Yugoslav Journal of Operations Research
29 (2019), Number 2, 177–192
DOI: https://doi.org/10.2298/YJOR180822032Y

GROUP APPROACH TO SOLVING THE TASKS OF RECOGNITION

Yedilkhan AMIRGALIYEV
Institute of Information and Computational Technologies, Suleyman Demirel University, Almaty
amir_ed@mail.ru

Vladimir BERIKOV
Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Novosibirsk State University
berikov@math.nsc.ru

Lyailya S. CHERIKBAYEVA
Alfarabi Kazakh National University, Almaty
lyailya_sh@mail.ru

Konstantin LATUTA
Suleyman Demirel University, Almaty
konstantin.latuta@sdu.edu.kz

Kalybekuuly BEKTURGAN
Institute of Automation and Information Technology, Kyrguz Republic
yky198@mail.ru

Received: July 2018 / Accepted: November 2018

Abstract: In this work, we develop CASVM and CANN algorithms for semi-supervised classification problem. The algorithms are based on a combination of ensemble clustering and kernel methods. A probabilistic model of classification with the use of cluster ensemble is proposed. Within the model, error probability of CANN is studied. Assumptions that make probability of error converge to zero are formulated. The proposed algorithms are experimentally tested on a hyperspectral image. It is shown that CASVM and CANN are more noise resistant than standard SVM and kNN.

Keywords: Recognition, Classification, Hyper Spectral Image, Semi-Supervised Learning.

MSC: 90B85, 90C26.

1. INTRODUCTION

In recent decades, there has been a growing interest in machine learning and data mining. In contrast to classical methods of data analysis, in this area much attention is paid to modeling human behavior, solving complex intellectual problems of generalization, revealing patterns, finding associations, etc. The development of this area was boosted by the ideas arising from the theory of artificial intelligence.

The goal of pattern recognition is to classify objects into several classes. A finite number of features describe each object. Classification is based on precedents; the objects, for which the classes they belong to are known. In classical supervised learning, the class labels are known for all the objects in the sample. New objects are to be recognized as belonging to the one of the known classes. Many problems arising in various areas of research can be reduced to problems of classification.

In classification problems, group methods are widely used. They consist in the synthesis of results obtained by applying different algorithms to a given source information, or in selection of optimal, in some sense, algorithms from a given set. There are various ways of defining group classifications. The formation of recognition as an independent scientific theory is characterized by the following stages:

- the appearance of large number of various incorrect (heuristic) methods and algorithms to solve practical problems, oftentimes applied without any serious justification;
- the construction and research of collective (group) methods, providing a solution to the problem of recognition based on the results;
- processing of initial information by separate algorithms [1-4].

The main goal of cluster analysis is to identify a relatively small number of groups of objects that are as similar as possible within the group, and as different as possible from other groups. This type of analysis is widely used in information systems when solving problems of classification and detection of trends in data: when working with databases, analyzing Internet documents, image segmentation, etc. At present, a sufficiently large number of algorithms for cluster analysis have been developed. The problem can be formulated as follows. There is a set of objects described by some features (or by a distance matrix). These objects are to be partitioned into a relatively small number of clusters (groups, classes) so that the grouping criterion would take its best value. The number of clusters can be either selected in advance or not specified at all (in the latter case, the optimal number of clusters must be determined automatically). A quality criterion usually is understood as a certain function, depending on the scatter of objects within the group and the distances between groups.

By now, considerable experience has been accumulated in constructing both separate taxonomic algorithms and their parametric models. Unlike the recognition problems in related areas, in this area universal methods for solving taxonomic problems have not yet been created, and the current ones are generally heuristic.

Current methods include: the construction of classes, based on the allocation of compact groups; separation of classes using separating surfaces; the construction of classes using auxiliary "masks", "signatures". The main criteria that determine the quality of classification based on the natural definition of the optimal partition are the following: the compactness of the classes to be formed, the separability of classes and the classification stability of objects forming the class.

Recently, cluster analysis has been actively developing an approach based on collective decision-making. It is known that algorithms of cluster analysis are not universal: each algorithm has its own specific area of application: for example, some algorithms can better cope with problems in which objects of each cluster are described by "spherical" regions of multidimensional space; other algorithms are designed to search for "tape" clusters, etc. In the case when the data are of a heterogeneous nature, it is advisable to use not one algorithm but a set of different algorithms to allocate clusters. The collective (ensemble) approach also makes it possible to reduce the dependence of grouping results on the choice of parameters of the algorithm, to obtain more stable solutions in the conditions of "noisy" data, if there are "omissions" in them [5-9].

