AN APPROACH FOR TRAINING DATA REDUCTION USING BILINEAR FORM OPTIMIZATION

Vasily Ryazanov

***Abstract**: An approach for solving the problem of dataset reduction is considered. The problem of selection an optimal subset of features and objects is an important task for every classification algorithm. Having smaller and more informative dataset one can perform training operation faster and study data visually. However nowadays most of algorithms select features and objects separately and are based on statistical or logical base. In this paper a method for training set reduction is presented. Using votes for each class in calculation estimation algorithm a bilinear form is constructed. Having optimized bilinear form one can find an optimal subset for features and objects at once. In order to fasten optimization a technique for linear local optimization is proposed. During bilinear form optimization one can select an optimal iteration with smaller dataset and acceptable classification quality. Prospects of this approach are confirmed by a series of experiments on various practical tasks.*

***Keywords**: classification, data mining, supervised learning, dataset reduction, bilinear form.*

***ACM Classification Keywords**: I.2.4 ARTIFICIAL INTELLIGENCE Knowledge Representation Formalisms and Methods – Predicate logic, I.5.1 PATTERN RECOGNITION Models – Deterministic, H.2.8 Database Applications, Data mining.*

1. Introduction

Currently interest in the problem of minimizing training data, i.e. the selection of smaller informative training subsets, is growing.

Performance of present algorithms depends on solving the problem of an adequate description of objects and classes, with the smaller object-feature space, the optimization problems arising in training are solved easier and more accurate. Therefore, the task of reducing the training information is always of great interest.

With training data increasing some classifiers can physically lose the ability to process the data and construct the prediction, so data reduction can solve even this problem. Smaller tables also allow experts to examine directly and visualize data, thereby to find new patterns.

Current methods of minimizing the data are mostly built to optimize the number of features. To solve this problem algorithms based on information theory and correlation [Lei Yu et al., 2003] are often used. They introduce criteria based on entropy, statistical heuristics, and study the mutual features correlation. Algorithms of feature selection based on logical methods (binary trees, logical regularities) are also widely spread today. For example, in [Norbert Jankowski et al., 2005] the classical approach SSV (Separability Split Value) to ranking criteria is proposed: the importance of the feature is higher, the more frequently it occurs in the casting tree nodes and sooner it happens on a partition. Methods based on alternately adding and removing features, and then validation on test sample [P. Pudil et al., 1994], are also popular. Their disadvantages include a durable working time, random during searching for the optimum, as well as dependence on the validation set.

Despite a somewhat less attention to the domain of objects selection, this problem is no less important during classification. In many cases simple empirical search for anomalies, as well as filtering objects by some criteria in order to exclude clearly irrelevant objects are used. There are approaches based on the k-nearest neighbors algorithm [P. Hart, 1968], as well as other algorithms such as SVM or genetic algorithms.

To get the maximum knowledge from the data, objects and features should be examined in combination. Thus prerequisites for creating algorithm of complex data optimization occur.

In this paper the formulation and approach of the problem of the simultaneous reduction of feature and object descriptions are proposed. The task is to select a subset of informative features and objects, reducing the sample size while preserving quality close to the original.

A modification of ‘calculation estimation’ algorithm [Zhuravlev, 1978] is used as base approach. Vectors $\mathbf{x} = (x_1, x_2, ..., x_n)$, which has the same dimension as number of features and $\mathbf{y} = (y_1, y_2, ..., y_m)$, which has equal size to number of objects are introduced. Parameters $x_i, y_i \in [0,1] \forall i$ denote features and objects importance.

Next, after classification of test data, linear functions of the object estimations for their own and others' classes, are transformed to the bilinear form $F(\mathbf{x},\mathbf{y})$. Maximizing this criteria, one obtain the optimal vectors $\mathbf{x}$ and $\mathbf{y}$, corresponding to the optimal subsample of the original data.

