Contribution to Multicriterial Classification of Spatial Data

Jana Výrostková, Eva Ocelíková

Department of Cybernetics and Artificial Intelligence
Technical University of Košice
Letná 9, 041 20 Košice
Slovak Republic
Jana.Vyrostkova@tuke.sk, Eva.Ocelikova@tuke.sk

Abstract: Problems of classification has a great meaning at the handling of information. Statistical approaches, decision trees and approaches of artificial intelligence (sphere of neuron network) belong to standard methods of classification. This paper deals with simple classifiers – k-nearest neighbours and bayesian classifier, also with component classifiers - boosting and stacked generalization applied on experimental artificially created data and also on real data from remote sensing of the Earth.

Keywords: classification, classifiers, k - nearest neighbours, bayesian classifier, architecture boosting, stacked generalization

1 Introduction

Every algorithm which solve a problem of classification prefers one solution before others. That means, every created classifier has a good precision of classification in certain cases, but at the same time there are cases in which precision of classification is worse.

Precison of classification could be increase by the best properties of few classifiers united to the one component classifier. The advantage of the component classifier is union of different approaches, for example statistical approaches [1], neuron network or decision trees [2]. For all these advantages, usage of component classifier is not still sufficient and effectively in event when data set is more unique at the beginning. In this case is better use sufficient simple classifier.

Next chapters deal with simple and component classifiers applied on artificially created data and also on real data.

2 Experimental Part

2.1 Artificial Data

Artificially created data set Squares, Fig. 1, consists of six different color squares.

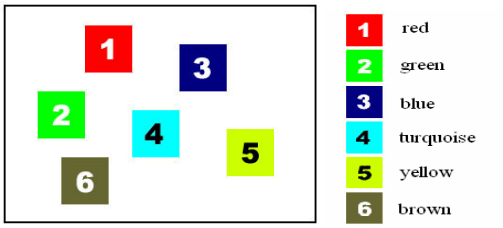

Figure 1
Test data set Squares

Train data set consists of three objects:

- object 1 described by three attributes (0.339 0.000 0.000), red - .
- object 2 described by three attributes (0.000 0.339 0.000), green - .
- object 3 described by three attributes (0.000 0.000 0.339), blue - .

Test data set Squares was classified by simple classifiers: k-nearest neighbours and bayesian classifier and by component classifiers: boosting a stacked generalization. The classifiers were learned on train data set mentioned thereinbefore. Results of classification are shown on Fig. 2, Fig. 3 and Fig. 4.

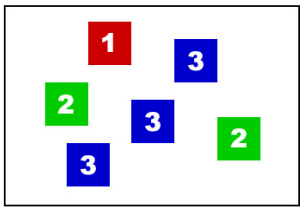

Figure 2
Classifier 3-NN

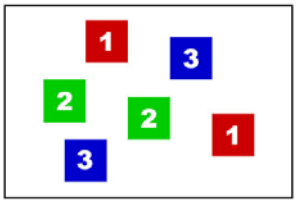

Figure 3
Bayesian classifier

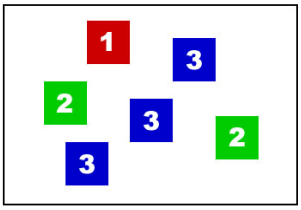

Figure 4
Architecture Boosting (C1:3-NN, C2:1-NN a C3:4-NN)

Figures shown results of classification of 6 squares of different color by mentioned train data set, which consists of only three objects of red, blue and green color. Difference between results of se methods is distinct. The best of simple classifiers is k-nearest neighbours, exactly 3-nearest neighbours and the best of component classifiers is architecture boosting.

2.2 Real Data

Experiments were realised on data from remote sensing of the Earth by satellite LANDSAT 7 ETM + (Enhanced Thematic Mapper Plus) of association NASA with seven sensors and their arguments [5].

The data set consists of 368 152 specimens surface of the Earth, where one of them represents area of 30 x 30 meters representing a total of 332 sq km of land. The specimen of the Earth surface is characterized by a 7- dimensional vector. These partial components are describing the brightness of the seven spectral bands. The data set has 475 rows and 775 columns. The figure of north part of Košice has dimension 475 x 775 pixels, Fig. 5.

Only a part of data set was use for following experiments, Fig. 5. That means, testing data set created with dimensional 83300 speciemens has 350 rows and 238 columns, that is circa 25% of whole data set, Fig. 6. This data set is smaller so calculation of individual methods is faster. At the same time information proceeds was observed, because these data consists of all important objecst for experimental realisation.

Figure 5
Segment of test data set for experiments

After the creating of test data set we create the train data set. The train set is as follows:

$$T = \{(x_1, \omega_r), (x_2, \omega_r), \ldots, (x_n, \omega_r)\}, \quad (1)$$

where ω_r is associated to a r - class. This means that to every object an equivalent class is associated. It is not easy to determine a proper association.

Well-done train data set leds to precise results of experiments. Objects were classified into seven thematic categories by expert, there are shown in Fig. 6.

