arterial blood gasometry, paralysis of diaphragm,
k-NN rule, classifiers, k means algorithm

Beata SOKOŁOWSKA[*], Adam JÓŹWIK[**,***]

RECOGNITION OF PATHOLOGICAL STATES IN ARTERIAL BLOOD BY DISTANCE BASED TECHNIQUES

The paper presents the application of some distance based pattern recognition algorithms for recognition of pathological states in respiratory system on the basis of the arterial blood gasometry (features pH, pCO_2, pO_2). In our biological model two experimental situations were considered: 1) the intact animals and 2) the main inspiratory muscles paralyzed (after acute of bilateral phrenicotomy). The comparison of the mentioned three features in the two conditions was the main goal of the present study. The analyzed biological data set contained 38 in class 1 (muscle function preserved) and 36 in class 2 (after diaphragm paralyzed) measurements. It was discovered that a significant part of the measurements could be correctly recognized as the ones coming from the first or the second class according to gasometric measurements.

1. INTRODUCTION

Surgery procedures may affected respiratory muscles function. Diaphragmatic dysfunction due to phrenic nerve damage is a well recognized complication of heart surgery. The incidence of diaphragm dysfunction in patients undergoing cardiac surgery varies from 10% to 85% according to the electrophysiological, radiological or other methods used for its detection, and is similar in adults, children and infants [1]. The aim of this paper is to estimate the influence of paralysis diaphragm on arterial blood gasometry. In experimental model of diaphragm paralysis performed on 19 anaesthetized cats were subjected to C5-C6 bilateral phrenicotomy [2, 3]. In the paper [2] we took into consideration only ventilation parameters for the evaluation breathing before and after bilateral paralysis of diaphragm. In this study only the features of arterial blood were measured: pH (feature no. 1), carbon dioxide pressure pCO_2 (feature no. 2) and oxygen pressure pO_2 (feature no. 3). Measurements of the blood samples were repeated two times for each of the animals before (class 1) and 1 hour after bilateral phrenicotomy (class 2). Our data set contained 74 of 3-dimensional measurement vectors. There were 38 measurements from the class 1 and 36 form the class 2 (one cat was rejected). The analysis of the arterial blood gas features has been a subject of our work [3], which concerned experimental conditions before and after application of chemical stimuli: hypercapnia and hypoxia.

[*] Medical Research Center PAS, Department of Neurophysiology, Pawińskiego 5, 02-106 Warsaw
[**] Technical University of Lodz, Computer Engineering Department, Al. Politechniki 11, 90-924 Lodz
[***] Institute Biocybernetics and Biomedical Engineering PAS, Trojdena 4, 02-109 Warsaw

To analyze the data we applied the well known techniques of statistical pattern recognition like the k nearest neighbor rule (k-NN) [4], k means clustering algorithm (k-MCA) [5] and the leave one out method [6].

2. PATTERN RECOGNITION METHODS

The statistical pattern recognition deals with two main kinds of methods: the supervised and the nonsupervised ones. The first one requires a set of objects with known class-membership called the training set. It is used for construction of the classification rule. The second one does not require the class-membership information. We deal then with the set of objects without being told their categories. The set of objects is divided into desired number of subsets (clusters), i.e. "clouds" of points. The goal of clustering is to distinct subgroups in such a way that any two objects within the same group are more similar than any two objects coming from the different groups. All considered objects are described by a set of features and may be treated as points or vectors in the n-dimensional feature space. In further considerations these points or vectors will be identified with the objects. All applied in the paper pattern recognition methods are based on the distance function. We decided to use the city distance measure.

2.1. THE K-NN RULE

The k-NN rule assigns the classified object, i.e. the point in the feature space, to the same class as the majority of its k "nearest neighbors" in the reference set (training set), in the sense of the assumed distance measure. Information contained in the reference set R is used for the determination of k in such a way that the probability of misclassification (error rate) reaches minimum. The probability of misclassification can be estimated by the leave one out method described below in the next section. The error rates e[k] are calculated with the use of the leave one out method for k=1,2,...,m, where m is number of objects in the training set. Next, the value of k that corresponds to the smallest error rate is chosen as the optimum one.

Attention: The training set X is the set used for the classifier construction. The reference set R is set that must be stored during the classification phase. It may be equal to the training set or it can be a subset (reduced set) of the training set or another set with lower size obtained from the training set (condensed set). In the considered problem we have assumed that X=R since we do not worry about the classification speed.

