A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 235, 2010

Marcin Pełka[*]

K-NEAREST NEIGHBOUR CLASSIFICATION FOR SYMBOLIC DATA

Abstract. The well-known kNN (*k* Nearest Neighbours) rule was proposed by Fix E. and Hodges J. L. [1951] and it is one of the best classifiers for classical data. In the most simple way, the *k*-nearest neighbours assign a classified object to a class that is mostly represented by its *k* nearest neighbours. If in the same distance as *k*-th neighbour are other objects, they also take part in voting. This paper presents an adaptation of KNN for symbolic data proposed by Marerba et al. (see: Malerba et al. [2004]). This research was conducted on symbolic data from a variety of models (generated by procedure cluster.Gen from package clusterSim for **R** software). These models contained a known number of classes. In addition, each model also contained a different number of noisy variables and outliers added to obscure the underlying cluster structure.

Key words: classification, k-nearest neighbors, symbolic data.

I. INTRODUCTION

Symbolic data is an extension of multidimensional analysis dealing with data in an extended form. Each symbolic object can be described by a single quantitative value, a categorical value, an interval, a multivalued variable, a multivalued variable with weights (Bock, Diday (2000), pp. 2-3). Due to this data representation symbolic data analysis introduces new methods and implements traditional methods in which symbolic data can be treated as an input.

The article presents an adaptation of the well-known classical *k*-nearest neighbour method to a symbolic data case and compares the quality of prediction in various scenarios with and without noisy variables and outliers.

The first part of this article is an introduction to the symbolic data analysis in which symbolic objects, symbolic variables are described. Also dissimilarity measures for symbolic objects are presented. The second part presents how classical *k*-nearest neighbour method may be applied for symbolic objects. The third part presents a computational simulation comparing results of the

[*] Ph.D., Chair of Econometrics and Informatics, University of Economics, Wrocław.

classification process with an application of KNN for symbolic objects in various scenarios with and without noisy variables and outliers. Finally some conclusions and remarks are given.

II. SYMBOLIC OBJECTS AND VARIABLES

Each symbolic object can be described by the following variables (Diday [2000], pp. 2-3):

1) single quantitative value,
2) categorical value,
3) quantitative variable of interval type (interval-valued variable),
4) set of values or categories (multivalued variable),
5) set of values or categories with weights (multivalued variable with weights).

Variables in the symbolic data analysis can also be (see: Bock and Diday (eds.) [2000], 2):

a) taxonomic dependant – which presents *a prior* known structure,

b) hierarchically dependent – rules which decide if a variable is applicable or not have been defined,

c) logically dependent – logical or functional rules which affect variable's values have been defined.

Symbolic data unlike classical data situation is more complex than tables of numeric values. While table 1 presents usual data representation with objects in rows and variables (attributes) in columns with a number in each cell, table 2 presents a table of symbolic objects.

Table 1: Classical data situation

Variable / Object	Car color	Price of a car (in PLN)	…	Car mark
1	grey	20000	…	Toyota
2	black	36000	…	Audi
3	blue	72000	…	Renault
…	…	…	…	…
n	red	12000	…	Fiat

Source: own research (artificial data).

Table 2: Symbolic data table

Variable / Object	Car color	Price of a car (in PLN)	…	Car mark
1	{grey, black}	(20000; 36000)	…	{Toyota (30%), Audi (70%)}
2	{blue, red}	(30000; 45000)	…	{Fiat (40%), Renault (60%)}
3	{red, white}	(46000; 65000)	…	{Honda (75%), Fiat (25%)}
…	…	…	…	…
n	{blue, red, white}	(135000; 166000)	…	{Mercedes (100%)}

Soruce: own research (artificial data).

There are four main types of dissimilarity measures for symbolic objects [Malerba et. al. (2001); Bock, Diday (2000), pp. 166-183]:

1. Gowda, Krishna and Diday – mutual neighbourhood value, with no taxonomic variables implemented;
2. Ichino and Yaguchi – dissimilarity measure based on operators of Cartesian join and Cartesian meet, which extend operators $\cup$ (sum of sets) and $\cap$ (product of sets) onto all data types represented in symbolic objects,
3. De Carvalho measures – extension of Ichino and Yaguchi measure based on a comparison function (*CF*), aggregation function (*AF*) and description potential of an object.
4. Hausdorff distance (for symbolic objects containing intervals).

III. ADAPTATION OF *K* NEAREST NEIGHBOUR FOR SYMBOLIC OBJECTS

Due to symbolic variable types some assumptions in *k* nearest neighbour method are made [Malerba et. al. (2001), pp. 23-24):

1. symbolic objects cannot be treated as points in a hyperdimensional space,
2. dissimilarity measure for symbolic objects is applied,
3. contribution of each neighbour is weighted with respect to its closeness to the classified symbolic object,
4. number of nearest neighbours can be extracted on the basis of a cross-validation of the training data.

As a result of *k* nearest neighbour classification for symbolic objects, we get *posterior* probabilities of assigning an object to clusters. This result may be treated as a fuzzy clustering or objects are assigned to a cluster with the highest probability.

Estimation of *posterior* probabilities is estimated in three ways [Malerba et. al. (2001), pp. 25):

1. If a distance from classified object (**O**) and all its *k* neighbours are equal to 0 and all neighbours are from the same class (C_j), *posterior* probability is estimated by:

$$P(\mathbf{O} \mid C_j) = 1. \tag{1}$$

2. If a distance from classified object and all its k neighbours are equal to 0, but neighbours are from different classes *posterior* probability is estimated by:

$$P(\mathbf{O} \mid C_j) = \frac{K_j}{K}, \forall j = 1,\ldots,J, \tag{2}$$

where:
K_j – number of neighbours from *j*-th class,
K – total number of neighbours,
$j = 1,\ldots,J$ – number of classes.