Ensemble approach allows improving the quality of clustering. There are several main directions in the methods of constructing ensemble solutions of cluster analysis: based on the consensus distribution, on the co-associative matrices, on the models of the mixture of distributions, graph methods, and so on. There are a number of main methods for obtaining collective cluster solutions: the use of a pairwise similarity/difference matrix; maximization of the degree of consistency of decisions (normalized mutual information, Adjusted Rand Index, etc.) Each cluster analysis algorithm has some input parameters, for example, the number of clusters, the boundary distance, etc. In some cases, it is not known what parameters of the algorithm work best. It is advisable to apply the algorithm with several different parameters rather than one specific parameter.

In this work semi-supervised learning is considered. In semi-supervised learning, the class labels are known only for a subset of objects in the sample. The problem of semi-supervised learning is important for the following reasons:

- Unlabeled data are cheap;
- Labeled data may be difficult to obtain;
- Using unlabeled data along together with labeled data may increase the quality of learning.

There are many algorithms and approaches to solve the problem of semi-supervised learning [10]. The goal of the work is to devise and test a novel approach to semi-supervised learning. The novelty lies in the combination of algorithms of collective cluster analysis [11,12] and kernel methods (support vector machines SVM [13] and nearest neighbor NN), as well as in theoretical analysis of the error probability of the proposed method. In the coming sections, a more formal problem statement will be given, some cluster analysis and kernel methods will be reviewed, the proposed methods will be described, and its theoretical and experimental ground will be provided.

2. FORMAL PROBLEM STATEMENT

2.1. Formal Problem Statement of Semi-Supervised Learning

Suppose we have a set of objects X to classify and finite set of class labels Y. All objects are described by features. A feature of an object is the following mapping f:X$\rightarrow D_f$, where D_f - set of values of a feature.

Depending on D_f features can be of the following types:

- Binary features: D_f =$\{0,1\}$.
- Numerical features: D_f = R.
- Nominal features: D_f - finite set.
- Ordered features: D_f - finite ordered set.

For a given feature vector f_1,...,f_m, vector x = ($f_1(\alpha)$,...,$f_m(\alpha)$) is called feature descriptor of object $\alpha \in$ X. Further, in the text we do not distinguish between an object and its feature descriptor. In the problem of semi-supervised learning at the input we have a sample $X_N = \{x_1,...,x_n\}$ of objects from X.

There are two types of objects in the sample:

- $X_c = \{x_1,...,x_k\}$ - labeled objects with the classes they belong to: $Y_c = \{y_1,...,y_k\}$;
- $X_u = \{x_{k+1},...,x_n\}$ - unlabeled objects.

There are two formulations of the classification problem statement. In the first, we are to conduct so-called inductive learning, i.e. build a classification algorithm a = X$\rightarrow$Y, which will classify objects from X_u and the new objects from X_{test}, which were unavailable at the time of building of the algorithm.

The second is so-called transductive learning. Here we get labels only for objects from X_u with minimal error. In this work, we consider the second variant of problem statement.

The following example shows how semi-supervised learning differs from a supervised learning.

Example: Label objects are given at the input X_c =$\{x_1,...,x_k\}$ with their respective classes $Y_c\{y_1,...,y_k\}$, where $y_1 \in \{0,1\}$, y_i =1,...,k. The objects have two features and their distribution is shown in Figure 1.

Unlabeled data is also given $X_u = \{x_{k+1},...,x_n\}$ as shown in Figure 2.

Suppose that a sample from a mixture of normal distributions is given. Let's estimate the density of the classes throughout the data set at only on the labeled data, after which we construct the separating curves. Then, from Figure 3 it can be seen that the quality of the classification using the full set of data is higher.

2.2. Ensemble Cluster Analysis

In the problem of ensemble cluster analysis, several partitions (clustering) S^1, S^2, Sr are considered. They may be obtained by:

- the results of various algorithms for cluster analysis;

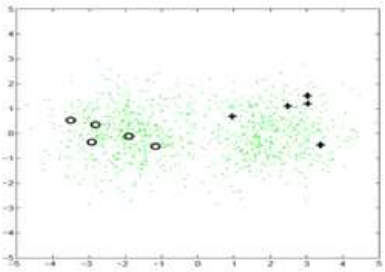

Figure 1: Features of objects

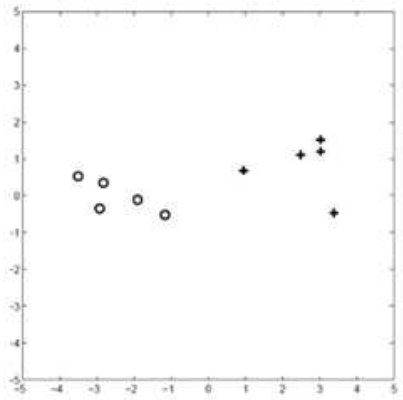

Figure 2: Labeled objects X_c with unlabeled objects X_u

- the results of several runs of one algorithm with different parameters.

