Multicategory classification of 11 neuromuscular diseases based on microarray data using support vector machine

Soo Beom Choi, Jee Soo Park, Jai Won Chung, Tae Keun Yoo, Deok Won Kim, *Life member, IEEE**

Abstract— **We applied multicategory machine learning methods to classify 11 neuromuscular disease groups and one control group based on microarray data. To develop multicategory classification models with optimal parameters and features, we performed a systematic evaluation of three machine learning algorithms and four feature selection methods using three-fold cross validation and a grid search. This study included 114 subjects of 11 neuromuscular diseases and 31 subjects of a control group using microarray data with 22,283 probe sets from the National Center for Biotechnology Information (NCBI). We obtained an accuracy of 100%, relative classifier information (RCI) of 1.0, and a kappa index of 1.0 by applying the models of support vector machines one-versus-one (SVM-OVO), SVM one-versus-rest (OVR), and directed acyclic graph SVM (DAGSVM), using the ratio of genes between categories to within-category sums of squares (BW) feature selection method. Each of these three models selected only four features to categorize the 12 groups, resulting in a time-saving and cost-effective strategy for diagnosing neuromuscular diseases. In addition, a gene symbol, SPP1 was selected as the top-ranked gene by the BW method. We confirmed relationships between the gene (SPP1) and Duchenne muscular dystrophy (DMD) from a previous study. With our models as clinically helpful tools, neuromuscular diseases could be classified quickly using a computer, thereby giving a time-saving, cost-effective, and accurate diagnosis.**

I. Introduction

Cells' genetic information is stored in DNA, and all cells in an organism have exactly the same genome. However, due to different tissue types, development stages, and even environmental conditions, genes from cells within the same organism can be expressed in different combinations and quantities during transcription from DNA to messenger RNA. These different gene expression patterns can distinguish people with and without diseases [1]. DNA microarray analysis is a highly promising technique with broad applications. Because of the high-throughput nature of microarray data and the very large number of genes, computational tools are essential in data analysis and mining, helping biomedical researchers to extract knowledge from experimental results [2].

Machine learning is an area of artificial intelligence research that uses statistical methods to classify data. Several machine learning techniques have been applied in clinical settings to predict diseases, and have shown higher diagnostic accuracy than classical methods such as clinical scoring systems or logistic regression analyses. Several machine learning techniques have been applied in the area of cancer diagnosis classifications based on gene expression profiling [3].

Understanding of neuromuscular diseases has improved as knowledge has expanded on basic neuromuscular processes. However, it is still quite difficult to discriminate various neuromuscular diseases with conventional examinations due to their indistinguishable symptoms. Sakellariou et al. used microarray data to perform the binary classifications of several neuromuscular diseases [4]. In this study, we developed multicategory classification models for reliable and discriminative diagnosis, and we identified biomarkers for 11 neuromuscular diseases using three machine learning algorithms for DNA microarray technology. The ability to classify 11 different neuromuscular diseases simultaneously with only one muscle biopsy could reduce both examination costs and pain for patients.

II. Materials and Methods

A. Data Acquisition

We collected 145 raw DNA microarray data for the 11 different neuromuscular disease groups and one control group from the gene expression omnibus repository at the National Center for Biotechnology Information (NCBI). Each microarray data sample was obtained from each skeletal muscle biopsy. The microarray data originated from HG_U133A GeneChips (Affymetrix Inc., Santa Clara, CA), and comprised 22,283 probe sets representing 14,500 well-characterized human genes.

The eleven neuromuscular diseases included amyotrophic lateral sclerosis (ALS), Becker muscular dystrophy (BMD), Duchenne muscular dystrophy (DMD), Emery-Dreifuss muscular dystrophy (EDMD), Fascioscapulohumeral muscular dystrophy (FSHD), juvenile dermatomyositis (JDM), limb-girdle muscular dystrophy by calpain 3 mutation (LGMD2A), limb-girdle muscular dystrophy by dysferlin mutation (LGMD2B) [5], dermatomyositis (DM) [6], mitochondrial encephalo-myopathy (MM) [7] and inclusion body myopathy (IBM) [8] (Table I). The inclusion of these 11

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2012R1A2A2A03045612).

S. B. Choi is with Brain Korea 21 PLUS Project for Medical Science, Yonsei University, Seoul, Korea (e-mail: plains7@yuhs.ac).

