An Integrated Algorithm for Dimension Reduction and Classification Applied to Microarray Data of Neuromuscular Dystrophies

Divya1, Babita Pandey2* and Devendra K. Pandey3

1School of Computer Science and Engineering, Lovely Professional University, Chaheru, Phagwara - 144411, Punjab, India; divyaanand.y@gmail.com
2School of Computer Applications, Lovely Professional University, Chaheru, Phagwara - 144411, Punjab, India; shukla_babita@yahoo.co.in
3School of Biosciences, Lovely Professional University, Chaheru, Phagwara - 144411, Punjab, India; dkpandey1974@yahoo.com

Abstract

Background/Objectives: Microarray technology allows the neuromuscular dystrophy to be predicted using gene expression patterns. Microarray gene expression data suffer from curse of high dimensionality i.e. tens of thousands of genes and few samples. So, it is necessitate reducing the dimension for accurate diagnosis. Methods/Statistical Analysis: Firstly, five-fold cross validation technique is applied to generate random results. Two feature selection techniques i.e. t-test and entropy are employed to select the genes. K-nearest neighbor and linear support vector machine are deployed for classification of diseased samples with the help of ranked genes. The performance of these integrated techniques is tested on the microarray dataset of neuromuscular dystrophies i.e. Juvenile Dermatomyositis (JDM) and Fascioscapulohumeral Muscular Dystrophy (FSHD). Findings: Effective disease specific genes are selected from thousand of genes. The value of various performance measures shows that the integration of entropy with k-nearest neighbor has outperformed on both datasets. It has given 89.47% accuracy on JDM dataset and 100% accuracy on FSHD dataset. The integration of these methods is first time application on these two diseases datasets. It can be applied on other neuromuscular disorder datasets as well.

Keywords: Dimension Reduction, Entropy, K-Fold Validation, K-Nearest Neighbor, Neuromuscular Dystrophy, Support Vector Machine

1. Introduction

Microarray technology is used to monitor the genome wide expression level of genes in a given organism. They provide us a way to obtain expression level of thousands of genes simultaneously under a particular condition1. In microarrays, gene expression matrix is formed where each row represents a gene and each column represents the expression level of genes in a sample. The value in gene expression matrix signifies the expression value of a gene in a sample. These days, microarrays are used to diagnose the diseases using gene expression values1.

The Gene Expression Value (GEV) generated using microarrays can be used to classify the muscular dystrophies. Each type of muscular dystrophy is formed due to an error in a specific gene associated with its muscle function. It is very difficult to classify the muscular dystrophy because the traditional methods only tells us that the person is suffering from neuromuscular disorders but they are deficient in telling us the kind/type of neuromuscular dystrophy. So, the neurologist always prefers the genetic diagnosis. The genetic diagnosis is performed by taking a blood sample from patient and experiment it on microarrays. The microarrays generate the data which is used further for diagnosis.

*Author for correspondence
The classification of muscular dystrophies using the microarray data is a bit challenging because of two reasons. Firstly, the microarray data is high dimensional as it contains a large number of genes i.e. tens of thousands of genes. Secondly, it contains very less samples i.e. a very few number of patients. As only a few genes are related to the diseases, rest other genes are noisy\(^1\). So, in order to do the proper diagnosis to get the accurate diagnosis results, we need to reduce the dimension to eliminate the noisy genes.

For the classification, dimensionality reduction techniques reduce noise in the data which minimizes redundancy and maximizes relevance\(^2\). These techniques increase the classification accuracy as the uninformative and redundant genes are removed in the feature selection phase. The dimensionality reduction techniques are classified into two main categories namely feature selection and feature extraction\(^2\). The feature selection techniques select a small subset of features from large dataset whereas feature extraction techniques transform features into a lower dimensional feature space.