For example, Figure 4 shows examples of different partitions for 4 sets. Different colors correspond to different clusters.

To construct a matrix of average differences, clustering of all available objects X = $\{x_1,...,x_N\}$ is done by an ensemble of several different algorithms $\mu_1,...,\mu_M$. Each algorithm gives L_m variants of partition, $m = 1, ..., M$. Based on the results of the algorithms, a matrix H of average differences is built for objects of X. The matrix elements are equal to:

$$h(i,j) = \sum_{m-1}^{M} \alpha_m \frac{1}{L_m} \sum_{i-1}^{L_m} h_m(i,j) \tag{1}$$

where $i, j \in \{1, ..., N\}$ - objects' numbers $(i \neq j), \alpha_m \geq 0$ - initial weights so that $\sum\limits_{m-1}^{M} \alpha_m = 1; h_m(ij) = 0$, if pair (ij) belong to different clusters in l - h variant of partition, given by algorithms μ_m and 1, if it belongs to the same cluster.

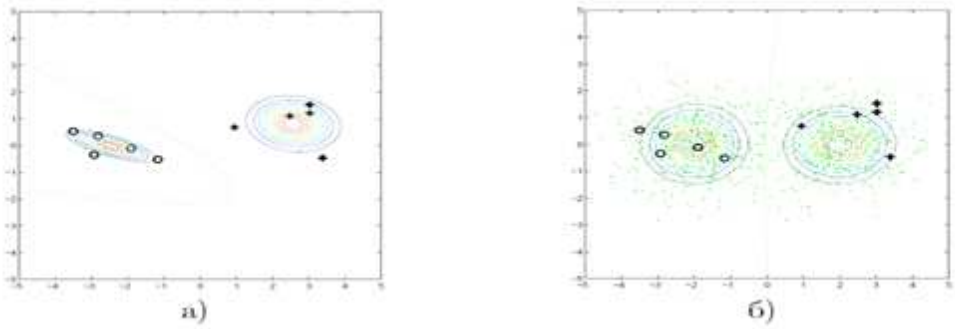

Figure 3: Obtained class densities:) - by labeled data; b) -by unlabeled data

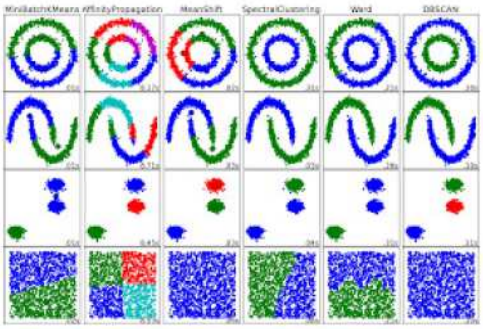

Figure 4: Examples of various distributions for 4 classes

Weights α_m may be same or, for example, may be set with respect to quality of each clustering algorithm. The selection of optimal weights is researched in [6].

The results of the ensemble work can be presented in the form of the following table 1, where for each partition and for each point the assigned cluster number is stored [2].

Table 1: Ensemble work

In this work semi-supervised learning is considered. In semi-supervised learning the classes are known only for a subset of objects in the sample. The problem of semi-supervised learning is important for the following reasons:

- unlabeled data is cheap;
- labeled data may be difficult to obtain;
- using unlabeled data along with some labeled data may increase the quality of learning.

There are many algorithms and approaches to solve the problem of semi-supervised learning [10]. The goal of the work is to devise and test a novel approach to semi-supervised learning. The novelty lies in the combination of algorithms of

	Cluster S^1	Cluster S^2	...	Cluster S^r
x_1	1	2		1
x_2	4	4		4
x_3	2	2		1
...				
x_n	3	2		3

collective cluster analysis [11,12] and kernel methods (support vector machines SVM [13] and nearest neighbor NN), as well as in theoretical analysis of the error of the proposed method. In the coming sections a more formal problem statement will be given, some cluster analysis and kernel methods will be reviewed, the proposed methods will be described, and its theoretical and experimental ground will be provided.

Cluster ensembles combine multiple clusters of a set of objects into one consolidated clustering, often called a consensus solution.

3. KERNEL METHODS OF CLASSIFICATION

To solve the classification problem, kernel methods are widely used, based on the so-called "kernel trick". To demonstrate the essence of this "trick", consider the support vector machine method (SVM) - the most popular kernel method of classification. SVM is a binary classifier, although there are ways to refine it for multiclassification.

