BAGGING ON SUBSPACES WITH BRAVERMAN 'S CLASSIFIERS[1]

Alexandra Kononova, Mikhail Alexandrov, Dmitry Stefanovskiy, Javier Tejada

***Abstract**: Non-compact and non-uniform object distributions in classes are the well-known reasons, which decrease quality of classification. To improve it we propose to use bagging on subspaces of object parameters with modified method of potential functions (Braverman's classifier). In the paper we shortly describe the proposed technology and end-user software. As an example of application we consider classification of Russian regions related to their investment attractivity.*

***Keywords**: bagging, classification, Braverman's method.*

***ACM Classification Keywords**: I.2 Artificial Intelligence.*

Introduction

Nowadays there are dozens of classification methods tested on many real examples [Bishop, 2006]. But modern approach in classification consists in combining methods instead of their individual use. Such an approach allows to perform successful classification in case of very complex data structures in parameter space. Two technologies realizing this approach are well-known: boosting and bagging.

Boosting is formation of a sequence of elementary classifiers, where each subsequent classifier corrects errors of the previous ones on learning sample. The result is one combined classifier [Schapire, 1999]. One should note that boosting uses elementary classifiers defined on the whole parameter space. The main advantage of boosting is taking into account non-compactness and non-uniformity of objects distribution in classes.

Bagging uses many classifiers, which independently assigns objects to its classes. The final decision is determined by the rule of consensus or the rule of majority. Bagging was firstly proposed in the monograph [Rastrigin, 1981], where the authors used different classifiers in the whole space of parameters. This technology were named collective recognition. Modern bagging technology was considered in [Breiman, 1996]. Here the author divided parameter space on subspaces with their own classifiers. The main advantage of bagging is the simplicity of its realization.

One should note that popular packages related to Data Mining include classification methods but this packages do not contain combinations of methods [Weka, http; Rapid Miner, http]. Boosting and bagging are included into the special package Adabag [Adabag, http]. Bagging in this package is based on the traditional method of decision tree.

In this paper we propose the technology that takes into account the advantages both boosting and bagging. Namely:

1. We use bagging on subspaces chosen by an expert. Here expert himself/herself groups parameters taking into account their mutual relations. We expect that object distributions in subspaces will prove to be more compact.
2. We use the modified method of potential functions in each subspace. Here classifiers learn individually in its subspaces. This procedure corrects errors and makes object distributions more uniform.

The method of potential functions and its modification were proposed in [Braverman, 1970]. But the latter was described without details and therefore by the moment the working version of modified Braverman's method

[1] The work done under partial support of the company Apoyo Total S.A. (subsidized by FINCYT–PERU, Foundation for innovation, science and technology)

is unknown. Also, by the moment neither Braverman's method nor its modification were used inside any bagging technology.

The paper is structured by the following way. Section 2 contains the short description of bagging technology, our version of modified Braverman's method, and developed software. In section 3 we demonstrate the results of several experiments with classification of Russian regions. Section 4 contains conclusions.

Classification Technology

Collective classification and process of decision-making

As we have already mentioned above non-compact object distribution in the space of all given parameters can lead to errors of classification. To reduce the number of errors the complete set of parameters is divided on groups, where each group forms its subspace. The classification is implemented separately in these subspaces. Figure 1 illustrates bagging in 3 subspaces (p1, p2), (p3, p4), (p5, p6). First classifier assigns object x to the class marked '1', the second one and the third one assign this object to the class ´0´ .

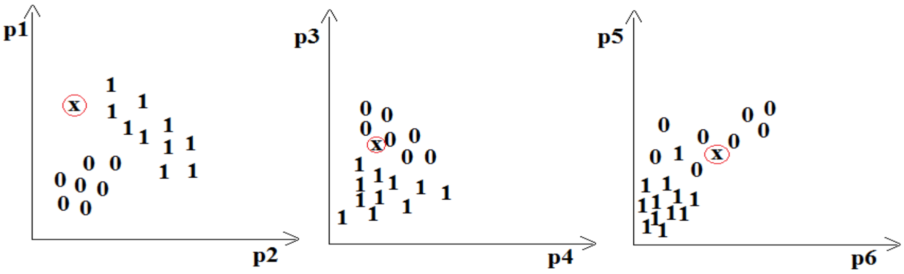


Figure 1. Bagging on subspaces

Subspace definition is implemented by an expert and here he/she tries to join together so-called related parameters. Such an approach is based on the natural hypothesis that compact classes are more probable in spaces with related parameters than in spaces with independent (unrelated) parameters.