This paper also proposes a local optimization algorithm of bilinear functional, having a linear complexity in the neighborhood of O^2 from the starting point. Usage of this criterion allow to speed up the optimization process.

2. Initial notations and problem statement.

Consider the following standard problem recognition by precedents. Let there is a set M of objects $\mathbf{z}$, defined with their feature descriptions. For simplicity assume that $\mathbf{z}\in R^n$. The set is $M=\bigcup_{i=1}^{l} K_i, K_i \cap K_j = \varnothing, i \neq j$, with classes $K_i, i=1,2,...,l$. Information of this partition is given by training sample $Z=\{\mathbf{z}_i, i=1,2,...,m\}$, which consists of representatives of each class: $Z=\bigcup_{i=1}^{l} K_i^*, K_i^* \subset K_i$, $\left|K_i^*\right| = n_i$. The task is assigning $\forall \mathbf{z}\in R^n$ to one of the classes.

Let's consider "Calculation Estimation" – baseline classification algorithm [Zhuravlev, 1978], which modification was used in this paper. It is considered that all training objects $Z=\{\mathbf{z}_i, i=1,2,...,m\}$ are divided into l disjoint classes. A natural $k, 1\le k \le n$ is fixed. Sample $\mathbf{z}$ is compared to all the objects from the training above all subsets of features $\Omega \subseteq \{1,2,...,n\}, |\Omega| = k$, having length k to calculate the "degree of proximity" of the object to each of the classes: $\Gamma_i(\mathbf{z}) = \frac{1}{\mathrm{n}_i} \sum_{\mathbf{z}_t \in K_i} \sum_{\Omega:|\Omega|=k} B_\Omega(\mathbf{z}_t, \mathbf{z})$.

Where the proximity function between two objects is defined as follows: $B_\Omega(\mathbf{z}_t,\mathbf{z}) = \begin{cases} 1, & \left|z_{tj} - z_j\right| \le \varepsilon_j, \forall j \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$.

In [Zhuravlev, 1978] it is shown that $\Gamma_i(\mathbf{z}) = \frac{1}{\mathrm{n}_i} \sum_{\mathbf{z}_t \in K_i} C_{d(\mathbf{z}_t,\mathbf{z})}^k$, where $d(\mathbf{z}_t,\mathbf{z}) = \left|\{j: \left|z_{tj} - z_j\right| \le \varepsilon_j, j=1,2,...,n\}\right|$. Here $\varepsilon_j, j=1,2,...,n$ - are parameters. Typically, they are set as $\varepsilon_j = \frac{2}{m(m-1)} \sum_{\substack{u,v=1,2,...,m \\ u>v}} \left|z_{uj} - z_{vj}\right|, j=1,2,...,n$. After calculating values $\Gamma_i(\mathbf{z}), i=1,2,...,l$ the object $\mathbf{z}$ is marked as having class K_j , having maximum score: $\Gamma_j(\mathbf{z}) > \Gamma_i(\mathbf{z}), i,j=1,2,...,l, i \neq j$. Otherwise, a rejection of the classification of the object $\mathbf{z}$ occurs.

Let's introduce the vector parameters $\mathbf{x} = (x_1, x_2, ..., x_n)$ called "feature weights" and the parameters $\mathbf{y} = (y_1, y_2, ..., y_m)$ called "sample weights".

