МРНТИ 28.23.15
УДК 004.93'1

https://doi.org/10.51889/2021-4.1728-7901.09

CONSTRUCTION OF AN OPTIMAL COLLECTIVE DECISION ON THE BASIS OF A CLUSTER ENSEMBLE

Amirgaliyev Ye.[1,4], Berikov V. [2,3], Cherikbayeva L.[1,4], Tulegenova B.[1,5], Daiyrbayeva E. [1,6]*

[1]Institute of Information and Computational Technologies CS MES RK, Almaty, Kazakhstan
[2] Sobolev Institute of Mathematics, Novosibirsk, Russia
[3]Novosibirsk State University, Novosibirsk, Russia
[4] Al-Farabi Kazakh National University, Almaty, Kazakhstan
[5]Kazakh National Technical University named after K.I.Satpaev, Almaty, Kazakhstan
[6]Academy of Logistics and Transport, Almaty, Kazakhstan
**e-mail: amir_ed@mail.ru*

Abstract

This article presents methods of image analysis based on supervised learning and an algorithm consisting of two stages of determining the optimal classifier using a cluster ensemble. At the first stage, the averaged co-association matrix is calculated using a cluster ensemble. In the clustering ensemble, we used a scheme of a single clustering algorithm that constructs base partitions with parameters taken at random. At the second stage, the optimal classifier is determined using the resulting kernel matrix as input data. Numerical experiments were carried out with real hyperspectral images. The experimental results showed that the proposed algorithm has classification accuracy comparable to some modern methods.

Keywords: machine learning, cluster ensemble, coassociation matrix, image, support vector machine.

Аңдатпа

Е. Амиргалиев[1,4], В. Бериков[2,3], Л. Черикбаева[1,4], Б. Тулегенова[1,5], Э. Дайырбаева[1,6]
[1]ҚР БҒМ ҒК Ақпараттық және есептеуіш технологиялар институты, Алматы қ., Қазақстан
[2] С.Л. Соболев атындағы математика институты РҒА СБ, Новосибирск қ., Ресей
[3] Новосибирск мемлекеттік университеті, Новосибирск қ., Ресей
[4] әл-Фараби атындағы Қазақ Ұлттық Университеті, Қазақстан
[5]К.И. Сатпаев атындағы Қазақ Ұлттық Техникалық Университеті, Алматы қ., Қазақстан
[6]Логистика және көлік академиясы, Алматы қ., Қазақстан

КЛАСТЕРЛІК АНСАМБЛЬ НЕГІЗІНДЕ ОҢТАЙЛЫ ТОПТЫҚ ШЕШІМ ҚҰРУ

Бұл мақалада бақыланатын оқытуға негізделген кескінді талдау әдістері және кластерлік ансамбльді қолдана отырып, тиімді классификаторды анықтаудың екі кезеңінен тұратын алгоритмі ұсынылған. Бірінші кезеңде кластерлік ансамбльді қолдана отырып орташа коассоциациялық матрица есептелді. Кластерлік ансамбльде біз кездейсоқ таңдалған параметрлермен негізгі бөлімдерді құратын бірыңғай кластерлік алгоритм схемасын қолдандық. Екінші кезеңдетабылған ядро матрицасын кіріс деректері ретінде қолдана отырып, оңтайлы классификатор анықталады. Сандық эксперименттер нақты гиперспектральды кескіндермен жүргізілді. Эксперимент нәтижелері ұсынылған алгоритмнің кейбір заманауи әдістермен салыстырғанда классификациялау дәлдігі жоғары екенін көрсетті.

Түйін сөздер: машиналық оқыту, кластерлік ансамбль, коассоциациялық матрица, кескін, тірек векторының алгоритмі.

