ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

Małgorzata Misztal[*], *Maciej Banach*[**]

ON DISTANCE-BASED ALGORITHMS IN MEDICAL APPLICATIONS

Abstract. Logistic regression is the most popular method used to classify patients into 2 selected subgroups in medical research. Distance-based algorithms, such as nearest neighbor algorithm, simple and intuitive, are rarely used in practice.

In the study some selected distance-based algorithms (NN, k-NN, DB and k-NN Tree) were applied to predict atrial fibrillation (AF) incidents among 300 patients with aortic valve defects, who underwent aortic valve replacement.

Key words: medical research, logistic regression, distance-based algorithms, atrial fibrillation.

I. INTRODUCTION

Let us consider a learning set U={($\mathbf{x}_1$, y_1), …, ($\mathbf{x}_N$, y_N)}, where **x** is the vector of independent variables $\mathbf{x}=[x_1, x_2, ..., x_p]^T$ and y is the response (dependent) variable. In our study, y is nominal variable, describing the number of the class the object belongs to.

Usually, in medical research, y is binary variable identifying one of the two classes the patient belongs to: e.g. good outcome or death, low risk subgroup or high risk subgroup, etc. The aim of the research is then to classify patients to the selected subgroups. The most popular method of classification in medical research is logistic regression. Other algorithms, such as distance-based ones are rarely used in practice although they are very simple and intuitive.

The objective of the study was to compare the prediction accuracy of some distance–based classifiers and logistic regression model used to classify patients with aortic valve defects, who underwent aortic valve replacement.

[*] Ph.D., Chair of Statistical Methods, University of Łódź.

[**] MD, Ph.D., Department of Cardiology, 1st Chair of Cardiology and Cardiac Surgery, Medical University of Łódź.

II. MATERIAL AND METHODS

The set of 300 case records of patients undergoing aortic valve replacement (AVR) due to aortic valve defect was analysed (Banach *et* al. 2006).

The most common complication following cardiac surgery is atrial fibrillation (AF) so the dependent variable y was the binary variable with two possible values: 0 for non – AF patients and 1 for AF patients.

Only preoperative risk factors were taken into account:

- BMI – body mass index [kg/m^2];
- Sex (Male, Female);
- Age [in years];
- EF – left ventricular ejection fraction [in %];
- LVESd - left ventricular end systolic dimension [in cm];
- LVEDd - left ventricular end diastolic dimension [in cm];
- ESIVST - end-systolic intraventricular septum thickness [in cm];
- EDIVST - end-diastolic intraventricular septum thickness [in cm];
- LAd - left atrium dimension [in cm].

The set of 300 patients was randomly divided into the learning sample (150 cases) and the test sample (150 cases).

The following classifiers $\psi(\mathbf{x})$ were used:

- Logistic regression model;
- The Nearest Neighbour Algorithm (NN – Tadeusiewicz, Flasiński 1991) - that classifies the unknown object **x** by calculating the distances between the object and all objects in the learning set, and assigning it to the class that the nearest learning object belongs to. So:

$$\psi^{\mathrm{NN}}(\mathbf{x}) = i \text{ if } d(\mathbf{x};\mathbf{x}_{i,l_i}) = \min_{g\in K} d(\mathbf{x};\mathbf{x}_{g,l_g}) \qquad (1)$$

$$i\in \mathrm{K},\ \mathrm{K}\in\{0,1\};\ \ l_i = 1,\dots,N_i \ \ l_g = 1,\dots,N_g$$

where $d(\mathbf{x}_\mathrm{m}; \mathbf{x}_\mathrm{n})$ is a distance measure between two objects.

- The k-Nearest Neighbours Algorithm (k-NN – Lacrose 2006, Kurzyński 1997) - that classifies the unknown object **x** by assigning it to the class that is most common among its k nearest neighbours:

$$\psi^{\mathrm{k\text{-}NN}}(\mathbf{x}) = i \ \text{ if } \ \mathrm{k}_i = \max_{g\in K} \mathrm{k}_g \ \ i\in \mathrm{K},\ \mathrm{K}\in\{0,1\} \qquad (2)$$

- The Distance - Based Algorithm (DB – Cuadras 1989) - that classifies the unknown object **x** to the class scoring the lowest value among the k classifying functions:

$$\psi^{\mathrm{DB}}(\mathbf{x}) = i \text{ if } {}^{DB}D_i(\mathbf{x}) = \min_{g \in \mathrm{K}}\{{}^{DB}D_g(\mathbf{x})\} \quad i \in \mathrm{K} \tag{3}$$

where:

$${}^{DB}D_i(\mathbf{x}) = \frac{1}{N_i}\sum_{m=1}^{N_i} d(\mathbf{x};\mathbf{x}_m) - \frac{1}{2N_i^2}\sum_{m=1}^{N_i}\sum_{n=1}^{N_i} d(\mathbf{x}_m;\mathbf{x}_n) \quad i \in \mathrm{K}, \quad \mathrm{K} \in \{0, 1\} \tag{4}$$

and $d(*)$ is a distance measure.