2.2. THE LEAVE ONE OUT METHOD

The leave one out method serves for experimental estimation of the misclassification probability on the basis of the training set X. Each object **x** of the training set is classified by the k-NN rule with the reference set that is decreased by the currently classified object, i.e. the object **x** is classified by the k-NN classifier with the set R-{**x**} as the reference set. The error rate er=r/m estimates the probability of misclassification, where r means the number of misclassified objects and m is the number of points in the set R. We do not need to dispose of the testing set in such an approach. For this reason the whole data set X can be used as the reference set, i.e. R=X.

2.3. THE K-MEANS ALGORITHM

To discover k clusters in the given set of object, i.e. data set of points in the feature space, it is necessary to initialize k start points $\mathbf{mv_1}$, $\mathbf{mv_2}$,…, $\mathbf{mv_k}$, each of one represents the separate cluster. Then all points **x** from the data set are assigned to the cluster that corresponds to the nearest representative. The points $\mathbf{mv}_i$, i=1,2,..k, are substituted by the cluster gravity centers and again all points **x** are classified to the cluster with the nearest representative. This procedure is stopped when no change in all $\mathbf{mv}_i$, i=1,2,..k, will be observed. Instead of controlling the component values of $\mathbf{mv}_i$, i=1,2,..k, we can control the contents of the distinguished data subsets.

2.4. THE TWO STAGE CLASSIFIER

This type classifier makes up the decision in 2 stages. In each stage it can assign the correct or wrong class number, refuse the decision or send the classified object to the second stage. The second stage can operate according to the different rule than the one used in the first stage. We find in the feature space certain regions A_i, i=1,2,…,nc, each of one contains only the points from the one class i. If the new point falls only into the area A_i then it is classified to the class i. The points from the class overlap area are clustered into nc subgroups. If the classified point falls into the class overlap area then the class i most heavily represented within the corresponding cluster is assigned.

To construct this classifier it is necessary to determine class areas A_i, i=1,2,…,nc. For each class i, i=1,2,…,nc, a separate area A_i is defined according to two given below formulas:

$$e_i = \max_{x_j \in X_i} d(\{X_i - x_j\},\{x_j\}), \qquad (1)$$

$$A_i = \{x: d(X_i,\{x\}) \le e_i\}, \qquad (2)$$

where X_i is a subset of the training set which corresponds to the class i and d denotes a distance between two sets. As a distance between two sets the city measure between two nearest points each of them belonging to a different set is taken. The symbol $\{x_j\}$ is a set containing a single x_j , similarly {x} denotes a single vector x. It is very easy to establish which of the A_i, i=1,2,…,nc, contain the classified object.

3. COMPUTATIONAL RESULTS

We start our computations with the standard k-NN rule. The use of all features offered 23.0% of misclassifications estimated by the leave one out method. Since we deal only with 3 features there is no problem to review all possible k-NN rules (k=1,2,..,74) and all feature combinations. The minimum error rate 18.9% was reached for k=1 and for the 2 features. The results for the k-NN rule are shown in the Tab. 1.

Tab.1. The confusion matrices for the k-NN rule

Features: pH and pCO_2, 1-NN rule is the optimum k-NN rule						
Error rate: 18.9%	Number of points from the class i (row) assigned to the class j (column)		Probability that point from the class i (row) will be assigned to the class j (column)		Probability that point assigned to the class i (row) is in fact from the class j (column)	
True class ↓	**Assigned class**		**Assigned class**		**Assigned class**	
	1	**2**	**1**	**2**	**1**	**2**
1	30	8	79.0%	21.0%	83.3%	16.7%
2	6	30	16.7%	83.3%	21.0%	79.0%

The Tab. 1 indicates that 8 (21%) out of 38 points of the class 1 and 6 (16.7%) out of 36 points of the class 2 were misclassified during the realization of the leave one out method. Thus, only 79.0% and 83.3% from the class 1 and the class 2 respectively were correctly classified. The decisions, which indicate the class 1 are slightly more confident (83.3%) than the assignments to the class 2 (79.0%). The most strange phenomenon is that 21% of the measurements concerned to the class 1, i.e. the control one, were classified as the pathological ones (class 2). It is difficult to agree with such a conclusion. Since only two features were selected, it is possible to illustrate the data.