Earlier, a number of techniques have been applied for feature selection on microarray datasets for classification of cancers but almost negligible work has been done on the genetic diagnosis of muscular dystrophies. The various ranking and selection techniques used for feature selection are t-test\(^3\)–7, chi-square test\(^8\), signal to noise ratio\(^9,10\), family wise error rate\(^11\), F-Score\(^12\) and ANOVA\(^13,14\). The various feature reduction methods namely PCA, t-test, class seperability measure and Fisher ratio are used for cancer classification\(^3\). The use of mutual information and statistical measures for feature selection for classification of four dataset using naive Bayes and SVM is shown\(^15\). Data mining techniques are also used for classification of neurodegenerative diseases\(^16\).

In the present paper, we use two feature selection techniques i.e. t-test and entropy and two classification algorithms i.e. K-Nearest Neighbor (K-NN) and linear Support Vector Machine (SVM) for the diagnosis of neuromuscular dystrophy. Two microarray datasets of neuromuscular dystrophies i.e. Juvenile Dermatomyositis (JDM) and Facioscapulohumeral Muscular Dystrophy (FSHD) is used for employing the integration of these techniques. The results show that the integration of entropy with K-NN has shown better performance in both cases.

JDM is an autoimmune disease which generally affects children. It usually affects 3000-5000 children in United States each year. It is a genetic disease which tends to run in the families of patients. In this, the muscle weakness may results in dysphonia, fatigue, weight loss, clumsiness and other issues. FSHD is an autosomal dominant neuromuscular dystrophy. The muscle weakness mostly occurs in the muscles of face, shoulder, arms and stomach. As such, there is no cure for these diseases. But a proper and accurate diagnosis of the disease could help the patient in many ways.

The paper is structured as follows: In section 2, we describe the different gene selection and classification methods used in the paper. In section 3, we describe the two datasets taken for the application of integrated algorithms and compare the results of these algorithms using various performance measures. Conclusion is given in section 4.

## 2. Methods

Classification is the process of identification of categories of test observations (unseen data) on the basis of the training observations (seen data). Feature selection for classification selects the subset of highly discriminative features from training data. The overall framework used in the present paper is shown in Figure 1. The division of data into training set and test set and to generate the unbiased results, five-fold cross validation strategy is used. Cross validation is a technique used to divide the data into two segments in which one segment is used to train the model and the other segment is used to evaluate or validate the model. In five-fold cross validation strategy, every time test dataset contains only one set while training dataset contain remaining four sets. Two gene selection techniques namely t-test and entropy are used on the training set. Two classification techniques K-NN and linear SVM are applied to classify the data. The performance of integrated algorithm is evaluated using various performance measures. These are accuracy, sensitivity, specificity, Positive Predicted Value (PPV) and Negative Predicted Value (NPV).

### 2.1 Gene Selection Methods

#### 2.1.1 T-Test

To rank and select the genes, firstly t-test is applied. The procedure for calculating p-value using t-test is: Calculate the Mean (M) of the GEV of the diseased samples for the first gene. Calculate the M of GEV of the non-diseased or
2.1.2 Entropy

Entropy is a measure of difference between two probability distributions\(^7\). Suppose the probability functions of two discrete distributions A and B are \(A_k\) and \(B_k\) respectively. Then the information divergence or relative entropy or Kullback-Leibler divergence distance of A with respect to B is given in Equation 2

\[
e = \sum_k A_k \log B_k
\]  

(2)

The relative entropy of all the genes is calculated and genes are arranged in descending order of their value. From them, top 500 genes are selected and are used for classification using K-NN and linear SVM.

2.2 Classification Methods

2.2.1 K- Nearest Neighbor

K-NN is a lazy\(^8\) and supervised learning technique. It is a non-parametric method which can be used for both classification and regression\(^9\). The input to the algorithm is given in the form of vectors as a training set with their selectors which must reside in the memory at the run time. So, during training, the inputs with their selectors are just stored in the memory. During classification or testing, any unlabeled/unseen vector is classified by assigning the selector which is for the most part used among the k training samples nearest to that unseen vector. That is why it is called memory-based classification.