3.1. Binary Classification with SVM

In the problem of dividing into two classes (the problem of binary classification), a training sample of objects $X = \{x_1, ..., x_n\}$ is at the input with classes $Y = \{y_1, ..., y_n\}$, $y_1 \epsilon \{+1, -1\}$, for $i = 1, ..., n$, where object are points in m - dimensional space of feature descriptors. We are to divide the points by hyperplane of dimension $(m-1)$. In the case of linear class separability, there exist an infinite number of separating hyperplanes. It is reasonable to choose a hyperplane, the distance from which to both classes is maximized. An optimal separating hyperplane is a hyperplane that maximizes the width of the dividing strip between classes. The problem of the support vector machine method consists in constructing an optimal separating hyperplane. The points lying on the edge of the dividing strip are called support vectors.

A hyperplane can be represented as $< w, x > +b = 0$, where $<,>$ - scalar product, w - vector perpendicular to separating hyperplane, and b - an auxiliary parameter. Support vector method builds decision function in in the form of

$$F(x) = sign(\sum_{i-1}^{n} \lambda_i c_i < x_1, x > +b) \tag{2}$$

It is important to note that the summation goes only along support vectors for which $\lambda_i \neq 0$. Objects $x \epsilon X$ with $F(x) = 1$ will be assigned one class, and objects with $F(x) = 0$ another.

With linear inseparability of classes, one can perform a transformation $\varphi : X \rightarrow G$ of object space X to a new space G of a higher dimension. The new space is called is called "rectifying", because the objects in the space can already be linearly separable.

Decision function $F(x)$ depends on scalar products of objects, rather that the objects themselves. That is why scalar products $< x, x' >$ can be substituted by products of $< \varphi(x), \varphi(x') >$ kind in the space G. In this case the decision function $F(x)$ will look like this:

$$F(x) = sign(\sum_{i-1}^{n} \lambda_i c_i < (\varphi x_1), \varphi(x) > +b) \tag{3}$$

Function $K(x, x') =< \varphi(x), \varphi(x') >$ is called kernel. The transition from scalar products to arbitrary kernels the "kernel trick". Selection of the kernel determines the rectifying space and allows to use linear algorithms (like SVM) to linearly non-separable data.

3.2. Mercer Theorem

Function K, defined on a finited set of objects X, can be set as $\mathrm{K} = (K(x_i, x_j))$, where $x_i, x_j \epsilon X$. In kernel classification methods, a theorem is widely known that establishes a necessary and sufficient condition for the matrix to define a certain kernel:

Theorem (Mercer). Matrix $\mathrm{K} = (K(x_i, x_j))$ of size $p \times p$ is the kernel matrix if and only if it symmetric $K(x_i, x_j) = K(x_j, x_i)$ and nonnegatively defined: for any $z \epsilon R^p$ the following condition holds: $z^T \mathrm{K} z \geq 0$.

4. PROPOSED METHOD

The idea of the method is to construct a similarity matrix (1) of all objects from the input sample X. The matrix will be compiled by applying different clustering algorithms to X. The more a pair of objects are classified as belonging to one class the more similar they will be. Two possible variants of prediction for unlabeled classes X_u will be proposed using similarity matrix. Further the idea of the algorithms will be described in detail. The following theorem holds:

Theorem 1. Let $\mu_1, ..., \mu_M$ - be algorithms of clustering analysis , each algorithm gives L_m variants of partition, $m = 1, ..., M, h_{lm}(x, x') = 0$, if a pair of objects (x, x') belongs to different clusters in l - th variant of partition, given by algorithm μ_m and 1, if it belongs to the same cluster. $\alpha_m \geq 0$ - initial weights such that $\sum\limits_{m-1}^{M} \alpha_m = 1$. Then function $H(x, x') = \sum\limits_{m-1}^{M} \alpha_m \frac{1}{L_m} h_{lm}(x, x')$ satisfies the condition of Mercer theorem.

Proof. It is obvious that function $H(x, x')$ symmetric. Let C_r^{lm} - be the set of indices of objects that belong to r - th cluster, given by m - th algorithmin l - th variant of partition. Let's show that $H(x, x')$ nonnegatively defined.