Аннотация

Е. Амиргалиев[1,4], В. Бериков [2,3], Л. Черикбаева [1,4], Б. Тулегенова [1,5], Э. Дайырбаева [1,6].
[1]Институт информационных и вычислительных технологий КН МОН РК, Казахстан
[2] Институт математики им. С.Л. Соболева СО РАН, г. Новосибирск, Россия
[3] Новосибирский государственный университет, г. Новосибирск, Россия
[4] Казахский Национальный Университет имени аль-Фараби, Казахстан
[5]Казахский национальный исследовательский технический университет им. К.И. Сатпаева, г. Алматы, Казахстан
[6]Академия логистики и транспорта, г. Алматы, Казахстан

ПОСТРОЕНИЕ ОПТИМАЛЬНОГО КОЛЛЕКТИВНОГО РЕШЕНИЯ НА ОСНОВЕ КЛАСТЕРНОГО АНСАМБЛЯ

В данной статье представлены методы анализа изображений, основанные на контролируемом обучении, и алгоритм, состоящий из двух этапов определения оптимального классификатора с использованием кластерного ансамбля. В этой работе представлен алгоритм состоящий из двух этапов и метод анализа изображений, основанный на полу-контролируемом обучении. На первом этапе вычисляется усредненная коассоциационная матрица с использованием кластерного ансамбля. На втором этапе определяется оптимальный классификатор с использованием полученной матрицы ядра в качестве входных данных. Численные эксперименты проводились с реальным гиперспектральным изображением. Результаты экспериментов показали, что предлагаемый алгоритм обладает высокой точностью классификации, сопоставимой с некоторыми современными методами, и во многих случаях превосходит их, особенно в условиях шума.

Ключевые слова: машинное обучение, кластерный ансамбль, коассоциационная матрица, изображение, машина опорных векторов.

Introduction

Currently, a huge number of classification algorithms and their modifications have been developed for various applied data analysis tasks of various nature and volumes: logistic regression, Bayesian classifier, decision trees, rest rules, neural networks, k nearest neighbors algorithm. There is no best criterion for the quality of clustering and universal algorithms for cluster analysis. Ensemble approach exploits the idea of collective decision making by usage of algorithms working on different settings such as subsets of parameters, subsamples of data, combinations of features, etc. Ensemble based systems usually yield robust and effective solution, especially in case of uncertainty in data model or when it is not clear which of algorithm's parameters are most appropriate for a particular problem. As a rule, properly organized ensemble (even composed from "weak" learners) significantly improves the overall quality of predictions [1,2].

Ensemble clustering is one of the successful implementations of the collective methodology. There are a number of major techniques for constructing the ensemble decision [3]. Following *evidence accumulation* approach [4], the decision is found in two steps. On the first step, a number of clustering results are obtained (for example, by usage of K-means for different number of clusters or with random initializations of centroids). For each partition variant, the co-association boolean matrix is calculated. The matrix elements correspond to the pairs of data objects and indicate if the pair belong to the same cluster or not. On the second step, the averaged co-association matrix is calculated over all variants; it is used for constructing the resultant partition: the matrix elements are considered as distances or similarity measures between data points and any clustering algorithm designed for such type of input information is applied to get a final clustering partition.

This paper introduces an algorithm of classifier construction using a combination of ensemble clustering and kernel based learning. The proposed methodic is based on the hypothesis that the preliminary ensemble clustering allows one to restore more accurately metric relations between objects under noise distortions and existence of complex data structures. The obtained kernel matrix depends on the outputs of clustering algorithms and is less noise-addicted than conventional similarity matrix. Clustering with sufficiently large number of clusters can be viewed as Learning Vector Quantization methodic [5] known for lowering the average distortion in data. These reasons, as supposed, eventually result in an increase of recognition accuracy of the combined method. The outline of the method is as follows. First of all, a number of variants of a dataset partitioning are obtained with base clustering algorithm. Then the averaged co-association matrix is calculated, where the averaging is performed with weights dependent on the obtained ensemble's characteristics. The matrix elements play the role of similarity measures between objects in the new feature space induced by implicit non-linear transformation of input features. On the second stage, a kernel classifier

is constructed by usage of the obtained co-association matrix as input kernel matrix (we used SVM in numeric experiments).

The aim of this paper is to substantiate the usefulness of the suggested combination with theoretical analysis and experimental evaluation.

There are two main types of cluster ensembles: homogeneous (when a single algorithm partitions data by varying its working settings) and heterogeneous ones (which includes a number of different algorithms). Heterogeneous cluster ensemble was considered in [6], where methods for its weights optimization were suggested. Homogeneous cluster ensemble was investigated in [7] with use of the probabilistic model assuming the validity of some key assumptions. In the current work, we follow a scheme of homogeneous ensemble and perform theoretical investigation of some of its properties using less restrictive assumptions.