- The k-NN Tree Algorithm (Buttrey & Karo 2002) – that is a combination of classification tree and k-NN algorithm. In the first step the feature space is divided into homogenous subspaces by classification tree, and in the second step – the test set objects are classified using the k-NN rule just among those training objects in the same leaf of the tree as the test object is.

The following distance measures were used:

- The Euclidean distance measure: $d(\mathbf{x}_m;\mathbf{x}_n) = \left[\sum_{r=1}^{p}|x_{mr} - x_{nr}|^2\right]^{\frac{1}{2}}$;
- The Manhattan distance measure: $d(\mathbf{x}_m;\mathbf{x}_n) = \sum_{r=1}^{p}|x_{mr} - x_{nr}|$;
- The Canberra distance measure: $d(\mathbf{x}_m;\mathbf{x}_n) = \sum_{r=1}^{p}\frac{|x_{mr} - x_{nr}|}{|x_{mr} + x_{nr}|}$;
- The Chebyshev distance measure: $d(\mathbf{x}_m;\mathbf{x}_n) = \max|x_{mr} - x_{nr}|$.

All the analyses were performed with STATISTICA PL Software ver. 7.0 and the R environment.

III. RESULTS

The most important aim of any classifier is that it should make accurate predictions for novel cases. For binary outcome the results can be summarized in a confusion matrix (Table 1).

In medical research a confusion matrix is used to calculate some accuracy measures (Table 2).

The most popular accuracy measures in medical research are sensitivity, specificity, positive predictive power and negative predictive power.

Sensitivity is the proportion of true positives (AF patients) that are correctly identified by the classifier. Specificity is the proportion of true negatives (non-AF patients) that are correctly identified by the classifier.

Table 1. A confusion matrix

Predicted class	**Actual class**		**Total**
	AF	**non-AF**	
AF	Correct True positive **a**	Incorrect False positive **b**	a+b
non-AF	Incorrect False negative **c**	Correct True negative **d**	c+d
total	a +c	b+d	N=a+b+c+d

Source: own elaboration.

Positive predictive power is the proportion of patients with positive test results (classifier predicts AF group) who are correctly recognized. Negative predictive power is the proportion of patients with negative test results (classifier predicts non-AF group) who are correctly recognized.

Table 2. Confusion matrix-derived accuracy measures

Measure	Calculation
Correct classification rate	$(a+d)/N$
Misclassification cost	$(b+c)/N$
Sensitivity	$a/(a+c)$
Positive predictive power	$a/(a+b)$
Specificity	$d/(b+d)$
Negative predictive power	$d/(c+d)$
False – positive rate	$b/(b+d)$
False – negative rate	$c/(a+c)$
Kappa statistics	$\frac{(a+d)-\{[(a+c)(a+b)+(b+d)(c+d)]/N\}}{N-\{[(a+c)(a+b)+(b+d)(c+d)]/N\}}$

Source: Fielding 2007.

The best results of application of mentioned algorithms are summarized in Table 3.

The classification tree obtained in the first step of the k-NN Tree algorithm is shown in Figure 1. The tree has 4 leaves. The NN algorithm with Euclidean distance measure was employed in every terminal node.