Let take arbitrary $z \epsilon R^P$ and show that $z^T H z \geq 0$

$$z^T H z = \sum_{i.j-1}^{p} \sum_{m-1}^{M} \alpha_m \frac{1}{L_m} \sum_{l-1}^{L_m} h_{lm}(i,j) z_i z_j = \sum_{m-1}^{M} \alpha_m \frac{1}{L_m} \sum_{l-1}^{L_m} \sum_{i.j-1}^{P} h_{lm}(i,j) z_i z_j =$$

$$\sum S_{m-1}^{M} \alpha_m \frac{1}{L_m} \sum_{l-1}^{L_m} (\sum_{i.j \epsilon C_l^{lm}} z_i z_j + ... + \sum_{i.j \epsilon C_{K_{lm}}^{lm}} z_i z_j) = \tag{4}$$

$$\sum_{m-1}^{M} \alpha_m \frac{1}{L_m} \sum_{l-1}^{L_m} ((\sum_{io C_l^{lm}} z_i)^2 + ... + (\sum_{ie C_{K_l m}^{lm}} z_i)^2) \geq 0.$$

Thus, function $H(x, x')$ can be used as a kernel in kernel methods of classification. For instance, in support vector machines (SVM) and in nearest neighbor method (NN). Further, the two variants of the algorithm that implement the proposed method are described:

Algorithm CASVM

Input: objects X_c with their classes Y_c and objects X_u, number of clustering algorithms M, number of clustering L_m by each algoritm $\mu_m, m = 1, ..., M$

Output: classes of objects X_u.

1. Cluster objects $X_c \cup X_u$ by algorithms $\mu_1, ..., \mu_M$, and get L_m variants of partitions from each algorithm $\mu_m, m = 1, ..., M$.
2. Computer matrix H for $X_c \cup X_u$ by formula (1).
3. Train SVM with labeled data X_c, using matrix H as kernel.
4. By means of SVM predict classes of unlabeled data X_u.

End of algorithm

Algorithm CANN

Input: objects X_c with given classes Y_c and objects X_u, number of clustering algorithms M, number for clusters L_m

Output: classes of objects X_u.

1. Cluster objects $X_c \cup X_u$ Cluster objects by algorithms $\mu_1, ..., \mu_M$, get L_m variants of partitions from each algorithm $\mu_m, m = 1, ..., M$.

2. Compute H for $X_c \cup X_u$ by formula (1).

3. Use NN: for each unlabeled object $x \epsilon X_u = \{x_{k+1}, ..., x_N\}$ assign the most similar class in sense $H(x, x`)$ of labeled object $x` \epsilon X_c = \{x_1, ..., x_k\}$.

Formally written: $x_i = argmaxH(x_i, x_j), i = k + 1, ..., Nj = 1...k$. **End of algorithm**

Note that in the proposed algorithms there is no need to store matrix H in memory $N \times N$ entirely: it is enough to store the clustering matrix of size $N \times L$, where $L = \sum_{l-1}^{M} L_m$, in this case H can be computed dynamically. In practice, $L << N$, for example, when working with image pixels.

5. THEOETICAL ANALYSIS OF THE METHOD

Let's recall the problem statement. At the input we have sample of objects $X_N\{x_1, ..., x_N\}$. There are two types of objects in the sample:

$X_c = \{x_1, ..., x_k\}$ - labeled objects with classes $Y_c = \{y_1, ..., y_k\}, I_c = \{1, ..., k\}$ - object indices

$X_u = \{x_{k+1}, ..., x_N\}$ - unlabeled objects, $I_u = \{k + 1, ..., N\}$ - indices of the objects.

For simplicity, suppose that the number of different algorithms in the ensemble is algorithms is $M = 1$, i.e. the algorithms $\mu = \mu_1$ makes $L = L_1$ clasterizations according to parameters $\Omega_1, ..., \Omega_L$, that are chosen from the given set. Let us consider these parameters as independent and equally distributed random variables.

Let's introduce the following notations for $x_i, x_j \epsilon X_N$:

$h_l(x_i, x_j) = \{1$, if algorithm μ in variant l unites the pair $(x_i x_j)$ 0 - otherwise$\}$

And the following quantities $L_1(x_i x_j) = \sum_{l=1}^{L} h_l(x_i x_j)$, $L_0(x_i x_j) = L - L_1(x_i x_j)$, which are the number of variants in whcih the algorithm voted for the union of pair $(x_i x_j)$, or against it, respectively. Let $Y(x)$ - be hidden from us true labels of unlabeled objects $x \epsilon X_u$.

Let's introduce a random variable:

$$Z(x_i x_j) = \begin{cases} \{1, if Y(x_i) = Y(x_j) \\ 0 if Y(x_i) \neq Y(x_j) \end{cases} \tag{5}$$

Denote

$$q_0(x_i x_j) = P[h_1(x_i x_j) = 0 | Z(x_i x_j) = 0], q_1(x_i x_j) = P[h_1(x_i x_j) = 1 | Z(x_i x_j) = 1]$$

conditional probabilities of correct decision for the partition (union) of the pair into the same cluster (different clusters) correspondently.