J. S. Park is with Dept. of Medicine, Yonsei University College of Medicine, Seoul, Korea (e-mail: sampark@yuhs.ac).

J. W. Chung is with the Graduate Program in Biomedical Engineering, Yonsei University, Seoul, Korea (e-mail: chjw0915@yuhs.ac).

T. K. Yoo is with Dept. of Medicine, Yonsei University College of Medicine, Seoul, Korea (e-mail: fawoo2@yuhs.ac).

*D. W. Kim is a Professor at Dept. of Medical Engineering, Yonsei University College of Medicine, Seoul, Korea (corresponding author; phone: 82-2-2228-1916; fax: 82-2-364-1572; e-mail: kdw@yuhs.ac).

TABLE I. DATASET FOR 11 NEUROMUSCULAR DISEASE GROUPS AND ONE CONTROL GROUP

Data source	Neuromuscular disease	No. of patients	No. of control group members
Bakay et al. [5]	ALS	9	18
	BMD	5	
	DMD	10	
	EDMD	8	
	FSHD	14	
	JDM	21	
	LGMD2A	10	
	LGMD2B	10	
Nagaraju et al. [6]	DM	5	-
Crimi et al. [7]	MM	12	3
Eisenberg et al. [8]	IBM	10	10
Total	(11 diseases + control)	114	31

ALS = amyotrophic lateral sclerosis, BMD = Becker muscular dystrophy, DMD = Duchenne muscular dystrophy, EDMD = Emery-Dreifuss muscular dystrophy, FSHD = Fascioscapulohumeral muscular dystrophy, JDM = juvenile dermatomyositis, LGMD2A = limb-girdle muscular dystrophy by calpain 3 mutation, LGMD2B = limb-girdle muscular dystrophy by dysferlin mutation, DM = dermatomyositis, MM = mitochondrial encephalo-myopathy, IBM = inclusion body myopathy.

neuromuscular disease groups in this study was based solely on the availability of their corresponding microarray data in the NCBI database.

For analysis by machine learning, we transformed the raw DNA microarray data format into the MATLAB data format. In particular, we performed $\log_2$ transformation and mean normalization over all the genes for better analysis. For validation of each machine learning method, the data were randomly separated into two independent data sets: training and testing sets (Fig. 1). The training set, comprising 66.7% (97 subjects) of the overall dataset, was used to construct three SVM models. The testing set, comprising 33.3% (48 subjects) of the overall dataset, was used to assess the model's ability to categorize subjects into the 11 neuromuscular disease groups and one control group.

B. *Feature selection*

Feature selection was necessary to reduce the very high dimensionality of the dataset. The most important objectives of feature selection are: 1) to avoid over-fitting and improve prediction performance; 2) to make quicker and more cost-effective models; and 3) to offer deeper insight into underlying processes [9]. Regardless of the classification methods, many microarray-based studies suggest that gene selection is vital for achieving a high level of generalization [10]. The feature selection methods ranked all gene probe sets based on the calculated weights of each gene. We used four feature selection methods, which are filter methods: (1) Kruskal–Wallis non-parametric one-way ANOVA (KW); (2) ratio of genes between-categories to within-category sums of squares (BW); (3) signal-to-noise scores applied in a one-versus-one (S2N-OVO) fashion; and (4) S2N scores

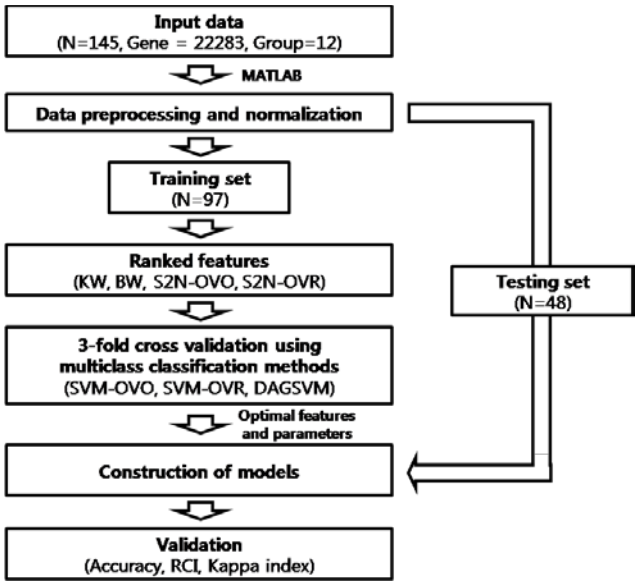

Figure 1. Flow chart for neuromuscular disease classification.

applied in a one-versus-rest (S2N-OVR) fashion. KW is a non-parametric method for testing whether samples originate from the same distribution. Dudoit et al. proposed the BW sum of squares across all paired classes for multicategory classification [11]. S2N is obtained by dividing the difference of the means of two groups by the sum of the standard deviations of those two groups [12].