In the present work, the cosine distance function is used and the value of k is chosen to be 5.

2.2.2 Support Vector Machine

SVM was invented in 1963 by Vladimir N. Vapnik and Alexey Ya. Chervonenkis. It is a supervised learning model used to discover informative patterns, classification and regression tasks. In a high-dimensional space, SVM constructs a set of hyperplanes which classifies the data into different classes. A hyperplane is a line through the origin which satisfies the Equation 3

\[
(z, y) = 0
\]

(3)

It is responsible for dividing the space into two halves as per the sign of f(x). There will be infinite number of hyper planes which proves the inequalities. But the optimal hyper plane will be that one hyper plane which accurately classifies training data with a large margin as
well as the test data which are unseen by the classifier at the time of training. From a set of hyper planes, the best hyper plane is chosen which neatly separates the data into different classes and the distance between the hyper plane and the classes can be maximized. Support vectors are the data points which lie nearest to the decision surface. SVM tries to maximize the margin between the separating hyper planes. In a linear classification task, each feature belonging to one of the either classes, SVM builds a model which separates the new data into their respective class. In the present paper, we have used linear kernel SVM. The linear SVM classifier is defined as the inner product between two vectors is given in Equation 4

\[ \langle z, y \rangle = \sum_{j=1}^{M} z_j y_j \]  

The decision function \( f(y) \) decides how to classify the data and assigns a score for the input \( y \). The decision function for linear classifier is of the form

\[ f(y) = \langle z, y \rangle + b \]  

where \( z \) is the weight vector and \( b \) is the bias. It should satisfy the following set of inequalities:

\[ f(x) = \begin{cases} > 0, & y_j \in c_1 \\ < 0, & y_j \in c_2 \end{cases} \]  

3. Results and Comparisons

The proposed integrated algorithms are evaluated on Juvenile Dermatomyositis (JDM) and Fascioscapulohumeral Muscular Dystrophy (FSHD) datasets. The dataset are downloaded from Gene Expression Omnibus (GEO) under experiment E-GEOD-3307 – Transcriptional profiling by array of 12 human muscle diseases\(^{20}\). The experiment was done on platform Affymetrix Human Gene 1.0 ST Array [transcript (gene) version].

3.1 Juvenile Dermatomyositis

The JDM dataset contains the GEV of human skeletal muscles. The microarray data contains the expression level of 22,645 genes and involves 21 samples which are affected by JDM and 18 healthy samples (not affected by JDM). We conduct five-fold cross validation experiments on the dataset, 34-35 training samples and 5-4 test samples in every run. We compare the performance measures of above mentioned integrated algorithms. These are t-test – linear SVM, t-test – K-NN, entropy – linear SVM and entropy – K-NN on JDM and FSHD datasets. Table 1 and Figure 2 illustrate the results of the integrations t-test – linear SVM and t-test – K-NN. We observe that the classifiers do not result same after ranking and selecting the genes using t-test. The integration t-test – K-NN is found to be more efficient as it is giving 100% accuracy in training dataset and 89.74% accuracy in test dataset.

We have shown the results of other integrated algorithms namely entropy – linear SVM and entropy – K-NN in Table 2 and Figure 3. Here also, the classifiers have shown different performance after ranking and selecting the genes using entropy. The classifier K-NN has shown the superior performance than linear SVM. It has given 100% performance measures in training dataset and better accuracy i.e. 92.31% in test dataset as compared to the previous integrated algorithm.