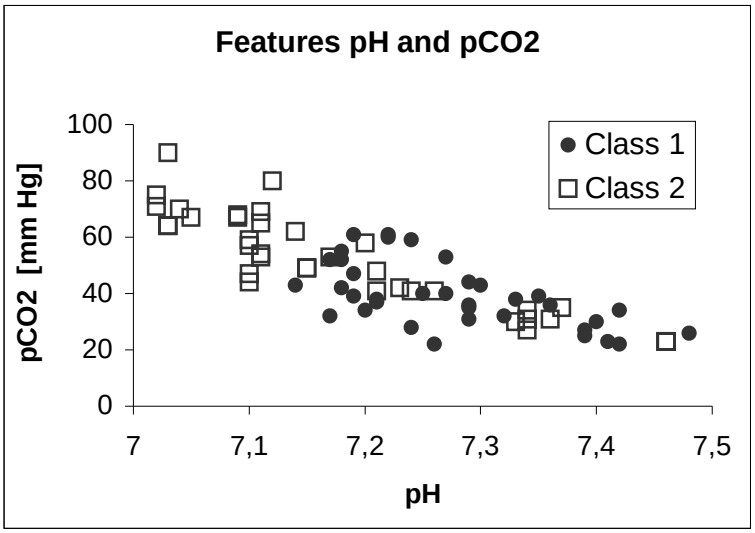

Fig.1. The data set projection on the feature space defined by pH and pCO_2

We can see that a pretty large number of points lie in the area free of points from the class 1. So, it is worth to find the areas A_i, i=1,2, of each class. The area A_1-A_2 will contain only the points from the class 1 and the area A_2-A_1 will contain only the points of the class 2. To maximize the number of points of the training set in the area $(A_1\text{-}A_2)\cup(A_2\text{-}A_1)$ we can review all possible feature combinations, as it took place in the case of the k-NN rule. The point **x** from the area $(A_1\text{-}A_2)\cup(A_2\text{-}A_1)$ is assigned to the class 1 if **x** is in A_1 and to the class 2 if it falls in A_2.

Tab.2. Number of points lying in the areas associated only with one class for the complete and selected feature combination

Feature combination	**Content of the area** A_1-A_2	**Content of the area** A_2-A_1
pH, pCO_2, pO_2	2 points	12 points
pCO_2, pO_2	4 points	17 points

This time the results are very promising, since 17 points out of 36 from the class 2 can be very confidently assigned to the proper class. The situation concerned to the class 1 is less optimistic and there are some doubts whether the area A_1-A_2 should be taken into account in the proposed two stage classifier. To our opinion the area A_1-A_2 contains not enough points from the data set to be considered separately. Furthermore, we have discovered that the area A_1-A_2 consists of two small disjunctive areas with the 2 points in each of them.

Also in this case only two features were selected, so we can again refer to the graphic illustration, see Fig. 2. Now we can deal with the construction of the second stage of the classifier.

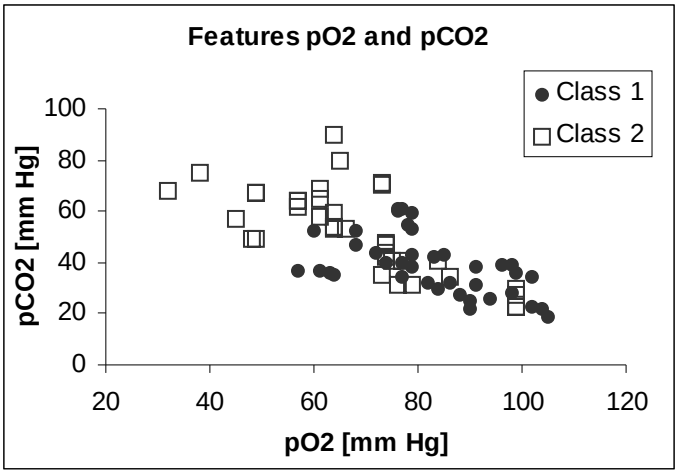

Fig.2. The data set projection on the feature space defined by pO_2 and pCO_2

Application of the k means clustering algorithm, with k=2, leads to the two clusters each of one contains points from the both classes. These clusters, jointly with the area A_2-A_1 as the third one, cover the whole our data set. In the space of the three features and in the space of any two or one single feature, the cluster 2 is positioned in the middle between the cluster 1 and the cluster 3 formed from the points lying in the A_2-A_1.

Tab.3. Results of the cluster analysis for the data decreased by points from the area A_2-A_1.