Figure 6
Seven thematic categories

Train data set created by expert was used for the training of classifiers. There were created three train data sets, namely *kosice-460.tr* consists of 460 objects, *kosice-3166.tr* consists of 3166 objects and *kosice-6331.tr* consists of 6331 objects, to which adequate classes are assigned.

Advantage of smaller train data sets is shorter time of classification. Advantage of larger train data sets is their higher information acquisition for classification process.

Fig. 7 shows sample of the train data set *kosice-6331.tr* and Fig. 8 shows sample of the test data set *kosice-83300.te*.

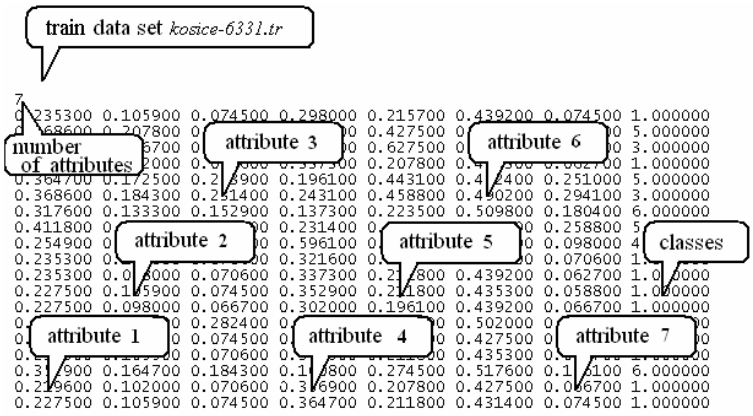

Figure 7
Sample of the train data set kosice-6331.tr

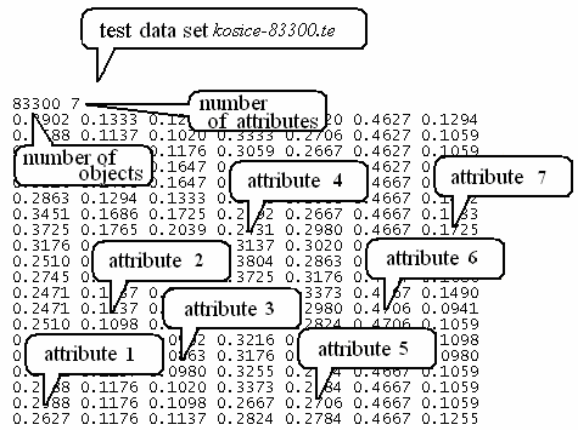

Figure 8
Sample of the test data set kosice-83300.te

The data from remote sensing of the Earth by satellite LANDSAT 7 ETM + (Enhanced Thematic Mapper Plus) of association NASA [5] with seven sensors and their arguments are presented in Table 1. These 7 sensors generate 7 attributes of every object.

Sensors 1, 2 and 3 are all together using to display a real world in RGB spectrum. Sensors 4, 5, 6 or 7 together with sensors 1, 2 and 3 demonstrated to conditions of vegetation. One pixel is represented of area 30x30 metres.

Table 1
Sensors of satellite LANDSAT 7 ETM+

	Wave-length	**Spectrum**	**Area**
Sensor 1	0,45 - 0,52 μm	Blue	30 x 30 m
Sensor 2	0,52 - 0,60 μm	Green	30 x 30 m
Sensor 3	0,63 - 0,69 μm	Red	30 x 30 m
Sensor 4	0,76 - 0,90 μm	Near IR	30 x 30 m
Sensor 5	1,55 - 1,75 μm	Middle IR	30 x 30 m
Sensor 6	10,40 - 12,50 μm	Thermal IR	120 x 120 m
Sensor 7	2,08 - 2,35 μm	Middle IR	30 x 30 m

The following figures shown results of classification by k-nearest neighbours, bayesian classifier, architecture boosting and stacked generalization.

Method **k-nearest neighbours** (**k-NN**) [3] contains index k where $1 < k < 10$, and so the using of same train data set can achieves 10 different results of classification. There were used train data set with a different size *kosice-460.tr, kosice-3166.tr* and *kosice-6331.tr*. Result of classification is shown on Fig. 9 for k=6 and *kosice-6331.tr*.

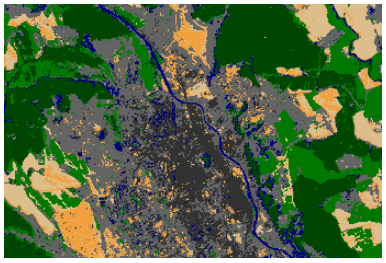

Figure 9
Method 6 – NN, train data set kosice-6331.tr

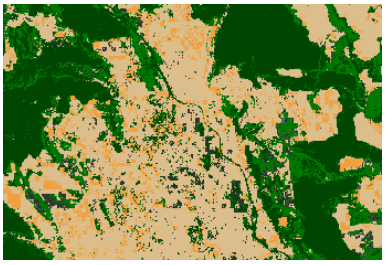

Figure 10
Bayesian classifier, train data set kosice-6331.tr

Bayesian classifier compares the attributes of train set with the attributes of the test set. If the attribute from train set is the same as the attribute selected from the test set, only then this attribute is appropriate for the classification process. Exactly this is the insufficiency of this method. If we insert into classification an attribut from the train set where the difference between the value of attribute from the train set and the test set is in an allowed interval, then we obtain more correct classified object. This classifier depends on data set. The result of classification obtained with bayesian classifier shows Fig. 10.