C. *Support vector machines*

We used three multicategory classification methods based on the binary SVM method: multiclass SVM one-versus-one (SVM-OVO), multiclass SVM one-versus-rest (SVM-OVR), and directed acyclic graph SVM (DAGSVM) [13]. Binary SVMs are learning and pattern recognition algorithms developed with the goal of separating classes by a function that is computed from available examples. The goal is to find a hyper plane that maximizes the separation or margin between two classes [14]. To solve multicategory problems using machine learning, classification methods actually use some combination of binary classifiers. For the OVO comparison approach, to make binary classifiers applicable to multicategory problems, c(c-1)/2 binary classifiers, where c is the number of classes, should be built for SVM to distinguish between every two class combination. Similarly, for the OVR approach, c binary classifiers should be built for SVM to distinguish one class from all the rest of classes [15]. In DAGSVM, the training phase of the algorithm is similar to that of the OVO approach using multiple binary SVM classifiers. However, the testing phase of DAGSVM requires construction of a rooted binary decision-directed acyclic graph using c(c-1)/2 binary classifiers [10].

D. *Model selection and validation*

We used the three-fold cross validation scheme to construct machine learning models (Fig. 1) instead of 10-fold cross validation, which is generally known to be the standard method in practical terms because there were only three samples in the DM disease group within the training set [16].

We determined the order of the variables with four feature selection methods (KW, BW, S2N-OVO, S2N-OVR) for multicategory classification, and then found the optimal variables to construct classification models by increasing the number of variables in the order of their importance using the so-called sequential forward selection (SFS) as the wrapper method [9].

To obtain the optimal results, we adopted a grid search in which a range of parameter values were tested using the three-fold cross validation strategy. For use with the multicategory SVM methods, we chose a radial basis function among the available kernel functions based on the recommendation of a practical guide [17]. The parameter values included a penalty parameter C and a scaling factor σ for the multicategory SVM methods. We found the best classification model and employed its parameters for prediction.

We evaluated the models' diagnostic ability based on accuracy, relative classifier information (RCI), and the kappa index for the testing set. RCI, a parameter of an entropy-based measure of classifier performance, can be measured by the difference in the prior and posterior uncertainties. The kappa index is a statistical measure of inter-rater agreement or inter-annotator agreement for categorical items. We used MATLAB 2012a (Mathworks Inc. Natick, MA) to analyze machine learning and SPSS 20.0 (SPSS Inc., Chicago, IL) for statistical analysis.

III. Result

Although four feature selection methods were performed to reduce the dimensionality of the dataset, we displayed results of BW and S2N-OVO only with good performance because of the space limitation. Table II shows the top 5 ranked probe sets and genes for two feature selection methods as applied to the training set. The secreted phosphoprotein 1 (SPP1) was selected as the top-ranked gene by the BW method. A recent study suggested that the SPP1 genotype is a determinant of disease severity in DMD, based on single nucleotide polymorphism (SNP) analyses in two DMD cohorts [18].

We obtained accuracy, RCI, and kappa index measurements for the testing set to evaluate the three SVM models with the optimal parameters and features determined.

TABLE II. The top 5 probe sets and genes for two feature selection method used in the training set

BW		**S2N-OVO**	
Probe set	***Gene symbol***	***Probe set***	***Gene symbol***
209875_s_at	SPP1	210926_at	ACTBL3
220295_x_at	DEPDC1	209714_s_at	CDKN3
213348_at	CDKN1C	214680_at	NTRK2
208070_s_at	REV3L	207117_at	ZNF117
216114_at	NCKIPSD	204534_at	VTN

BW = ratio of genes between-categories to within-category sums of squares, S2N = signal-to-noise scores, OVO = one-versus-one

TABLE III. Classification performance of each machine learning method with two feature selection method when applied to the testing set