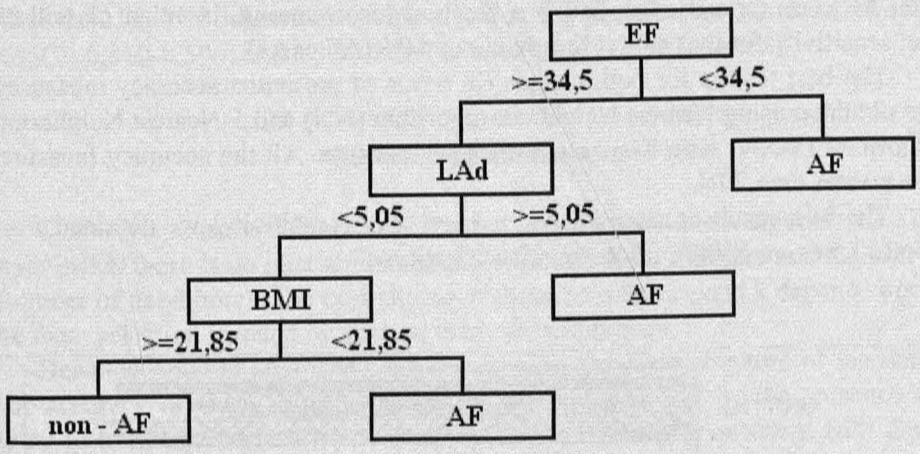

Fig. 1. Classification tree for patients undergoing AVR

Table 3. Comparison of classifiers on the basis of the test sample

Accuracy measure	Classifier						
	logistic regression	NN (Euclidean distance)	NN (Manhattan distance)	7-NN (Euclidean distance)	5-NN (Manhattan distance)	DB algorithm (Chebyshev distance)	k-NN Tree (NN with Euclidean distance in the leaves) – Fig. 1.
Correct classification rate (%)	60,67	65,33	72,67	68,67	74,00	63,33	72,00
Misclassification cost (%)	39,33	34,67	27,33	31,33	26,00	36,67	28,00
Sensitivity (%)	43,84	65,75	72,60	69,86	76,71	57,53	80,82
Positive predictive power (%)	64,00	64,00	71,62	67,11	71,79	63,64	67,82
Specificity (%)	76,62	64,94	72,73	67,53	71,43	68,83	63,64
Negative predictive power (%)	59,00	66,67	73,68	70,27	76,39	63,10	77,78
False - positive rate (%)	23,38	35,06	27,27	32,47	28,57	31,17	36,36
False - negative rate (%)	56,16	34,25	27,40	30,14	23,29	42,47	19,18
Kappa statistics	0,21	0,31	0,45	0,37	0,48	0,26	0,44

Source: authors' calculations.

The selected accuracy measures for logistic regression model and distance-based algorithms are compared in Figure 2.

Logistic regression model classified correctly 76,62% of test objects from non-AF class (specificity), and it is the best result among the other classifiers, but sensitivity for that model is only about 44% (AF cases).

The best results for both classes (in terms of presented accuracy measures) we obtained using Nearest Neighbour algorithm (NN) and 5-Nearest Neighbours algorithm (5-NN) with Manhattan distance measure. All the accuracy measures are greater than 70%.

The best result concerning AF patients only (sensitivity) we obtained from k-NN Tree – over 80% of correct classifications.