To use bagging it is necessary to define a rule of decision-making about the status of object under consideration. Ideally, it is consensus. In this case all classifiers have the same opinion about the object class. If consensus is not reached then the decision should be based on a simple majority. Expert's preferences to certain classes can be taken into account by means of weights to be assigned to classifiers.

Braverman's method of potential functions

The idea of the Braverman's method is the following. There are representatives of each given class in training set. We use here term 'point' to name objects to be classified. Each object creates potential of its class in a test point. Different formulae can be proposed for computing the total potential of a class. Here is one of them:

$$\varphi_i = \frac{1}{n}\sum_{j=1}^{n} \frac{w_j}{(1+ar_j)}$$

where w_j is a weight of object from a training set, r_j is the Euclidean distance between this object and a test point, n is a quantity of objects in a class under consideration and α is a coefficient.

As we have mentioned above one can propose the other formulae. For example, $(\alpha r_j)^2$ can be used instead of αr_j etc. The formula contains the parameter $\alpha=1/R$, where R is some typical size related to subspace. For example, it can be equal 50% of distance between the farthest objects in training set.

Figure 2 is the illustration of the method of potential functions. Point distribution is shown on the flat and potentials generated by the objects are volumetric figures.

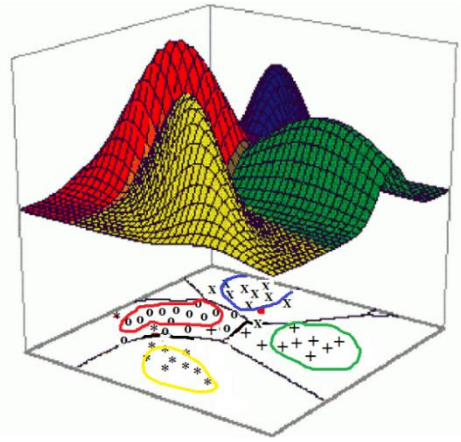


Figure 2. Potentials of classes

In real cases the distribution of points within classes is often essentially non-uniform that needs modification of the method. To do this one corrects weights using procedure of cross validation. The procedure includes 2 steps:

Step 1. Here all representatives of classes are reclassified once more. If an object is classified incorrectly then it is marked as 'incorrect object'. The other ones on this iteration are considered as 'correct objects'.

Step 2. Here all incorrect objects are considered. The nearest correct object to incorrect object increases its weight. Naturally both objects are to belong to the same class.

Step 3. If the quantity of incorrect objects does not decrease then the process stops. Otherwise the new iteration is repeated.

The described procedure looks like augmentation of objects of classes, to which incorrect objects belong. As a result we have more uniform object distribution. Graphic illustration of the method is presented below on Figure 3:

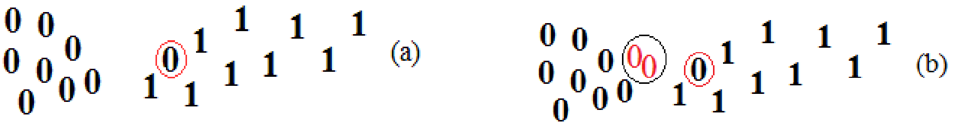


Figure 3. Interpretation of modified method

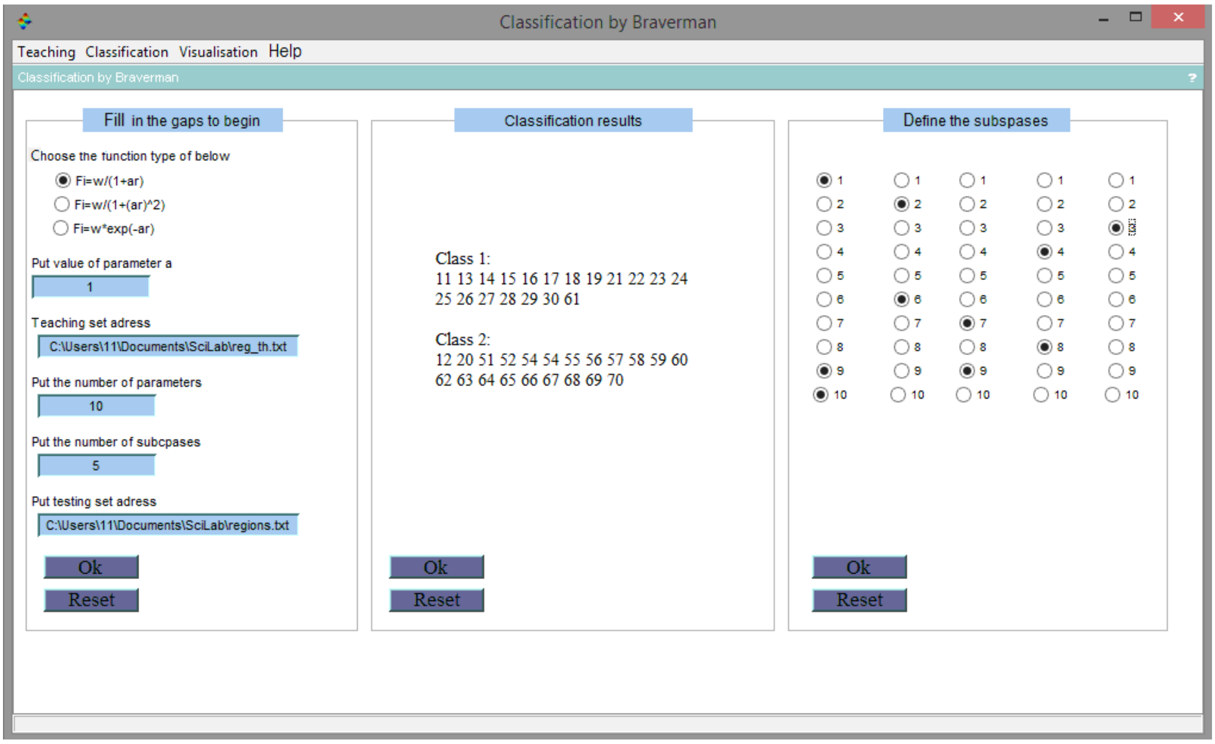
(a) two classes with the incorrectly classified object from the class '0'

(b) the nearest object from the class '0' doubles its weight

Software

The program was developped on the free-share platform SciLab [SciLab, http]. It contains all computational procedures and friendly interface. Preprocessing (normalization and outliers determination) is implemented in a separate module. User defines: number of parameters, contents of subspaces, sources of initial data with representatives of classes and objects to be classified. The program allows to use 2-10 parameters, 1-5 subspaces, 2-3 classes. The quantity of objects is limited by 10000 units. It is sufficient for many applications. Interface for subspace definition is presented on Figure 4 to the right.

Structures of input files for training and testing samples are almost similar. The only difference is: each object of training sample has the class label. The results are reflected in window of the interface. The results are also saved in output file. Figure 4 shows the program interface. It is used to manage the process of classification and to control the results


Classification by Braverman
Teaching Classification Visualisation Help
Classification by Braverman
?

Fill in the gaps to begin

Choose the function type of below

● Fi=w/(1+ar)
○ Fi=w/(1+(ar)^2)
○ Fi=w*exp(-ar)

Put value of parameter a: 1

Teaching set adress: C:\Users\11\Documents\SciLab\reg_th.txt

Put the number of parameters: 10

Put the number of subcpases: 5

Put testing set adress: C:\Users\11\Documents\SciLab\regions.txt

Classification results

Class 1:
11 13 14 15 16 17 18 19 21 22 23 24
25 26 27 28 29 30 61

Class 2:
12 20 51 52 54 54 55 56 57 58 59 60
62 63 64 65 66 67 68 69 70

Define the subspases

● 1	○ 1	○ 1	○ 1	○ 1
○ 2	● 2	○ 2	○ 2	○ 2
○ 3	○ 3	○ 3	○ 3	● 3
○ 4	○ 4	○ 4	● 4	○ 4
○ 5	○ 5	○ 5	○ 5	○ 5
○ 6	● 6	○ 6	○ 6	○ 6
○ 7	○ 7	● 7	○ 7	○ 7
○ 8	○ 8	○ 8	● 8	○ 8
● 9	○ 9	● 9	○ 9	○ 9
● 10	○ 10	○ 10	○ 10	○ 10

Figure 4. Program interface

Experiments

Source data

To demonstrate the proposed technology we completed experiments with classification of the Russian regions related to their investment atractivity. The initial data contains 80 regions. Each regions is described by 10 parameters. The part of data is presented in Table 1. Here all regions are ordered according their rating.