Let us assume that parameters $\Omega_1, ..., \Omega_L$ stay independent under any value of $Z(x_i x_j)$. Let us consider any arbitrary pair of points $x = \epsilon X_u$ and $x`\epsilon X_c$. Let the algorithms of the ensembl assign the pair to one cluster by majority of votes, and the object $x$ is given label $y = y`$, where $y`$ is class label, corresponding to object $x`$.
The following theorem holds.

Theorem 2. Assume for any element of the ensemble the following condition holds: $COV_{Z,\Omega_1,...,\Omega_L}[Z(x, x`), h_1(x, x`)] > 0, l\epsilon\{1, ..., L\}$. Then under the above assumption of the model, conditional probability of error classification of point x tends to zero, when $L_1(x, x`) \to \infty$ and $L_0(x, x`) = const$.

The last condition means that the overwhelming majority of voices in the ensemble are given for the unification of this pair into one cluster. The condition of positivity of covariance implies that the clustering algorithm tends to make a correct decision with respect to a given pair of points. The proof of Theorem 2 is given in the Appendix.

Corollary. Let the following holds for a pair of points : $q_0(x, x`) > \frac{1}{2}$, $q_1(x, x`) > \frac{1}{2}$. Then $P_{er}(x) \to 0$ under $L_1 \to \infty$.

Proof. Let's show, that under the given conditions the following holds: $cov[Z(x_i, x_j), h(x_i, x_j)] > 0$. Let's omit arguments x_i, x_j for simplicity of writing. Thus we have:
$q_0 \frac{p_{00}}{p_{00}+p_{01}} > \frac{1}{2}$, therefore $p_{00} > p_{01}$; similarly from $q_1 = \frac{p_{11}}{p_{10}+p_{11}} > \frac{1}{2}$ follows that $p_{11} > p_{10}$. According to Bernulli distribution property, $cov[Z, h] = p_{00}p_{11} - p_{01}p_{10}$. It means $cov[Z, h] > 0$.

The corollary shows that the probability of classification error tends to zero under the assumption that the used algorithms correctly assign pairs of objects to one or different clusters with probability more 1/2, i.e., do not guess.

6. EXPERIMENTAL SETUP

A typical RGB image contains three channels: the intensity values for each of the three colors. In some cases, this is not enough to get complete information about the characteristics of the object being shot. To obtain data on the properties of objects that are indistinguishable by the human eye, hyper spectral images are used.

For an experimental analysis of the developed algorithm, we used picture Salinas-A [17]. The image was collected by the 224-band AVIRIS sensor over Salinas Valley, California. The image size is 83 x 86; each pixel is characterized by the vector of 204 spectral intensities; the image spatial resolution is 3.7 m. The scene contains six types of vegetation and a bare soil. Figure 5a) illustrates the image: a grayscale representation is obtained from the 10th channel. Figure 5b) shows the ground truth image; different classes are presented with different colors

(the colors do not match any vegetation patterns; they are used just to distinguish between classes).

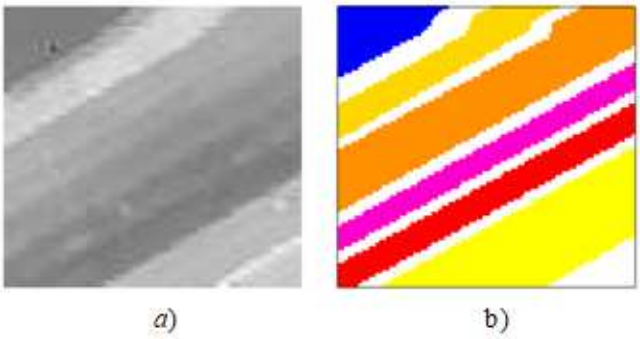

Figure 5: Salinas-A hyperspectral image: 10th channel (); groundtruth classes (b).

In an experimental analysis of the algorithm, 1% of the pixels selected at random for each class made up the labeled sample; the remaining ones were included in the unlabeled set. To study the effect of noise on the quality of the algorithm, randomly selected $r\%$ of the spectral brightness values of the pixels in different channels were subjected to a distorting effect: the corresponding value x was replaced by a random variable from interval $[x(1-p), x(1+p)]$, where r, p - initial parameters.

The noisy data table containing the spectral brightness values of the pixels across all channels was fed to the input of the CASVM algorithm, and the K-means algorithm was chosen as the basic algorithm for constructing the cluster ensemble.

Different variants of partitions were obtained by random selection of three channels. Number of clusters $K = 7$. To speed up the operation of the K-means algorithm and to obtain more diverse groupings, the number of iterations was limited to 1.

Since the proposed algorithm implements the idea of distance metric learning, it would be natural to compare it with a similar algorithm (SVM method), which uses the standard Euclidean metric, under similar conditions (the algorithm parameters recommended by default in Matlab environment).