The rest of the paper is organized as follows. Section 2 briefly overviews related works. Section 3 introduces necessary notions in the field of kernel based classifiers and ensemble clustering. In the next section we prove that the weighted co-association matrix obtained with clustering ensemble is a valid kernel matrix. The proposed algorithm of classifier design is also presented and some details of the optimization procedure are given. Section 5 provides a probabilistic analysis of the ensemble clustering stage. The final section describes the results of numerical experiments with the algorithm. The conclusion summarizes the work and describes some of the future plans.

Research methodology

The idea of combining cluster analysis and pattern recognition methods is rather well-known in machine learning. There are several natural reasons for the combination:

– Cluster analysis can be viewed as a tool for data cleaning to eliminate outliers or noisy items from learning sample.

– Joint learning and control sample provides additional information on data distribution that can be utilized to improve the classifier performance (this way of reasoning is sometimes called the trunsductive learning). For example, the authors of [8] make a partition of the united sample into clusters which are used to design more accurate decision rule.

– In semi-supervised learning context [9], usage of small amount of labeled data in combination with a large volume of unlabeled examples is useful for constructing more efficient classifier.

A connection between cluster analysis and kernel based classifiers was established in [10], where *cluster kernels* were proposed implementing the *cluster assumption* in the form: “two points are likely to have the same class label if there is a path connecting them passing through regions of high density only”. Three types of kernels were presented: kernels from mixture models, random walk kernels and kernels induced by a cluster representation with spectral clustering algorithm [11].

The usage of a certain similarity function (which not necessarily possesses positive semi-definiteness property) instead of kernel function was proposed in. A classifier is finding in two stages. On the first stage, the choice of some ”supporting” points is performed. With regard to these points, according to the defined similarity function, initial observations are mapped into metric space of small dimensionality. On the second stage, a linear classification rule is constructed in the new space implementing SVM-type algorithm to find the classification margin of maximum width.

Following the idea of combining cluster ensembles and supervised classification, the authors of [12] construct new feature space by usage of the degree of belonging of objects to clusters in the obtained variants of data partitioning with cluster ensemble. The transomed data table is utilized as input training set for classification using conventional techniques such as Decision Tree, Naive Bayes, K-nearest neighbors, Neural Network. The method showed its effectiveness in comparison with a number of state-of-the-art procedures.

Unlike the above mentioned works, we apply completely different combination scheme based on the notion of kernel function.

Basic preliminaries: Suppose we are given a data set $A = \{a_1, \dots, a_N\}$ consisting of N objects (examples), $A \subset \Gamma$, where Γ is a statistical population. Information about the objects is presented in the form of a feature matrix $Z = (X, Y) = (x_i, y_i)_{i=1}^{N}$, where $x_i = \left(x_{i,1}, \dots, x_{i,d}\right) \in R^d$ is input feature vector (d is feature space dimensionality), $x_{i,m} = X_m(a_i)$ is a value of feature X_m for object a_i; y_i is a class label attributed to ith object, $i = 1, \dots, N$. For binary classification task we assume $y_i \in \{-1, 1\}$. In multi-class classification problem, the set of class labels $\{\omega_1, \dots, \omega_T\}$ $(T > 2)$ is defined.

On the basis of the information about A (training sample), it is required to find a classifier (predictor, decision function) $y = f(x)$, optimal in some sense, e.g. having minimal expected losses for unseen examples. To examine the performance of the classifier, it is possible to use test sample $B = b_1, \dots, b_{N_t}, B \subset \Gamma$ described with feature matrix $\mathrm{X_{Test}}$. We shall presume that the objects in A and B are independent and identically distributed (iid), that is, the sets are collected on the basis of independent random choice of objects from Γ without replacement following a fixed distribution.

Kernel classifiers make use of the notion of kernel function $K(x_i, x_j) \geq 0$, where K is a kind of similarity measure between two data points. Linear kernel classifier exemplifies a decision function introduced within this approach:

$$f(x) = sign\left(\sum_{x_i \in X} \alpha_i y_i K(x, x_i)\right)$$

where sign is the sign function, $\alpha_1, \dots, \alpha_N$ are weights. A number of methods for determining weights (Support Vector Machine, Kernel Fisher Discriminate, etc.) exist.