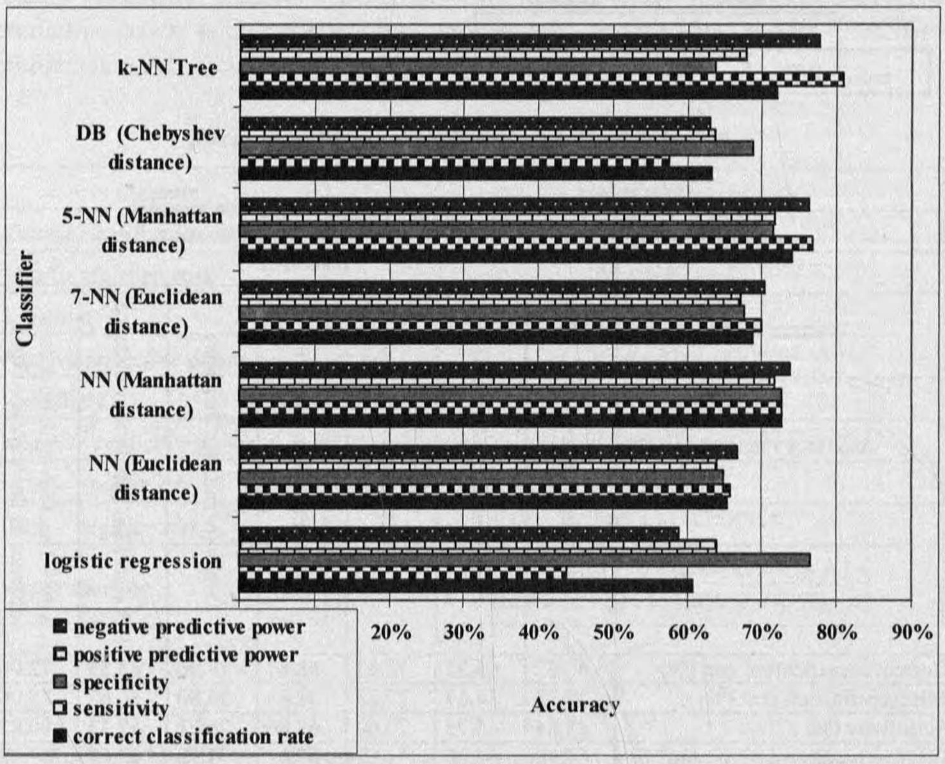

Fig. 2. Prediction accuracy for logistic regression and distance-based algorithms

Generally, for almost all classifiers we obtained better results than for logistic regression model.

According to kappa statistics (an index which compares the agreement against that which might be expected by chance) the best results were gained from 5-NN algorithm with Manhattan distance measure (κ=0,481; 95%CI: 0,341-0,621), NN algorithm with Manhattan distance measure (κ=0,453; 95%CI: 0,310-0,596) and k-NN Tree (κ=0,442; 95%CI: 0,302-0,583).

IV. CONCLUSIONS

There are some problems connected with distance-based algorithms. The main is that there is no instruction about the best distance measure and the best k (number of neighbors). The best choice of distance measure and k depends upon the data; selection is made by various heuristic techniques.

Distance – based algorithms are sensitive to the local structure of the data. The accuracy of these algorithms can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance.

Distance – based methods are examples of instance – based learning – training data set is stored so that classification for a new unclassified object may by found simply by comparing it to the most similar records in the training set (Lacrose 2006). But for instance-based learning algorithms it is important to have a rich database full of as many different combinations of attribute values as possible.

It is also important to represent rare classifications sufficiently, so that the algorithm does not only predict common classifications. That's why the data set should be balanced (e.g. by reducing the proportion of the cases with more common classifications. For more details see: Lacrose 2006).

Almost all classifiers achieved better results than logistic regression model. Distance – based algorithms are simple and intuitive. They can be recommended when we are interested in accurate prediction rather than providing insight into data.

REFERENCES

Banach M., Rysz J., Drożdż J., Okoński P., Misztal M., Barylski M., Irzmański R., Zasłonka J. (2006), Risk Factors of Atrial Fibrillation Following Coronary Artery Bypass Grafting. A Preliminary Report, *Circulation Journal* 2006; 70: 438 – 441.

Banach M., Goch A., Misztal M., Rysz J., Jaszewski R., Goch J. H., (2007), Predictors of Paroxymal Atrial Fibrillation in Patients Undergoing Aortic Valve Repalcement, *The Journal of Thoracic and Cardiovascular Surgery* (in press).

Buttrey S. E., Karo C. (2002), Using k-nearest-neighbor classification in the leaves of a tree, *Computational Statistics & Data Analysis* 40 (2002), 27-37.

Cuadras C. M. (1989), *Distance Analysis in Discrimination and Classification Using Both Continuous and Categorical Variables*, (in:) *Statistical Data Analysis and Inference*, (Dodge ed.), Elsevier Science Publishers B. V., North Holland, 459 - 473.

Fielding A. H. (2007), *Cluster and Classification Techniques for the Biosciences*, Cambridge University Press, Cambridge.

Kurzyński M. (1997), *Rozpoznawanie obiektów. Metody statystyczne*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

Larose D. T. (2006), *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, PWN, Warszawa.

Tadeusiewicz R., Flasiński M. (1991), *Rozpoznawanie obrazów*, PWN, Warszawa.

Małgorzata Misztal, Maciej Banach

O ALGORYTMACH MINIMALNOODLEGŁOŚCIOWYCH W ZASTOSOWANIACH MEDYCZNYCH

W badaniach medycznych do przewidywania przynależności pacjentów do jednej z wyróżnionych dwóch klas zwykle wykorzystuje się model regresji logistycznej. Algorytmy minimalnoodległościowe, takie jak np. algorytm najbliższego sąsiada, mimo ich prostoty i intuicyjnej interpretacji, są wykorzystywane bardzo rzadko.

W referacie podjęto próbę zastosowania algorytmów opartych na odległościach (NN, k-NN, DB oraz k-NN Tree) do prognozowania wystąpienia migotania przedsionków wśród 300 pacjentów po zabiegu wymiany zastawki aortalnej.