Next modify the CE algorithm and it's scores $\Gamma_j(\mathbf{z})$ for test object $\mathbf{z}=(z_1,z_2,...,z_n)$ to the class K_j having added parameters $\mathbf{x},\mathbf{y}$: $\Gamma_j(\mathbf{z})=\frac{1}{|K_j|}\sum_{\mathbf{z}_\nu\in K_j} y_\nu \sum_{\Omega\in\Omega_A}(\sum_{i\in\Omega} x_i)B_\Omega(\mathbf{z}_\nu,\mathbf{z})$, where $|K_j|$ - number of objects in the class K_j , Ω - a reference set (subset of attributes) form a plurality of reference sets Ω_A of CE algorithm, $B_\Omega(\mathbf{z}_\nu,\mathbf{z})$ - the proximity of the object $\mathbf{z}$ to the training object $\mathbf{z}_\nu$ on the support set. One can show that $\Gamma_j(\mathbf{z})=\frac{1}{|K_j|}\sum_{\mathbf{z}_i\in K_j} y_i(\sum_{t\in J(\mathbf{z},\mathbf{z}_i)} x_t)C^{k-1}_{d(\mathbf{z},\mathbf{z}_i)-1}$ where $J(\mathbf{z},\mathbf{z}_i)=\{\nu:|z_{i\nu}-z_\nu|\le\varepsilon_\nu,\nu=1,...,n\}$, $d(\mathbf{z},\mathbf{z}_i)=|J(\mathbf{z},\mathbf{z}_i)|$, $k,\varepsilon_1,\varepsilon_2,...,\varepsilon_n$ - some of the parameters of the recognition algorithm. In such a setting various subsets of k signs represent all the support sets.

Let's introduce an aggregated functional that will characterize the generalized classification quality. For this let's sum all the scores for "own" classes with a positive coefficient τ_1 and sum all the scores for "foreign" classes with coefficient τ_2 . Without loss of generality one can set $\tau_1=1,\tau_2=-t$.

The final functional will be following: $f(\mathbf{x},\mathbf{y},t)=\sum_{\alpha=1}^{l}\sum_{\mathbf{z}'_\tau\in K_\alpha}\Gamma_\alpha(\mathbf{z}'_\tau)-t\sum_{\alpha=1}^{l}\sum_{\mathbf{z}'_\tau\notin K_\alpha}\Gamma_\alpha(\mathbf{z}'_\tau)$. One can easy show that it is a bilinear form on the parameters $x_1,x_2,...,x_n,y_1,y_2,...,y_m$.

So, the following optimization problem is considered:

$$f(\mathbf{x},\mathbf{y},t)=\sum_{i=1}^{n}\sum_{j=1}^{m}c_{ij}(t)x_i y_j\to\max_{\mathbf{x}\in X,\mathbf{y}\in Y}$$

Two types of areas X and Y can be considered during optimization of this functional: $1\ge x_i\ge 0, i=1,2,...,n, 1\ge y_j\ge 0, j=1,2,...,m,$ or $x_i\in\{0,1\}, i=1,2,...,n, y_j\in\{0,1\}, j=1,2,...,m$, for continuous and discrete case. In the second case, the choice of parameters $x_1,x_2,...,x_n,y_1,y_2,...,y_m$ means choosing a sub-table from training table. Here t serves as a parameter that is not involved in the optimization process directly.

3. Local step of optimization.

Consider the proposed method of local optimization algorithm. Fix some $t = t_0 = const$. Then the functional takes the following form:

$$f(\mathbf{x},\mathbf{y},t_0) = \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}(t_0)x_i y_j = F(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}x_i y_j \to \max_{\mathbf{x}\in X,\mathbf{y}\in Y}$$

Let $\tilde{\mathbf{x}} = (\tilde{x}_1,...,\tilde{x}_n) \in \{0,1\}, \tilde{\mathbf{y}} = (\tilde{y}_1,...,\tilde{y}_m) \in \{0,1\}$ is a current point in the process of optimization, $d(\mathbf{x},\tilde{\mathbf{x}})$ is the Hamming distance between vectors. Introduce the following notation: $O_{k_x k_y}(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = \{\mathbf{x},\mathbf{y} : d(\mathbf{x},\tilde{\mathbf{x}}) = k_x, d(\mathbf{y},\tilde{\mathbf{y}}) = k_y\}$, i.e. $O_{k_x k_y}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ is the set of vectors $\mathbf{x},\mathbf{y}$ that they differ in exactly k_x feature vectors and k_y object vectors. Denote:

$$F^{20}(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = \max_{O_{20}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})} F(\mathbf{x},\mathbf{y}),\ \ F^{02}(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = \max_{O_{02}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})} F(\tilde{\mathbf{x}},\mathbf{y}),\ \ F^{11}(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = \max_{O_{11}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})} F(\mathbf{x},\mathbf{y})$$

$$u_j(\mathbf{x}) = \sum_{i=1}^{n} C_{ij}x_i,\ \ v_i(\mathbf{y}) = \sum_{j=1}^{m} C_{ij}y_j$$

Next, show that each of the values $F^{20}(\tilde{\mathbf{x}},\tilde{\mathbf{y}}), F^{02}(\tilde{\mathbf{x}},\tilde{\mathbf{y}}), F^{11}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ can be calculated with $O(n+m)$ time , so, the step of local extremum search in $O^2(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ has linear complexity with respect to the sum of number of features and number of objects. Consider three possible cases

1) $\mathbf{x},\tilde{\mathbf{y}} \in O_{20}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$, where $x_i = \tilde{x}_i, i \notin \{i_1,i_2\}$. It's obvious that $x_i \in \{0,1\}, i \in \{i_1,i_2\}$. This implies:

$$\Delta F(\mathbf{x},\tilde{\mathbf{y}}) = F(\mathbf{x},\tilde{\mathbf{y}}) - F(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = v_{i_1}(\tilde{\mathbf{y}})(x_{i_1} - \tilde{x}_{i_1}) + v_{i_2}(\tilde{\mathbf{y}})(x_{i_2} - \tilde{x}_{i_2})$$

$$\varphi_1 = \min_{\mathbf{x},\tilde{\mathbf{y}}\in O^2(\tilde{\mathbf{x}},\tilde{\mathbf{y}})} \Delta F(\mathbf{x},\tilde{\mathbf{y}})$$

2) Similarly, when $\tilde{\mathbf{x}},\mathbf{y} \in O_{02}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$, $y_i = \tilde{y}_i, i \notin \{i_1,i_2\}$, $y_i \in \{0,1\}, i \in \{i_1,i_2\}$

$$\Delta F(\tilde{\mathbf{x}},\mathbf{y}) = F(\tilde{\mathbf{x}},\mathbf{y}) - F(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = u_{i_1}(\tilde{\mathbf{x}})(y_{i_1} - \tilde{y}_{i_1}) + u_{i_2}(\tilde{\mathbf{x}})(y_{i_2} - \tilde{y}_{i_2})$$

$$\varphi_2 = \min_{\tilde{\mathbf{x}},\mathbf{y}\in O^2(\tilde{\mathbf{x}},\tilde{\mathbf{y}})} \Delta F(\tilde{\mathbf{x}},\mathbf{y})$$

Obviously, the search of φ_1 and φ_2 values has complexity $O(n+m)$.

3) Third case where $\mathbf{x},\mathbf{y} \in O_{11}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$

$$x_i = \begin{cases} \tilde{x}_i, & i \neq i_1, \\ 1-\tilde{x}_i, & \text{otherwise.} \end{cases} \quad y_i = \begin{cases} \tilde{y}_i, & i \neq i_2, \\ 1-\tilde{y}_i, & \text{otherwise.} \end{cases}$$

where i_1, i_2 is an arbitrary fixed coordinate pair. In this case:

$$\Delta F(\mathbf{x},\mathbf{y}) = F(\mathbf{x},\mathbf{y}) - F(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = v_{i_1}(\tilde{\mathbf{y}})(x_{i_1} - \tilde{x}_{i_1}) + u_{i_2}(\tilde{\mathbf{x}})(y_{i_2} - \tilde{y}_{i_2}) + C_{i_1 i_2}(x_{i_1} - \tilde{x}_{i_1})(y_{i_2} - \tilde{y}_{i_2})$$

