pattern recognition, feature selection,
linear classifier, k-NN rule, pair-wise classifier,
biomarkers, amyotrophic lateral sclerosis

Beata SOKOŁOWSKA[1], Adam JÓŹWIK[2], Irena NIEBROJ-DOBOSZ[3], Piotr JANIK[4]

THE PAIR-WISE LINEAR CLASSIFIER AND THE *K*-NN RULE IN APPLICATION TO ALS PROGRESSION DIFFERENTIATION

The two kinds of classifier based on the *k*-NN rule, the standard and the parallel version, were used for recognition of severity of ALS disease. In case of the second classifier version, feature selection was done separately for each pair of classes. The error rate, estimated by the leave one out method, was used as a criterion as for determination the optimum values of *k*'s as well as for feature selection. All features selected in this manner were used in the standard and in the parallel classifier based on *k*-NN rule.

Furthermore, only for the verification purpose, the linear classifier was applied. For this kind of classifier the error rates were calculated by use the training set also as a testing one. The linear classifier was trained by the error correction algorithm with a modified stop condition.

The data set concerned with the healthy subjects and patients with amyotrophic lateral sclerosis (ALS). The set of several biomarkers such as erythropoietin, matrix metalloproteinases and their tissue inhibitors measured in serum and cerebrospinal fluid (CSF) were treated as features. It was shown that CSF biomarkers were very sensitive for the ALS progress.

1. INTRODUCTION

The different biomarkers/markers are searched for evaluating the progress of ALS and for monitoring the treatment effects [2, 7, 12]. In our studies on ALS we marked erythropoietin (EPO) [3], matrix metalloproteinases (MMPs), such as membrane type matrix metalloproteinase-1 (MT-MMP-1), gelatinases A (MMP-2) and B (MMP-9) and their tissue inhibitors TIMP-1, TIMP-2 [8]. Differences in EPO concentration between the mild and severe ALS cases were not significant [3]. However, combining the EPO with disease duration and patient age and using the pattern recognition methods, it was possible to detect course of ALS [5,6]. Our reports [8,10,11] concerning only MMPs demonstrated that the most useful features were MT-MMP-1, MMP-2 and MMP-9. The main goal of the paper was investigation and evaluation of all these ALS biomarkers in serum/CSF in differentiating disease progression.

2. MATERIAL AND METHODS

2.1. PATIENS AND BIOMARKER MEASUREMENTS

Thirty patients with ALS and fifteen healthy subjects were studied. According to their clinical status, ALS patients were divided into two groups: (a) with mild steady progressing and (b) the severe ALS with rapidly developed symptoms [1]. The serum and lumbar cerebrospinal (CSF) samples were collected during laboratory/diagnostic procedures. The biomarkers such as EPO, MT-MMP-1, MMP-2, MMP-9, TIMP-1, TIMP-2 were determined by immunoassay methods, which are in detail described elsewhere [3,8]. These biomarkers were used as features for evaluation the disease severity. The three

[1] Beata Sokołowska, beta.sokolowska@imdik.pan.pl, Bioinformatics Laboratory, Mossakowski Medical Research Center, Polish Academy of Sciences, 02-106 Warszawa, Pawińskiego 5.
[2] Adam Jóźwik, adam.jozwik@ibib.waw.pl,Department for Mathematical Modelling of Physiological Processes, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, 02-109 Warszawa, Trojdena 4
[3] Niebrój-Dobosz, dobosz@imdik.pan.pl Neuromuscular Unit, Mossakowski Medical Research Center, Polish Academy of Sciences, 02-106 Warszawa, Pawińskiego 5
[4] Piotr Janik, piotr.janik@wum.edu.pl, Department of Neurology, Warsaw Medical University, 02-097 Warszawa, Banacha 1A

groups as classes of the disease progress were defined: healthy subjects (control) - class I, mild ALS patients - class II and severe ALS patients - class III.

2.2. METHODS

The data were analyzed using two kinds of classifiers based on *k*-NN rule. One is the standard *k*-NN classifier and the second one is a classifier composed of two-decision *k*-NN classifiers [4]. Each of the component classifiers corresponds to a different pair of classes. Optimum numbers of nearest neighbors were determined separately for each of the component classifiers, using as criterion the error rates estimated by the leave one out method. The error rate, calculated in the above mentioned manner was also used as a feature selection criterion that was performed separately for each of the component classifiers. The obtained results were additionally verified by the modified pair-wise linear classifier, trained by the well known error correction algorithm [9], applied to the feature sets selected for the considered parallel *k*-NN classifier.