Algorithm	**Feature selection**							
	BW				***S2N-OVO***			
	ACC (%)	***RCI***	***Kappa***	***No. of features***	***ACC (%)***	***RCI***	***Kappa***	***No. of features***
SVM -OVO	100	1.000	1.000	4	100	1.000	1.000	15
SVM -OVR	100	1.000	1.000	4	100	1.000	1.000	27
DAG SVM	100	1.000	1.000	4	100	1.000	1.000	15

BW = ratio of genes between-categories to within-category sums of squares, S2N = signal-to-noise scores, OVO = one-versus-one, ACC = accuracy, RCI = relative classifier information, Kappa = kappa index, SVM = support vector machine, OVR = one-versus-rest, DAG = directed acyclic graph.

Table III summarizes the overall classification performance of the three machine learning methods when each of the two feature selection methods was used on the testing set. Using the SVM-OVO, SVM-OVR, DAGSVM learning methods with the BW and S2N-OVO feature selection methods all showed perfect performance, with accuracy of 100%. However, the S2N-OVO used more number of features, 15 and 27, for all three SVM methods than the BW method, which used only 4 to categorize subjects into the 12 groups, providing an opportunity for time-saving and cost-effective diagnosis of neuromuscular diseases.

The optimal models of SVM-OVO and DAGSVM with the BW method were found using a radial basis function with a penalty parameter C of 300 and scaling factor σ of 100. The optimal parameters of SVM-OVR with BW were C of 300 and σ of 300. For the training set, the accuracy of the cross-validation when using the optimal parameters and features with the BW method were 99.0%, 100.0%, and 99.0% for SVM-OVO, SVM-OVR, and DAGSVM, respectively; these results are not included in Table III. These results indicate that the trained models were not over-fitting, considering the accuracies for the training and testing sets.

IV. Discussion and Conclusion

Using the three SVM algorithms for multicategory classification, we searched an optimal approach for clinical decision support based on microarray data on 11 neuromuscular disease groups and one control group. Our models would be clinically helpful in diagnosing patients with variety of neuromuscular diseases whose symptoms are quite similar. Due to the low prevalence of these diseases and their overlapping symptoms, it is currently quite difficult for clinicians to diagnose accurately. Furthermore, objective diagnostic tools for these diseases are generally restricted to biochemical tests, which require human and material resources, and still carry the risk of an inaccurate diagnosis. With our models, neuromuscular diseases could be classified quickly using a computer, thereby giving a time-saving, cost-effective, and accurate diagnosis. Although further

studies are necessary for validation of the genes selected herein, it would be very promising to utilize the selected genes as biomarkers for neuromuscular diseases. One example of a successful accomplishment in identifying biomarkers and applying them in clinics is the oncotype molecular test for breast cancer, which uses a 21-gene molecular signature [19].

The classification models of SVM-OVO, SVM-OVR, and DAGSVM, when used with the BW feature selection method, discriminated neuromuscular disease and control groups with 100% accuracy using only four features. Above, we presented the number of selected genes for each model along with their accuracy, RCI, and kappa index scores. The fewer features are selected, the more efficient the model is; the use of a small number of features represents a great reduction in the high-dimensional microarray data, implying a time-saving and cost-effective analysis. It would be costly to use high-throughput arrays for clinical practice, since synthesizing the necessary polymerase chain reaction (PCR) primers for such a large number of genes increases production costs drastically.

Although the SVM-OVO, SVM-OVR, and DAGSVM methods using BW also yielded 100% accuracy, among them, only SVM-OVR discriminated 12 groups with 100% accuracy of cross-validation for the training set as well as for the testing set. Binary SVMs have been shown to perform well in various areas of biological analysis, and are well suited to working with high-dimensional data, such as microarray data, which have proven to be problematic for many traditional methods. Multicategory SVMs based on binary SVM have also been demonstrated to accurately classify a gene expression data set. In multicategory SVMs, however, when the number of classes increases, the complexity of the overall classifier also increases [15].

Sakellariou et al. investigated the minimum number of genes for classifications of 10 neuromuscular disease groups and one control group using five binary feature selection methods and four binary machine learning methods [4]. They separately categorized each neuromuscular disease group with the control group, showing accuracies of 100% using several binary models. Our study simultaneously classified 11 neuromuscular disease groups and one control group based on multicategory machine learning methods. Even though multiclass classification is much more complex than binary classification, and the classification accuracy may drop dramatically with increasing number of classes [3], the all SVM methods with BW in the present study showed a perfect accuracy of 100% when applied to the testing set.

