh E, Furuya K, Zhang L, Maeda S, and Naganuma H. Expression of syndecans, a heparan sulfate proteoglycan, in malignant gliomas: participation of nuclear factor-kappaB in upregulation of syndecan-1 expression. *Neuro Oncology* 2006;**77**:25-32.