Table 1. Parameters for Russian regions

Number	***Region***	***p1***	***p2***	***p3***	***p4***	***p5***	***p6***	***p7***	***p8***	***p9***	***p10***
1	Moscow	95,2	87,4	74,8	69,1	39,7	76,2	91,5	100	84,4	61
2	St.Petersburg	74,4	79,2	74,4	57,7	57,4	56,9	81,3	93,1	72,1	58,8
3	Moscow Region	57,2	77,7	48,6	54,8	48,8	57,1	66,1	80,6	71,6	66,7
4	Tatarstan	56,9	75,6	53,1	61	60	50,7	60	45,7	53,4	53,3
5	Krasnodar region	42,9	76,1	47,4	79,7	48,3	52,1	60,2	49,8	52,6	68,5
6	Belgorod region	51,6	68,2	44,1	52,3	63	39,8	61,2	52,6	53	38,4
7	Yugra	64	74	49,5	36,8	59,5	62,1	60,9	19,7	71,9	42,4
8	…										

Here: p1 is income level, p2 is living conditions, p3 is social infrastructure, p4 is ecology, p5 is safety, p6 is demography, p7 is education and health, p8 is transport infrastructure, p9 is economy development level, p10 is entrepreneurship development level.

We deal with 2 classes, the positive and the negative ones. To form these clases and to select the training and testing sets we prepare data according to Table 2.

Table 2. Data preparation

Rating	***%%***	***Category***
1-8	10	Training set (good)
9-24	20	Testing set (good)
25-56	40	To be excluded
57-72	20	Testing set (bad)
73-80	10	Training set (good)

Classification of regions

We completed 4 experiments under different conditions to study the quality of classification. The results are presented in Table 3. In all experiments we used decision-making based on majority rule.

Table 3. Results of experiments

Experiment conditions	***Accuracy***
Classification in the complete 10D space Here we have no subspaces	62,5%
Bagging on 5 subspaces All 10 parameters are included	85%
Bagging on 3 subspaces Parameters related to economy, health and education are excluded	75%
Bagging on 3 subspaces Parameters related to safety and ecology are excluded	90%

The table shows that:

- Bagging essentially improves the results of classification in all cases. It means that our technology leads to higher compactness and uniformity of object distribution in classes.
- Compactness and uniformity are related rather to parameters reflecting economy, health and education than safety and ecology.

Conclusions

In the paper:

- Interactive bagging is developed in the form of end-user software.
- Braverman's method with the nearest neighbor correction is realized and tested.
- Proposed technology demonstrates its advantages on the real example.

In the future we plan to include preprocessing and procedures of visualization to the program.

Bibliography

[Adabag, http] Inside-R Adabag electronic resours http:// www.inside-r.org/packages/cran/adabag

[Bishop, 2006] Bishop. C. Pattern Recognition and Machine Learning, Springer, 2006

[Braverman, 1970] Arcadev A., Braverman E. Learning machine for classification of objects. Science, 1971 (rus)

[Breiman, 1996] Breiman L. Bagging predictors. Machine Learning , 1996.

[RapidMiner, http] electronic resource: http:// rapid-i.com.

[Rastrigin, 1981] Rastrigin L., Erenstein R. The method of collective detection. Energoisdat, 1981.(rus)

[Sci Lab, http] electronic resource: http:// www.scilab.org/

[Schapire, 1999] Freund Y., Schapire R, A Short Introduction to Boosting, Shannon Lab.USA,1999.,pp.771-780

[Weka, http] electronic resource: http:// www.cs.waikato.ac.nz/ml/weka/

Authors' information

Alexandra Kononova – *M.Sc student, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; Moscow Institute of Physics and Technology (State Research University); Institutskii per 9., Dolgoprudny, Moscow Region, 141700, Russia;*

e-mail: kononova@ phystech.edu

Major Fields of Scientific Research: data mining, classification

Mikhail Alexandrov – *Professor, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain;*

e-mail: malexandrov@ mail.ru

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

Dmitry Stefanovskyi – *Assoc. Prof., Ph.D, The Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russian Federation;*

e-mail: dstefanovskiy@ gmail.com

Major Fields of Scientific Research: mathematical modeling, world economy

Javier Tejada – *Professor of Computer Science Department, San Pablo Catholic University; Campus Campiña Paisajista s/n Quinta Vivanco, Barrio de San Lázaro, Arequipa, Perú;*

e-mail: jtejada@ itgrupo.net

Major Fields of Scientific Research: Natural Language Processing, Business Intelligence