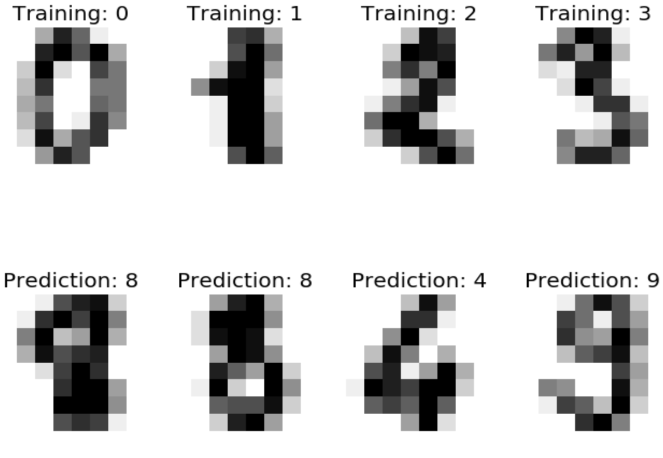
$$\Delta F(\mathbf{x},\mathbf{y}) \le v_{i_1}(\tilde{\mathbf{y}})(x_{i_1} - \tilde{x}_{i_1}) + u_{i_2}(\tilde{\mathbf{x}})(y_{i_2} - \tilde{y}_{i_2}) + \max_{i_2}\left|C_{i_1 i_2}\right| = \Delta_1 F(\mathbf{x},\mathbf{y})$$

$$\Delta F(\mathbf{x},\mathbf{y}) \le v_{i_1}(\tilde{\mathbf{y}})(x_{i_1} - \tilde{x}_{i_1}) + u_{i_2}(\tilde{\mathbf{x}})(y_{i_2} - \tilde{y}_{i_2}) + \max_{i_1}\left|C_{i_1 i_2}\right| = \Delta_2 F(\mathbf{x},\mathbf{y})$$

Obviously, in this case the bust of all $\Delta_1 F(\mathbf{x},\mathbf{y}), \Delta_2 F(\mathbf{x},\mathbf{y})$ values and search for the optimum has complexity $O(n+m)$. In the case $\mathbf{x},\mathbf{y} \in O^1(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ the complexity is linear for obvious reasons. Finally, the total complexity of the algorithm is $O(n+m)$.

4. Results of numerical experiments

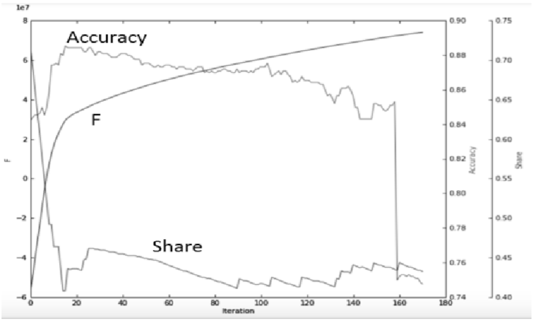
In this paper we tested the approach on different datasets. The first considered set is named "Digits" and was taken from scikit-learn [Pedregosa et al., 2011] database. It is dedicated to the problem of handwritten digits classification. The set consists of 1797 objects that represent 8x8 black-and-white pictures of numbers. Thus, each object has 64 features – color intensity at each pixel. All objects are divided into 10 classes.



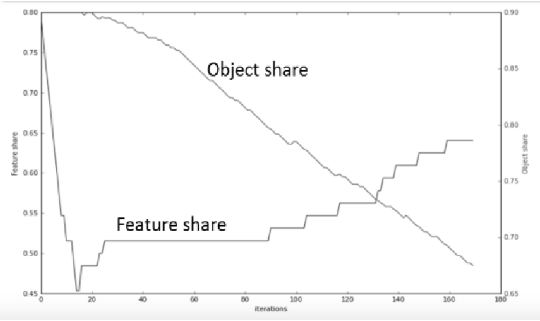
During the experiment, the dependency of classification quality (accuracy, or share of correctly recognized objects) on the F value, and share of saved features and objects, with different parameters t was considered.