The error correction algorithm ends after a final number of steps only when the sets are linearly separable. As a result weights of a separating hyperplane are found. When the sets are linearly inseparable the algorithm would never stop. Therefore, the number of corrections must be constrained. This algorithm can be very useful also when the sets are not linearly separable. It is enough, after each correction, to calculate a number of incorrectly separated samples and to keep in the computer memory the hyperplane offered a minimum of misclassified samples. The number of steps was constrained each time to thirty millions.

Since the linear classifier concerns only two class problem, to solve the three (or more) class task, one must apply a separate linear classifier for each class pair, i.e. to use a pair-wise linear classifier. In our study the error rates for the pair-wise linear classifiers were estimated using the same training set in the role of the testing one.

Independently of the classifier type, the values of all features were standardized by subtracting their mean values and dividing the outcomes by standard deviations.

3. RESULTS

The lower error rates were obtained for the CSF features as compared with the same features measured in serum. The biomarker TIMP-2 was excluded because its level remained on the normal range [8]. Finally, five features-biomarkers were analyzed in serum and in CSF. The considered features and classes are presented in Table 1.

Table 1. Description of classes and features.

CLASSES	FEATURES
I: Healthy subjects II: Mild ALS III: Severe ALS	1: MT-MMP1 2: MMP-2 3: MMP-9 4: TIMP-1 5: EPO

It is obvious that the costs of measurements are lower in case of serum. The error rates for all three considered classifiers, build for five features measured in serum, are presented in the Table 2. We can see that the same features, but measured in CSF promise lower misclassifications rates as compared to the features measured in serum.

Table 2. Error rates for the three classifiers: standard *k*-NN, parallel *k*-NN and pair-wise linear, for three classes and five features measured independently in serum and cerebrospinal fluid (CSF).

CLASSIFIER	SERUM	CSF
Standard (S)	0.200	0.067
Parallel (P)	0.200	0.089
Linear (L)	0.067	0.022

Feature selection can decrease the error rates. As it was mentioned earlier, in case of the parallel *k*-NN classifier, the features were selected separately for each pair of classes. But for the standard version of *k*-NN classifier feature selection was performed simultaneously for all three classes. The results for the component two-decision classifiers and for the global standard *k*-NN (S), parallel *k*-NN (P) and pair-wise linear classifiers (L) are presented in the Table 3.

Table 3. Error rates for the component and the global three class classifiers for serum features.

Classes	Selected features	After selection	Feature 1	Feature 2	Feature 3
[I, II]	{1,3}	0.000 (3-NN)	0.033	0.033	0.100
[I, III]	{1,3}	0.100 (3-NN)	0.200	0.400	0.233
[II, III]	{2}	0.067 (13-NN)	0.333	0.067	0.200
[I, II, III] S	{1,3}	0.178 (2-NN)	0.400	0.333	0.267
[I, II, III] P	{1,2,3}	0.111	0.356	0.311	0.289
[I, II, III] L	{1,2,3}	0.111	0.356	0.267	0.267

The component *k*-NN classifiers of parallel version exploited the selected features listed in the three upper rows of the second table column, i.e. jointly the features 1, 2 and 3. The standard *k*-NN classifier required only two features 1 and 3. Each component classifier of the pair-wise linear classifier utilized the same three features 1, 2 and 3. Although the parallel *k*-NN classifier and the pair-wise linear classifier offers the same error rate values, it is necessary to take into account that error rate for these two types of classifiers was calculated in a different manner. However, estimation of the misclassification rate by the leave one out method seems to be slightly more reliable. For this reason, in our further consideration, the result obtained the parallel *k*-NN classifier will be treated as the main one. The last three columns of Table 3 contain results obtained for the single features out of selected ones. Table 4 presents more detailed characteristic of the parallel *k*-NN classifier. We can see that all healthy subjects (class I) were correctly classified (Table 4A). In case of the patients with mild ALS (class II) the rate of correct classification was equal 93.3% (Table 4B), so it was also high. The patients with severe ALS form the most difficult class, only 73.3% cases from this class were correctly classified.

Table 4. Confusion matrix (panel A), probabilities a priori (panel B) and probabilities a posteriori (panel C) for selected serum features.