Cluster number →	**Cluster 1**		**Cluster 2**		**Cluster 3** (A_2-A_1)	
Class number →	**Class 1**	**Class 2**	**Class 1**	**Class 2**	**Class 1**	**Class 2**
Number of points	18	8	20	11	0	17
Class contributions	69.2%	29.8%	64.5%	35.5%	0.0%	100.0%
Mean vectors	[7.4, 30.1, 92.0]		[7.2, 46.6, 72.2]		[7.1, 66.2, 55.4]	
Fuzzy decisions	[0.692, 0.298]		[0.645, 0.355]		[0.000, 1.000]	

Now, the proposed two stage distance based classifier can be fully defined and it operates according to the following formula:

1. *If the classified point falls into the cluster 3 then the class 2 is assigned;*
2. *If the classified point falls into cluster 2 then the fuzzy decision [0.645, 0.355] is assigned, i.e. in our case we assume that this point belongs to the classes 1 and 2 with the probabilities indicated in the decision vector (with the probabilities of 0.645 and 0.355 to the class 1 and the class 2 respectively);*
3. *If the classified point fall into the cluster 1 then the fuzzy decision [0.692, 0.298] is assigned;*
4. *If the classified point falls outside of each of the three clusters then the decision is refused.*

The fuzzy decisions can be transformed into the crisp (nonfuzzy) ones. Thus, 8 points from the cluster 1 and 11 points from the cluster 2 were misclassified to the class 1 (see Tab. 3), i.e. the error rate was equal (8+11)/74=25.7%.

4. CONCLUSIONS

The standard classifiers, as it was noticed previously, are unusable for the considered problem. Application of the k-NN classifier based on the same selected feature set offers high error rate (18.9%) and none of its decisions is sufficiently confident. In our case the most important are decisions concerning to the class 2 (after paralysis of diaphragm). The proposed classifier can extract the pathological area in the feature space that contains approximately 17/74=23% of measurements. This part of the measurements will be classified very confidently. The fact that the (74-17)/74=77% measurements are very difficult to be recognized indicates the significant early respiratory compensation processes in some paralyzed animals. When the diaphragm is paralyzed then breathing assumes a new pattern that is underlain by the action of respiratory accessory and chest wall muscles. Thus, the arterial blood features were not too sensitive to recognize whether we deal with the intact or paralyzed respiratory muscle in the rest breathing.

However, the proposed classifier can assign to the part of measurements very confident decisions, i.e. the 23% of the measurements will be recognized correctly with the probability near 100%. In the case of the standard k-NN rule, the correct decision would be assigned with the probability of 100%-18.9%=81.1% for each classified measurement. If we combine the proposed approach with the k-NN rule then 77% of measurements will be correctly recognized with the mentioned probability of 81.1% and 23% of measurements with the probability close to 100%.

The fact that 77% of measurements, 51.3% from the class 1 (intact animals) and 25.7% from the class 2 (inspiratory muscles paralyzed) cannot be surely recognized denotes that 25.7% of the animals were able to compensate the diaphragm paralysis. The ventilation features used in [2] make possible good recognizing the paralysis of this muscle (99% of correct classifications), while the arterial blood features enable evaluation whether the consequences of this paralysis are serious or not.

BIBLIOGRAPHY

[1] SIAFAKAS N.M., MITROUSKA I., BOUROS D., GEORGOPOULOS D., Surgery and the respiratory muscles, Thorax 54, pp.458-465, 1999

[2] JÓŹWIK A., SOKOŁOWSKA B., BUDZIŃSKA K., An example of computer aided decision-making system for recognition of respiratory pathology, MIT vol. 2/2001: MI41-48

[3] SOKOŁOWSKA B., JÓŹWIK A., BUDZIŃSKA K., Analiza prężności gazów oddechowych krwi tętniczej z zastosowaniem algorytmu rozpoznawania obrazów, Materiały konf. Techniki Informatyczne w Medycynie, str. BP21-Bp33, Ustroń, 1999 (in Polish)

[4] FIX E., HODGES J. L., Discriminatory Analysis: Nonparametric Discrimination Small Sample Performance, project 21-49-004, Report Number 11, pp. 280-322,USAF School of Aviation Medicine, Randolph Field, Texas, 1952

[5] TOU J. T., GONZALES R. C., Pattern recognition principles, Addison-Wesley Publishing Company, London, 1974

[6] VAPNIK W. N., TCHERVONENKIS A. YA., Teoria rozpoznawania obrazow, Moskwa, 1974 (in Russian)