3. If a distance from the classified object and its neighbours differs from 0 *posterior* probability is estimated by:

$$P(\mathbf{O} \mid C_j) = \frac{\frac{K_j}{K} \cdot \Omega_j}{\sum_{j=1}^{J} \frac{K_j}{K} \cdot \Omega_j}, \forall j = 1,\ldots,J, \tag{3}$$

$$\Omega_j = w_i \cdot \delta(C_j, C_k) \tag{4}$$

where:

$w_i = \frac{1}{d(\mathbf{O}, \mathbf{X}_i)}$ – weights,

$\delta(C_j, C_k) = 1$ – if the *k*-th neighbor class is the same class to which we assign object **O**,

$\delta(C_j, C_k) = 0$ – if the *k*-th neighbor class is not the same class to which we assign object **O**,

$d(\mathbf{O}, \mathbf{X}_i)$ – distance measure between an object (**O**) to be classified and *i*-th object.

IV. SIMULATION

Five different symbolic data sets have been generated for simulation purposes (with the application of cluster.Gen from clusterSim package for R). Parameters of each model are described in table 3.

Table 3: Models of simulation

Model	Number of variables	Number of clusters	Variable types	Learning set	Training set
1	2	2	interval	200	40
2	2	5	interval	250	50
3	4	3	interval	200	40
mushrooms[1]	2	2	interval and categorial	106	27
patients[1]	2	2	interval and categorial	308	77

Source: own research.

Table 4 presents results of clustering for every model with no noisy variables, (2, 3, 5 noisy variables). Table 5 present result of clustering for each model with no outliers, (5%, 10%, 20% of outliers). For each model and scenario error ratio is calculated.

Table 4: Error ratio

Number of noisy variables and number of neighbours / Model	0 K= 10	2 K= 10	3 K= 11	5 K= 12
1	0,00%	48,78%	37,50%	50,00%
2	3,00%	71,70%	78,31%	81,58%
3	0,00%	37,50%	47,50%	50,00%
mushrooms	7,00%	36,54%	56,12%	75,32%
patients	30,00%	47,43%	64,12%	87,34%

Source: own research

Table 5: Error ratio

Percent of outliers and number of neighbours / Model	0% K= 10	5% K= 10	10% K= 11	20% K= 12
1	0,00%	4,88%	7,32%	9,32%
2	3,00%	16,83%	28,45%	47,12%
3	0,00%	14,56%	29,45%	75,33%
mushrooms	7,00%	19,21%	30,03%	57,75%
patients	30,00%	36,02%	46,00%	68,15%

Source: own research

[1] Mushrooms and patients data was prepared by Malerba et. al. for symbolic KNN purposes.

V. FINAL REMARKS

For artificially generated and real data with no noisy variables or outliers, KNN clustering for symbolic data usually gives classification results with a low error ratio. Error ratio rises rapidly when noisy variables are added. KNN clustering for symbolic data gives also classification results with a quite low error ratio for data with a low number of outliers (usually up to 5-10% of a test set).

An open issue for further research is a comparison of KNN classification for symbolic data with other clustering methods for symbolic data, for example, kernel discriminant analysis and symbolic classification trees.

REFERENCES

Bock H.-H., Diday E (Eds.) (2000), *Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*, Springer Verlag, Berlin.

Fix E., Hodges J. L. (1951), *Discriminatory analysis – nonparametric discrimination: consistency properties,* Project 21-49-004, Report no. 4, USAF School of Aviation Medicine, Randolph Field, 261-279.

Ichino M., Yaguchi H. (1994), Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 24, No. 4, 698-707.

Malerba D., Espozito F, Giovalle V., Tamma V. (2001), Comparing Dissimilarity Measures for Symbolic Data Analysis, *New Techniques and Technologies for Statistics* (ETK-NTTS'01), 473-481.

Malerba D., Esposito F., D'Amato C., Appice A. (2004), K-nearest neighbor classification for symbolic objects [in:] P. Brito, M. Noirhomme-Fraiture (Ed.), *Symbolic and spatial data analysis: mining complex data structures*, University of Pisa, Pisa, 19-30.

Marcin Pełka

KLASYFIKACJA METODĄ K-NAJBLIŻSZYCH SĄSIADÓW DLA DANYCH SYMBOLICZNYCH

Reguła *k*NN (*k* Nearest Neighbours) została zaproponowana w pracy (Fix E., Hodges J. L. [1951]) i jest jednym z najlepszych klasyfikatorów dla danych w ujęciu klasycznym. W najprostszym ujęciu metoda *k*-najbliższych sąsiadów polega na tym, że klasyfikowany obiekt jest zaliczany do klasy najliczniej reprezentowanej wśród jego *k* „najbliższych sąsiadów". Jeżeli w tej samej odległości, co *k*-ty „sąsiad" znajdą się jeszcze inne elementy, to wszyscy ci „sąsiedzi" biorą udział w głosowaniu.

W artykule zaprezentowano adaptację metody KNN dla danych symbolicznych, którą zaproponował zespół pod kierownictwem D. Malerby (por. Malerba i in. [2004]). Badania przeprowadzono na danych symbolicznych w różnych modelach (generowanych za pomocą procedury cluster. Gen z pakietu clusterSim dla programu **R**). Modele te zawierały znaną liczbę klas. Dodatkowo do każdego modelu dodano różną liczbę zmiennych zakłócających i wartości odstających, które zniekształcają oryginalną strukturę klas.