A. Numbers of cases from the class *i* (row) assigned to the class *j* (column)				B. Probabilities that a case from the class *i* (row) will be assigned to the class *j* (column)				C. Probabilities that a case assigned to the class *i* (row) comes in fact from the class *j* (column)			
True class	Assigned class			True class	Assigned class			Assigned class	True class		
	I	II	III		I	II	III		I	II	III
I	15	0	0	I	1.000	0.000	0.000	I	0.789	0.000	0.211
II	0	14	1	II	0.000	0.933	0.067	II	0.000	1.000	0.000
III	4	0	11	III	0.267	0.000	0.733	III	0.000	0.083	0.917

More important than the table of probabilities a priori (Table 4B), from the practical point of view, is the table of probabilities a posteriori (Table 4C). Only 78.9% of cases among those assigned to the class I (healthy one) were correctly classified and 21.1% patients assigned to the class I suffered in fact the severe ALS. The most reliable diagnoses were assignments to the mild ALS disease, i.e. to the class II (100% correct). Taking into account the results contained in Table 3 one can expect significantly lower error rates, after feature selection, if the features would be measured in CSF. The error rates for the component and global classifiers are shown in Table 5.

Table 5. Error rates for component and total classifiers for CSF features.

Classes	Selected features	After selection	Feature 2	Feature 5
[I, II]	{2,5}	0.000 (3-NN)	0.100	0.067
[I, III]	{2,5}	0.000 (3-NN)	0.033	0.033
[II, III]	{2}	0.033 (12-NN)	0.033	0.267
[I, II, III] S	{2,5}	0.022 (3-NN)	0.089	0,244
[I, II, III] P	{2,5}	0.022	0.089	0.222
[I, II, III] L	{2,5}	0.022	0.133	0.311

This time only two features were selected as for standard as well as for parallel version of the *k*-NN classifier. The healthy subjects (class I) were correctly distinguished from the mild (class II) and from the severe ALS group (class III) using only two features 2 and 5. Furthermore, only one feature 2 was selected to differentiate between the patients with mild and severe ALS. All types of the considered classifiers offered the same value 2.2% of the error rate. Similarly, as it took place in case of features measured in serum, the more detailed characteristic was determined. It is given below in Table 6. This time only one case with severe ALS was misclassified as the patient suffering mild ALS (Table 6A). All cases from the classes I and II were correctly classified (Table 6A and 6B). The rate of correct classification, in case of the class III, was equal to 93.3% (Table 6B). As implies from Table 6C, classifications were very reliable when the case was assigned to the class I or to the class III, i.e. diagnosed as the healthy subject or as the patient suffering the severe ALS, so the extreme classes were easier for the diagnosis.

Table 6. Confusion matrix (panel A), probabilities a priori (panel B) and probabilities a posteriori (panel C) for selected CSF features.

A. Numbers of cases from the class *i* (row) assigned to the class *j* (column)

True class	Assigned class		
	I	II	III
I	15	0	0
II	0	15	0
III	0	1	14

B. Probabilities that a case from the class *i* (row) will be assigned to the class *j* (column)

True class	Assigned class		
	I	II	III
I	1.000	0.000	0.000
II	0.000	1.000	0.000
III	0.000	0.067	0.933

C. Probabilities that a case assigned to the class *i* (row) comes in fact from the class *j* (column)

Assigned class	True class		
	I	II	III
I	1.000	0.000	0.000
II	0.000	0.937	0.063
III	0.000	0.000	1.000

Less confident were assignments to the class II, the mild one, since among cases classified to this class, one case, i.e. 6.3% (Table 6C) were misclassified. The correct classification rate equaled to 93.7%.

4. FINAL REMARKS

The concentrations of the matrix metalloproteinases (as MT-MMP-1, MMP-2, MMP-9) form the set of the best features among those measured in serum, for ALS course evaluation. There were selected from among five features, i.e. the omitted two features TIMP-1 and EPO were useless. The parallel *k*-NN classifier or the pair-wise linear classifier operating with three above mentioned features can be suitable for ALS progress evaluation.

If the measurements were performed in the cerebrospinal fluid then different features were selected than it took place in case of serum. Only two features, MMP-2 and EPO, were chosen in this case. This time all three considered classifiers were nearly equivalent since the error rate was the same.