Architecture **Boosting** is based on sufficient choosing objects from train data set for learning of individual classifiers. In first event classification was done by used these classifiers: 3-NN, 6-NN a 9-NN, Fig.11 and in second event classification was done by used classifiers were 1-NN, 3-NN a 4-NN, Fig. 12.

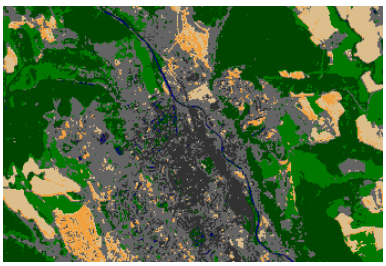

Figure 11
Architecture Boosting, C1:3-NN, C2:6-NN
a C3:9-NN

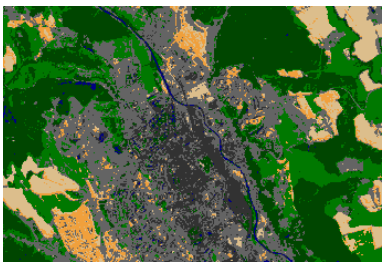

Figure 12
Architecture Boosting C1:1-NN, C2:3-NN
a C3:4-NN

For the architecture **Stacked generalization** [4] it is hard to decide what and how many classifiers to use on zero level. The structure of data affects the selection of specific classifier type also. Some combinations of classifiers are successful on specific data set but the results on others data sets are not so suitable. Fig. 13 illustrates the successful result of classification using the combination of classifiers 1-NN, bayesian classifier, 3-NN and 4-NN. Fig. 14 shows the less successful result obtained by classifier 1-NN, decision tree, 3-NN and 4-NN.

Figure 13
Architecture Stacked Generalization, C1:1-NN, C2: Bayesian classifier, C3:3-NN a C4:4-NN

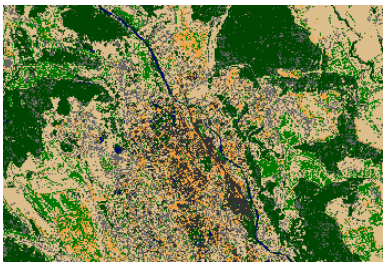

Figure 14
Architecture Stacked Generalization, C1:1-NN, C2: Decision tree, C3:3-NN a C4:4-NN

Conclusions

In this paper were presented simple classifiers (k- nearest neighbours and bayesian classifier) and component classifiers (boosting and stacked generalization) on artificially created data and real data, too.

On the basis of mentioned experiments, k-nearest neighbours is the best of all simple classifiers. This classifier achieves high percentage success and obtained picture by classification is most similar to the real picture also.

Experiments were realised for different values k where $1 < k < 10$, but the best classification results were from interval k: $3 < k < 6$. This method is very simply and as well precise. Bayesian classifier achieved worse results than classifier k-nearest neighbours.

Architecture boosting achieved worse results than stacked generalization. This method can be improved by usage algorithm AdaBoost, which allows to attache a weak trainees while classification error is not minimal. Results of stacked generalization show that lower precision of one classifier doesn't prove or minimal proves the precision of classification process.

There are a lot of unsolved tasks in classification area as elimination of distorted data, classification of incomplete samples, correction of wrong samples or elimination of time and memory difficulties. These problems require increased

attention because they are daily problems of our life, e.g. public health, agriculture, industry, economy, banking, geology, etc.

Acknowledgement

This work is supported by the VEGA project No 1/2185/05.

References

[1] Michie, D., Spiegelhalter, D. J., Taylor, C. C.: Machine Learning Neural and Statistical Classification, Ellis Horwood, 1994

[2] Ocelíková, E., Krištof, J.: Classification of Multispectral Data. Journal of Information and Organizational Sciences, Vol. 25, Number 1, Varaždín, 2001, pp. 35-41

[3] Skalak, D. B.: Prototype Selection for Composite Nearest Neighbor Classifiers. CMPSCI Technical Report 96-89, 1997

[4] Ting, K. M., Witten, I. H.: Issues in Stacked Generalization. Journal of Artificial Inteligence research 10 (1999) 271-289

[5] NASA: <http://landsat.gsfc.nasa.gov/>

[6] Výrostková, J., Ocelíková, E., Klimešová, D.: Fuzzy Clustering of the Multidimensional Data. In: Proceedings of 7th International Scientific-Technical Conference Process Control 2006, ŘÍP 2006, June 13-16, 2006, Kouty nad Desnou, Czech Republic, p. 133

[7] Ocelíková, E., Výrostková, J.: Object Fuzzy Clustering. In: Proceedings of 6th International Symposium of Hungarian Researchers on Computational Intelligence, Budapest, November 18-19, 2005, pp. 650-657