![Figure 2. Integration of t-test – linear SVM and t-test – k-NN on JDM.](image)

Table 1. Performance measures of integration of t-test – linear SVM and t-test – k-NN on JDM

| Classification Technique | Data | Accuracy | Sensitivity | Specificity | PPV | NPV |
|--------------------------|------|----------|-------------|-------------|-----|-----|
| Linear SVM               | Training | 78.87%  | 78.95%      | 78.79%      | 81.80% | 76.47% |
|                          | Test   | 74.36%  | 76.19%      | 72.22%      | 76.19% | 72.22% |
| K-NN                     | Training | 100%   | 100%        | 100%        | 100%  | 100% |
|                          | Test   | 89.74%  | 90.48%      | 88.89%      | 90.48% | 88.89% |
3.2 Fascioscapulohumeral Muscular Dystrophy

The FSHD dataset also contains the gene expression data of human skeletal muscles. It consists of total 32 samples in which 14 samples are affected by FSHD and 18 samples are healthy (not affected by FSHD). Here also, each sample contains 22,645 genes. In this also, we conduct 5 fold cross-validation experiments on the dataset i.e. 27-28 training samples vs. 5-4 test samples are used in every run. The performance of t-test – linear SVM, t-test – k-NN, entropy – linear SVM and entropy – k-NN on FSHD datasets is evaluated. Table 3, Figure 4, Table 4 and Figure 5 visualize the results of the integrations performed on FSHD dataset. Table 3 and Figure 4 show the performance measures of first two integrations t-test – linear SVM and t-test – k-NN on FSHD dataset.

Here also, the integrated algorithm t-test – k-NN has outperformed than t-test – linear SVM. The training dataset has given 100% in all the performance measures whereas the accuracy of test dataset is 84.38%. The best results on FSHD datasets were found when the entropy is integrated with k-NN. Like the previous integration, the training dataset has given 100% of all the performance measures and test dataset has given 96.88% accuracy.

Table 2. Performance measures of integration of entropy – linear SVM and entropy – k-NN on JDM

| Classification Technique | Data | Accuracy | Sensitivity | Specificity | PPV   | NPV   |
|--------------------------|------|----------|-------------|-------------|-------|-------|
| Linear SVM               | Training | 76.06%   | 86.84%      | 63.64%      | 73.33%| 80.77%|
|                          | Test    | 74.36%   | 88.10%      | 58.33%      | 71.15%| 80.77%|
| K-NN                     | Training | 100%     | 100%        | 100%        | 100%  | 100%  |
|                          | Test    | 92.31%   | 90.48%      | 94.44%      | 95%   | 89.47%|

Figure 3. Integration of entropy – linear SVM and entropy – k-NN on JDM.

Table 3. Performance measures of integration of t-test – linear SVM and t-test – k-NN on FSHD

| Classification Technique | Data | Accuracy | Sensitivity | Specificity | PPV   | NPV   |
|--------------------------|------|----------|-------------|-------------|-------|-------|
| Linear SVM               | Training | 72.41%   | 81.82%      | 60%         | 72.97%| 71.43%|
|                          | Test    | 68.75%   | 77.78%      | 57.14%      | 70%   | 66.67%|
| K-NN                     | Training | 100%     | 100%        | 100%        | 100%  | 100%  |
|                          | Test    | 84.38%   | 88.89%      | 78.57%      | 84.21%| 84.62%|

Figure 4. Integration of t-test – linear SVM and t-test – k-NN on FSHD.

Table 4. Performance measures of integration of entropy – linear SVM and entropy – k-NN on FSHD

| Classification Technique | Data | Accuracy | Sensitivity | Specificity | PPV   | NPV   |
|--------------------------|------|----------|-------------|-------------|-------|-------|
| Linear SVM               | Training | 81.03%   | 90.91%      | 68%         | 78.95%| 85%   |
|                          | Test    | 79.69%   | 91.67%      | 64.29%      | 76.74%| 85.71%|
| K-NN                     | Training | 100%     | 100%        | 100%        | 100%  | 100%  |
|                          | Test    | 96.88%   | 100%        | 92.86%      | 94.74%| 100%  |
4. Conclusion

For neuromuscular dystrophies, it is a complex task to find specific genes which assists in diagnosing the diseases. In the present paper, the proposed integrated algorithms select genes and classify the microarray data of Juvenile Dermatomyositis (JDM) and Fascioscapulohumeral Muscular Dystrophy (FSHD). The genes are ranked and selected using t-test and entropy which can successfully classify the gene expression values using K-Nearest Neighbor (K-NN) and linear Support Vector Machine (SVM). The integration of entropy with K-NN has given the best performance measures for gene selection and classification of both JDM and FSHD and it can also be applied for the diagnosis of other neuromuscular dystrophies using microarray data.