Table 2 shows the accuracy values of the classification of the unlabeled pixels of the Salinas-A scene for some values of the noise parameters. The running time of the algorithm was about 3 seconds on a dual-core Intel Core i5 processor with a clock speed of 2.8 GHz and 4 GB of RAM. As it is shown in the table, CASVM algorithm has better noise resistance than SVM algorithm.

Table 2: Accuracy of CASVM and SVM under various noise values.

Parameters r, p	0%, 0	10%, 0.1	20%, 0.2	30%, 0.3
CASVM	0.742	**0.737**	**0.704**	0.639
CANN	0.703	0.703	0.694	**0.657**
SVM	**0.769**	0.709	0.693	0.387
kNN	0.710	0.671	0.517	0.362

7. CONCLUSION AND DISCUSSION

The paper has considered one of the variants of the problem of pattern recognition: the task of semi-supervised learning. The algorithms CASVM and CANN were developed to solve this problem. They use a combination of collective cluster analysis and kernel based classification. Both theoretical grounds and experimental confirmations of the usefulness of the proposed methodology were presented. The proposed combination allows one to use positive features of both approaches: receive stable decisions in noise conditions, in the presence of complex data structures.

In our theoretical study, we a) proved that the co-association matrix obtained with clustering ensemble is a valid kernel matrix and can be applied in kernel based classification; b) proved that the conditional probability of classification error for CANN tends to zero then increasing the number of elements in the ensemble, under the condition of positivity of covariance between ensemble decisions and the true status of the pair of data points. In the latter case, a probabilistic classification model for clustering ensemble was proposed. The suggested model expands theoretical concepts of classification and forecasting.

An experimental study of the proposed algorithms on a hyperspectral image was performed. It was shown that the CASVM and CANN algorithms are more noise-resistant than standard SVM and kNN.

Our theoretical investigation was limited by the assumed validity of a number of assumptions, such as: availability of independent random choice of clustering algorithms learning settings; positive covariance between clustering decisions and the true status of data points. Of course, the truthfulness of these assumptions can be criticized. In real clustering problems, the ensemble size is always finite and the assumptions lying at the basis of limit theorems can be violated. However, our study can be considered as a step to obtaining validating conditions which ensure success of semi-supervised methodology because it is a yet unsolved problem.

The authors plan to continue studying theoretical properties of clustering ensembles and their application in machine learning and data mining, in particular, for regression problems and hyperspectral image analysis. The designed methods will be used for genome-wide search for regulatory SNPs (rSNPs) associated with susceptibility to oncohematology diseases based on ChIP-seq and RNA-seq experimental data.

Acknowledgement: The work was carried out in accordance with the Memorandum on scientific and technical cooperation between the Sobolev Institute of mathematics of the SB RAS and the Institute of Information and Computing Technologies of the Ministry of Education and Science of the Republic of Kazakhstan. The research was carried out within the framework of the research program "Mathematical Methods of Pattern Recognition and Prediction" of the Math. S.L. Sobolev SB RAS and the project of grant financing of the GF INN 05132648 MES RK. The study was also partially supported by the RFBR grants 18-07-00600, 18-29-0904mk and partly by the Ministry of Science and Education of the Russian Federation within the framework of the 5-100 Excellence Program.

REFERENCES

[1] Amirgaliev, .N., Mukhamedgaliev, A.F., "On optimization model of classificatio n algorithms", *USSR Computational Mathematics and Mathematical Physics*, 25 (6)(1985) 95-98.
[2] Aidarkhanov, M.B., Amirgaliev, E.N., La, L.L., "Correctness of algebraic extensions of models of classification algorithms", *Cybernetics and Systems Analysis*, 37 (5) (2001) 777-781.
[3] Amirgaliyev, Y., Hahn, M., and Mussabayev, T., "The speech signal segmentation algorithm using pitch synchronous analysis", *Open Computer Science*, 7 (1) (2017) 1-8.
[4] Amirgaliyev, Y., Nusipbekov, A., Minsoo Hahn, "Kazakh Traditional Dance Gesture Recognition", *Journal of Physics: Conference Series*, 495 (4) (2014) 012036.
[5] Ghosh, J., Acharya, A., "Cluster ensembles", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1 (4) (2011) 305315.
[6] Domeniconi, C., and Al-Razgan, M., "Weighted cluster ensembles: Methods and analysis", *ACM Transactions on Knowledge Discovery from Data*, 2 (4) (2009) 17.
[7] Nimesha, M., Patil, Dipak, V., Patil, "A Survey on K-means Based Consensus Clustering", *IJETT*, 3 (1) (2016) 40-44.
[8] Topchy, A., Law, M., Jain, A., Fred, A., "Analysis of consensus partition in cluster ensemble", *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM04)*, (2004) 225-232.
[9] Vega-Pons, S., Correa-Morris, J., Ruiz-Shulcloper, J., "Weighted cluster ensemble using a kernel consensus function", *LNAI*, 5197 (2008) 195-202.
[10] Zhu, X., "Semi-supervised learning literature survey", Tech. Rep., Department of Computer Science, Univercity of Wisconsin, Madison, 2008.
[11] Berikov, V. B., "Weighted ensemble of algorithms for complex data clustering", *Pattern Recognition Letters*, 38 (2014) 99-106.
[12] Berikov, V., Pestunov, I., "Ensemble clustering based on weighted co-association matrices: Error bound and convergence properties", *Pattern Recognition*, 63 (2017) 27-436.
[13] Vapnik, V.N., "Restoration of dependencies according to empirical data", Information Statistics and Statistics, Springer, 1979. 448 p.