Comparing the classifiers based on serum features with the classifiers utilized features measured in the cerebrospinal fluid, one can noticed, looking in Tables 4 and 6, that CSF features are more significant for evaluation the progression of the ALS disease. They offer five time lower error rate, i.e. 2.2% versus 11.1% expected for the serum features. The misclassifications appeared in case of the CSF features are less danger since only one severe case (6.7%) was classified as the mild one. In case of the serum features 26.7% of severe cases were misclassified to the healthy subjects, what seems to be much more danger. The remaining misclassifications (6.7%) were less danger since they concern classification the mild cases as the severe ones.

The presented results have shown that CSF biomarkers are very sensitive for the ALS progress.

5. ACKNOWLEDGMENTS

In this study, the infrastructure and services developed by the Biocentrum-Ochota project (POIG.02.03.00-00-003/09) were used.

BIBLIOGRAPHY

[1] BROOKS B.R., MILLER R.G., SWASH M., MUNSAT T.L., El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis, Amyotrophic Lateral Sclerosis and other Motor Neuron Disorder, 2000, Vol. 1, pp. 293-299.

[2] CUDKOWICZ M., QURESHI M., SHEFNER J., Measures and Markers in Amyotrophic Lateral Sclerosis, NeuroRx, 2004, Vol. 1, No. 2, pp. 273-283.

[3] JANIK P., KWIECINSKI H., SOKOLOWSKA B., NIEBROJ-DOBOSZ I., Erythropoietin concentration in serum and cerebrospinal fluid of patients with amyotrophic lateral sclerosis, Journal of Neural Transmission, 2010, Vol. 117, pp. 343–347.

[4] JÓŹWIK A., SERPICO S., ROLI F., A parallel network of modified 1-NN and *k*-NN classifiers -application to remote-sensing image classification, Pattern Recognition Letters, 1998, Vol. 19, pp. 57-62.

[5] JÓŹWIK A., SOKOŁOWSKA B., NIEBOJ-DOBOSZ I., JANIK P., KWIECIŃSKI H., Extraction of biomedical traits for patients with amyotrophic lateral sclerosis using parallel and hierarchical classifiers, International Journal of Biometrics, 2011, Vol. 3, No. 1, pp. 85-94.

[6] JÓŹWIK A., SOKOŁOWSKA B., NIEBROJ-DOBOSZ I., JANIK P., KWIECIŃSKI H., Pattern recognition approach to differentiation of disease severity in patients with amyotrophic lateral sclerosis, Journal of Medical Informatics & Technologies, 2008, Vol. 12, pp. 143-147.

[7] KUDO L.C., PARFENOVA L., VI N., LAU K., POMAKIAN J., VALDMANIS P., ROULEAU G.A., VINTERS H.V., WIEDAU-PAZOS M., KARSTEN S.L., Integrative gene-tissue microarray-based approach for identification of human disease biomarkers: Application to amyotrophic lateral sclerosis, Human Molecular Genetics, 2010, Vol. 19, No. 16, pp. 3233-3253.

[8] NIEBROJ-DOBOSZ I., JANIK P., SOKOŁOWSKA B., KWIECIŃSKI H., Matrix metalloproteinases and their tissue inhibitors in serum and cerebrospinal fluid of patients with amyotrophic lateral sclerosis, European Journal of Neurology, 2010, Vol. 17, pp. 226–231.

[9] NILSSON N., Learning machines, McGraw-Hill, New York, 1965.

[10] SOKOŁOWSKA B., JOZWIK A., NIEBROJ-DOBOSZ I., JANIK P., KWIECIŃSKI H., Evaluation of matrix metalloproteinases in serum of patients with amyotrophic lateral sclerosis with pattern recognition methods, Journal of Physiology and Pharmacology, 2009, Vol. 60, Suppl. 5, pp. 117-120.

[11] SOKOŁOWSKA B., JOZWIK A., NIEBROJ-DOBOSZ I., JANIK P., KWIECIŃSKI H., Analysis of matrix metalloproteinases (MMPs) in cerebrospinal fluid of patients with amyotrophic lateral sclerosis (ALS), Journal of Medical Informatics & Technologies, 2009, Vol. 13, pp. 147-150.

[12] STRONG M.J., The basic aspects of therapeutics in amyotrophic lateral sclerosis, Pharmacology &Therapeutics, 2003, Vol. 98, pp. 379–414.