5. References

1. Babu MM. Introduction to microarray data analysis. Computational Genomics: Theory and Application. 2004; 17(6):225–49.
2. Sharma A, Paliwal KK. Cancer classification by gradient LDA technique using microarray gene expression data. Data and Knowledge Engineering. 2008 Aug; 66(2):338–47.
3. Chu F, Wang L. Applications of support vector machines to cancer classification with microarray data. International Journal of Neural Systems. 2005 Dec; 15(6):47–84.
4. Tang J, Alelyani S, Liu H. Exploiting social relations for sentiment analysis in microblogging. Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM’13); 2013. p. 537–46.
5. Yendrapalli K, Basnet RB, Mukkamala S, Sung AH. Gene selection for tumor classification using microarray gene expression data. World Congress on Engineering. 2007 Jul; p. 290–5.
6. Wang S, Wang Y, Du W, Sun F, Wang X, Zhou C, Liang Y. A multi-approaches-guided genetic algorithm with application to opeor prediction. Artificial Intelligence in Medicine. 2007 Oct; 41(2):151–9.
7. Zheng CH, Chong YW, Wang HQ. Gene selection using independent variable group analysis for tumor classification. Neural Computing and Applications. 2011 Mar; 20(2):161–70.
8. Chen YC, Ke WC, Chiu HW. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Computers in Biology and Medicine. 2014 May; 48:1–7.
9. Ziaei L, Mehri AR, Salehi MA. Application of artificial neural networks in cancer classification and diagnosis prediction of a subtype of lymphoma based on gene expression profile. Journal of Research in Medical Sciences. 2006 Jan; 11(1):13–7.
10. Chen AH, Lin CH. A novel support vector sampling technique to improve classification accuracy and to identify key genes of leukaemia and prostate cancers. Expert Systems with Applications. 2011 Apr; 38(4):3209–19.
11. Zhang JG, Deng HW. Gene selection for classification of microarray data based on the Bayes error. BMC Bioinformatics. 2007 Oct; 8(1):370.
12. Sahu SS, Panda G, Barik RC. A hybrid method of feature extraction for tumor classification using microarray gene expression data. Int J Comput Sci Informatics India. 2011; 1(1):1–5.
13. Lee ZJ. An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer. Artificial Intelligence in Medicine. 2008 Jan; 42(1):81–93.
14. Ananda Kumar K, Punithavalli DM. Efficient cancer classification using Fast Adaptive Neuro-Fuzzy Inference System (FANFIS) based on statistical techniques. IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence. 2011; 132–7.
15. Das K, Ray J, Mishra D. Gene selection using information theory and statistical approach. Indian Journal of Science and Technology. 2015 Apr; 8(8):695.
16. Suganya P, Sumathi CP. A novel metaheuristic data mining algorithm for the detection and classification of Parkinson disease. Indian Journal of Science and Technology. 2015 Jul; 8(14):1–1.
17. Qian H. Relative Entropy: Free Energy Associated with Equilibrium Fluctuations and Nonequilibrium Deviations. 2000. Available from: http://arxiv.org/abs/math-ph/0007010
18. Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition. 2007 Jul; 40(7):2038–48.
19. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992; 46(3):175–85.
20. Bakay M, Wang Z, Melcon G, Schiltz L, Xuan J, Zhao P, Sartorelli V, Seo J, Pegoraro E, Angelini C, Shneiderman B. Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb–MyoD pathways in muscle regeneration. Brain. 2006 Apr; 129(4):996–1013.