[14] Mercer, J., "Functions of positive and negative type and their connection with the theory of integral equations", Proceedings of the Royal Society, London, 1909.

Appendix

Proof of Theorem 2. Let $h_1^0(x,x`),...,h_L^0(x,x`)\epsilon\{0,1\}$ be the ensemble decisions for the pair. For short, let us skip arguments $x, x`$ until the proof end. Then the conditional probability of error in classifying x equals:

$$P_{er}(x) = P[Y(x) \neq Y(x`)|h_1 = h_1^0,...,h_L = h_L^0] =$$
$$= P[Z = 0|h_1 = h_1^0,...,h_L,...,h_L^0] = \tag{6}$$
$$= \frac{P[Z = 0, h_1 = h_1^0,...,h_L = h_L^0]}{P[h_1 = h_1^0,...,h_L = h_L^0]} =$$

$$= \frac{\prod\limits_{l=1}^{L_0} P[h_l = 0|Z = 0] \prod\limits_{l=1}^{L_1} [h_l = 1|Z = 0]P[Z = 0]}{\prod\limits_{l=1}^{L_0} P[h_l = 0] \prod\limits_{l=1}^{L_1} P[h_l = 1]} = \frac{q_0^{L_0}(1-q_0)^{L_1}P = [Z = 0]}{(P[h_l = 0])^{L_0}(P[h_l = 1])^{L_1}}$$

.

Let us denote

$p_{00} = P[Z = 0, h = 0], p_{01} = P[Z = 0, h = 1], p_{10} = P[Z = 1, h = 0], p_{11} = P[Z = 1, h = 1]$, where h is a statistical copy of h_l.

Random vector (Z, h) follows two-dimensional Bernulli distribution $Ber(p_{00}, p_{01}, p_{10})$.

Then

$$q_0 = \frac{p_{00}}{p_{00} + p_{01}}, P[h = 0] = p_{00} + p_{10}, P[Z = 0] = p_{00} + p_{01}. \tag{7}$$

One may suppose that $0 < p_{00}, p_{01}, p_{10}, p_{11} < 1$.

Thus

$$P_{er}(x) = \frac{p_{00}^{L_0} p_{00}^{L_1}(p_{00} + p_{01})}{[(p_{00} + p_{01})(p_{00} + p_{10})]^{L_0}[(p_{00} + p_{01})(1 - p_{00} + p_{10})]L_1} = \tag{8}$$
$$= \frac{(p_{00} + p_{01})^{1-L_0} p_{00}^{L_0}}{(p_{00} + p_{10})^{L_0}} \frac{p_{01}^{L_1}}{(p_{00} + p_{01})^{L_1}(1 - p_{00} - p_{10})^{L_1}}.$$

Denote $A(L_0) = \frac{(p_{00}+p_{01})^{1-L_0} p_{00}^{L_0}}{(p_{00}+p_{10})^{L_0}} = const$ under fixed L_0. Because $1 - p_{00} - p_{10} = p_{01} + p_{11}$, we have

$$P_{er}(x) = A(L_0) = \left[\frac{p_{01}}{(p_{00}+p_{01})(p_{01}+p_{11})}\right]^{L_1} = A(L_0)\left[\frac{P[Z=0,h=1]}{P[Z=0]P[h=1]}\right]^{L_1}. \tag{9}$$

From the condition of positiveness of the covariance between Z and h, one may obtain: $cov[1-Z,h] = -\ cov[Z,h] < 0$. On the other hand,

$$cov[1-Z,h] = E[(1-Zh)] - E[1-Z]E[h] = P[Z=0,h=1] - P[Z=0]P[h=1]$$

.

Henceforth $\frac{P[Z=0,h=1]}{P[Z=0]P[h=1]} < 1$ and $P_{er}(x) \to 0$ as $L_1 \infty$

The Theorem is proved.