During each experiment a certain proportion of the component feature and object vectors are set as 0. Next, a greedy optimization process described above is run. The local step each time shifts to the most optimal point of the neighborhood. Below, on the left graph 3 plots are shown: share of new table to the old, quality of the classifier while training on the reduced data, and F - the value of optimized functional at certain optimization iteration. On the right side the proportion of saved attributes and objects in each iteration is shown.

Experiment 1, $\frac{\sum x_i}{|x_i|} \approx 0.8$, $\frac{\sum y_i}{|y_i|} \approx 0.9, t = 0.23$:

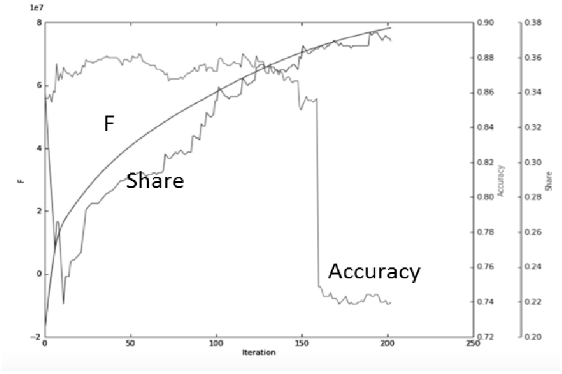


Accuracy, F and Table Share

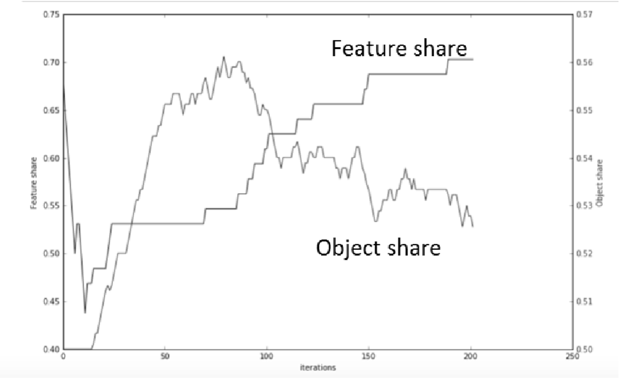


Feature share and object share

Experiment 2, $\frac{\sum x_i}{|x_i|} \approx 0.67$, $\frac{\sum y_i}{|y_i|} \approx 0.5, t = 0.23$:

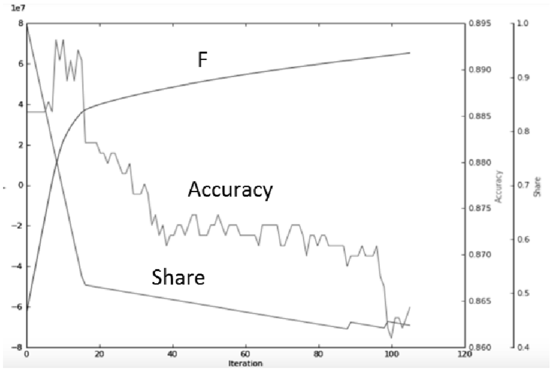


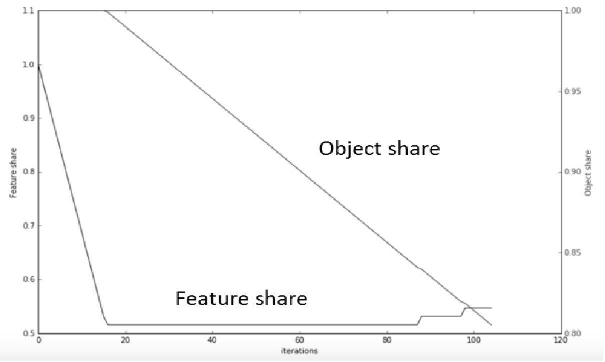
Accuracy, F and Table Share



Feature share and object share

Experiment 3, $\frac{\sum x_i}{|x_i|} = 1$, $\frac{\sum y_i}{|y_i|} = 1, t = 0.23$:





Accuracy, F and Table Share

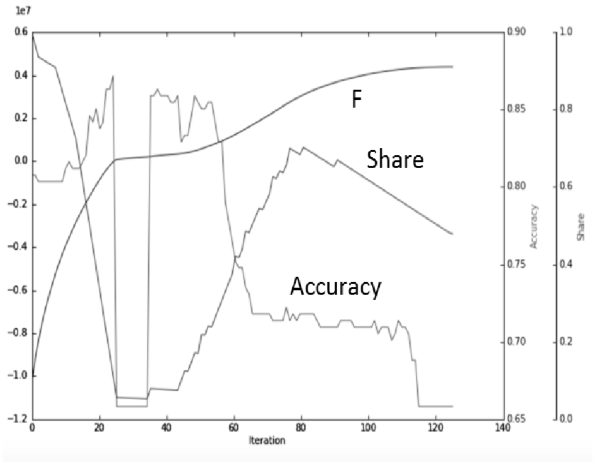
Feature share and object share

Following patterns can be seen during each experiment: after a certain number of iterations, an optimal sub-sample appears, which size is substantially smaller comparing to the original and at the same time the quality of the classification decreases slightly or even increases. Considering the right plot one can see that objects and features are examined by the algorithm simultaneously.

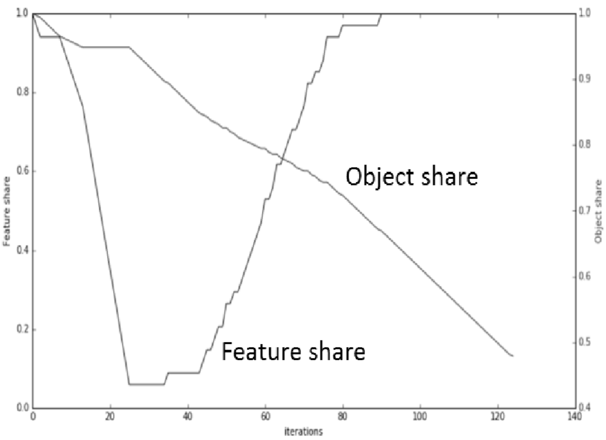
The second problem called "ionosphere", is taken from the repository [Lichman, 2013]. A system of 16 high-frequency antennas explores the properties of the ionosphere. The task is to distinguish between two types of signals - "good", having free electrons and carrying useful information of the structure of the ionosphere and "bad", which pass through the ionosphere without reflection. The electromagnetic signals are characterized by a set of 17 pulsations, each of which has two attribute, so the number of features equals 34. Two tables are given – training and validating, in each of the tables approximately 200 objects of each class persist.

The experiment also demonstrates simultaneous selection of attributes and objects and in search of an optimal iteration, also accuracy is risen comparing to baseline and training table size is significantly reduced both in number of features and the number of objects.

Experiment 4, $\frac{\sum x_i}{|x_i|}=1$, $\frac{\sum y_i}{|y_i|}=1, t=4$:



Accuracy, F and Table Share

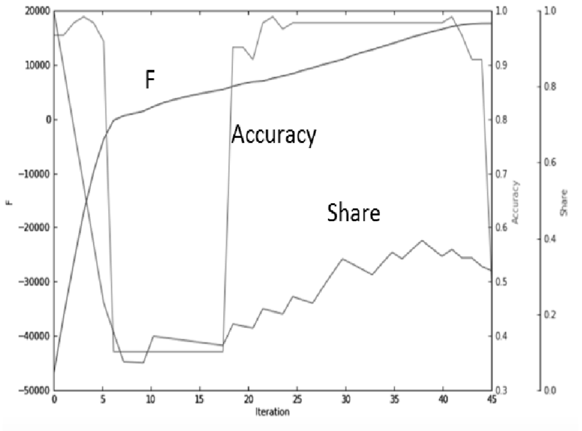


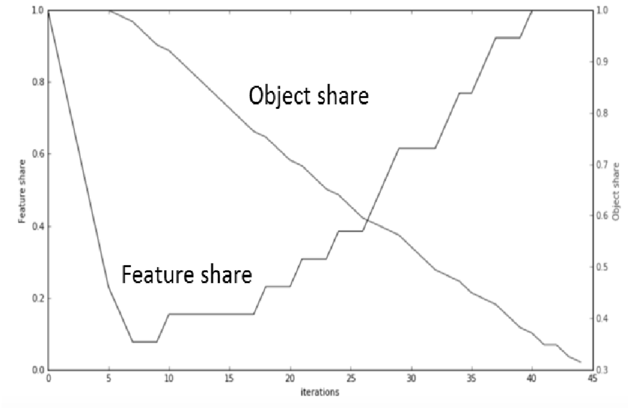
Feature share and object share

And the last of the tasks - the problem of classifying wines by chemical analysis. Sample objects are the result of chemical analysis, expressed in the 13 symptoms, such as alcohol content, malic acid, magnesium, and other hue. The table is divided into 3 classes corresponding to 3 grades of wine made from grapes grown in the same region of Italy.

Similarly to the previous classification problem quality is improved by reducing the training sample. In the last two experiments, an interesting pattern appears - at some iteration the quality drops sharply and then returns to the original values. This is due to the removal "by mistake" some important features, and then returning them back to the task.

Experiment 5, $\frac{\sum x_i}{|x_i|}=1, \frac{\sum y_i}{|y_i|}=1, t=3.3$:





Accuracy, F and Table Share

Feature share and object share

5. Acknowledgements

This work was supported by program RFBR № 16-31-00443 mol_a.

6. Conclusion

These experiments demonstrate the ability of the algorithm to solve its initial task - simultaneous selection of features and objects, thereby it leads to training set reduction and improves the quality of the classification. This result was obtained on different data and different starting parameter options.

It should also be noted that the proposed method of local optimization is a cheap procedure, thereby it allows to search for optimal subsamples faster, accelerating the speed of learning.

Bibliography

[Lei Yu et al., 2003] L. Yu, H. Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 856-863, 2003

[Norbert Jankowski et al., 2005] N. Jankowski, K. Grabczewski. Feature Selection with Decision Tree Criterion. Hybrid Intelligent Systems, International Conference on, vol. 00, no. , pp. 212-217, 2005

[P. Pudil et al., 1994] P. Pudil, J. Novovičová, J. Kittler. Floating search methods in feature selection. Pattern Recognition Letters, vol. 15, issue 11, pp. 1119-1125, 1994

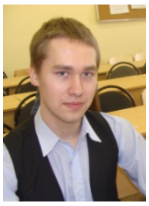
[P. Hart, 1968] P. Hart. The condensed nearest neighbor rule. IEEE Transactions on Information Theory, vol. 14, issue 3, pp. 515-516, 1968

[Zhuravlev, 1978] Yu.I. Zhuravlev. On the algebraic approach to solving the problems of recognition and classification. Problems of Cybernetics. M .: Nauka, Issure.33. pp. 5-68, 1978.

[Pedregosa et al., 2011] F. Pedregosa et al. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830., 2011

[Lichman, 2013] UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Authors' Information



Vasily Ryazanov *– PhD student; Moscow Institute of Physics and Technologies, Russia, 141700, Moscow reg., Dolgoprudny, Institutsky per. 9; e-mail: vasyarv@mail.ru*

Major Fields of Scientific Research: Data Mining, Missing Data, Multiclass Classification