One of the limitations of our study was that we did not propose any pathogeneses of our selected genes for neuromuscular diseases. Another limitation was the relatively small sample size for each disease; this was because of the very low prevalence of the neuromuscular diseases, and difficulty in collecting their raw DNA microarray data. The direction of our future research will include investigating the pathogenesis of the selected genes by PCR with animal models for neuromuscular diseases, which could validate our selected genes as verified biomarkers.

REFERENCES

[1] G. Piatetsky-Shapiro and P. Tamayo, "Microarray data mining: facing the challenges," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 1-5, Dec. 2003.

[2] S. Saviozzi, G. Iazzetti, E. Caserta, A. Guffanti, and R. A. Calogero, "Microarray data analysis and mining," *Methods Mol. Med.*, vol. 94, pp. 67-90, Dec. 2003.

[3] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics,* vol. 20, no. 15, pp. 2429-2437, Apr. 2004.

[4] A. Sakellariou, D. Sanoudou, and G. Spyrou, "Investigating the minimum required number of genes for the classification of neuromuscular disease microarray data," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 3, pp. 349-355, May 2011.

[5] M. Bakay, Z. Wang, G. Melcon, L. Schiltz, and J. Xuan et al., "Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration," Brain, vol. 129, pp. 996-1013, Feb. 2006.

[6] K. Nagaraju, L. G. Rider, C. Fan, Y. Chen, and M. Mitsak et al., "Endothelial cell activation and neovascularization are prominent in dermatomyositis," *J. Autoimmune Dis.*, vol. 3, no. 2, pp. 1-8, Feb. 2006.

[7] M. Crimi, A. Bordoni, G. Menozzi, L. Riva, and F. Fortunato et al., "Skeletal muscle gene expression profiling in mitochondrial disorders," *FASEB J.,* vol. 19, no. 7, pp. 866-868, Feb. 2005.

[8] I. Eisenberg, N. Novershtern, Z. Itzhaki, M. Becker-Cohen, and M. Sadeh et al., "Mitochondrial processes are impaired in hereditary inclusion body myopathy," *Hum. Mol. Genet.*, vol. 17, no. 23, pp. 3663-3674, Aug. 2008.

[9] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, Aug. 2007.

[10] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics,* vol. 21, no. 5, pp. 631-643, Mar. 2005.

[11] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J. Am. Stat. Assoc.*, vol. 97, no. 457, pp. 77-87, Mar. 2002.

[12] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, and M. Gaasenbeek et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531-537, Oct. 1999.

[13] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGS for multiclass classification," *In Advances in Neural Information Processing Systems 12,* MIT Press, pp. 547–553, 2000.

[14] T. Srivastava, B. T. Darras, J. S. Wu, and S. B. Rutkove, "Machine learning algorithms to classify spinal muscular atrophy subtypes," *Neurology*, vol. 79, no. 4, pp. 358-364, Jul. 2012.

[15] R. Zhang, G. Huang, N. Sundararajan, and P. Saratchandran, "Multicategory classification using an Extreme Learning Machine for microarray gene expression cancer diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 4, no. 3, pp. 485-495, Jan. 2007.

[16] C. J. Feng, Z. Yu, U. Kingit, and M. P. Baig, "Threefold vs. fivefold cross validation in one-hidden-layer and two-hidden-layer predictive neural network modeling of machining surface roughness data," *J. Manuf. Syst.*, vol. 24, no. 2, pp. 93-107, 2005.

[17] C. Hsu, C. Chang, and C. Lin, "A practical guide to vector classification," Dept. Comput. Sci., National Taiwan Univ., Taipei, Taiwan, 2003.

[18] E. Pegoraro, E. P. Hoffman, L. Piva, B. F. Gavassini, and S. Cagnin et al., "SPP1 genotype is a determinant of disease severity in Duchenne muscular dystrophy," *Neurology*, vol. 76, no. 3, pp. 219-226, Jan. 2011.

[19] J. A. Sparano, and S. Paik, "Development of the 21-gene assay and its application in clinical practice and clinical trials," *J. Clin. Oncol.*, vol. 26, no. 5, pp. 721-728, Feb. 2008.