For the SVM classifier, the weights are found as a solution to the constrained quadratic optimization problem of maximizing the width of a margin (separation region) between two classes in Hilbert's space induced by kernel mapping.

KFD is a kernelized version of Fisher's linear discriminant analysis (LDA) which aims at finding such a position of a straight line in feature space, for which the object's projections are separated as better as possible in the sense of a functional minimizing within-class scatter of projections and maximizing between-class distance.

The general multi-class classification problem can be solved by the application of a series of binary classification tasks for SVM or KFD, e.g., one-againstall, one-against-one or Error Correcting Output Codes (ECOC) methods [13].

Kernel k-NN classifier assigns data points according to k Nearest Neighbor rule, where neighboring points are determined with respect to similarity measure defined by kernel function.

Cluster analysis aims at determining a partition of a dataset on natural clusters using objects descriptions and a certain criterion of compactness-remoteness of groups. There exist a large number of clustering methods (see, e.g., [3]). A number of methods for obtaining the ensemble solution can be found in the literature.

Let a clustering algorithm μ be running a number of times under different conditions such as initial cluster centroids coordinates, subsets of features, number of clusters or other parameters. The joined data set $A \cup B$ is the input for the algorithm. In each lth trial, it creates a partition of the given dataset composed of K_l clusters, where $l = 1, \dots, L$, and L is the given number of runs. For each variant of clustering, we define some evaluation function γ_l (cluster validity index or diversity measure). We suppose that the values are standardized so that $0 \leq \gamma_l \leq 1$; and the better are the found variants according to some criterion, the larger are the function values.

For a pair of different data objects $a_i, a_j \in A \cup B$, we define the value $h_l(i,j) = \mathbf{I}[\mu_l(a_i) = \mu_l(a_j)]$, where $\mathbf{I}[\cdot]$ is the indicator function: $\mathbf{I}[true] = 1$; $\mathbf{I}[false] = 0$; $\mu_l(a)$ is the cluster label assigned by algorithm μ to object a in lth run. Ensemble matrix M stores the results of clusterings: $M = \left(\mu_l(a_i)\right)_{i=1,\dots,N+N_t}^{l=1,\dots,L}$.

The averaged co-association matrix $\mathbf{H} = \left(\bar{h}(i,j)\right)$ is defined over all generated variants:

$$\bar{h}(i,j) = \sum_{l=1}^{L} u_l h_l(i,j) \quad (1)$$

where the standardized weights $u_l, \dots, u_L$ indicate the quality of clustering for the given variants,

$$u_l = \frac{\gamma_l}{\sum \gamma_l}, l = 1, \dots L \quad (2)$$

Kernel classification with averaged co-association matrix

Let $K(x,\acute{x})$: $D \times D \to \mathbf{R}$ be a symmetric function, either continuous or having a finite domain, D be a closed subset in $\mathbf{R}^d$. According to Mercer's theorem, $K(x,\acute{x})$ is kernel function (i.e., it defines inner product in some metric space), if and only if for any finite set of m points $\{x_i\}_{i=1}^m$ in D and real numbers $\{c_i\}_{i=1}^m$, matrix $\mathbf{K} = \big(K(i,j)\big) = \big(K(x_i,x_j)\big)_{i,j=1}^m$ is nonnegativity definite: $\sum_{i,j=1}^m c_i c_j K(i,j) \geq 0$. Let us prove the following

Proposition. ***The averaged co-association matrix satisfies Mercer's condition.***

The symmetric property of **H** is obvious. The domain of **H** is a finite set $A \cup B$. Let $I_r^{(l)}$ be the set of indices for data points belonging to rth cluster in lth variant of partitioning. Then for any $\{c_i\}_{i=1}^m$ it holds true:

$$\sum_{i,j=1}^{m} c_i c_j \bar{h}(i,j) = \sum_{i,j=1}^{m} c_i, c_j \sum_{l=1}^{L} u_l h_l(i,j) = \sum_{l=1}^{L} u_l \sum\nolimits_{k=1}^{K_l} \sum_{i,j \in I_k^{(l)}} c_i c_j = \sum_{l=1}^{L} u_l \sum_{k=1}^{K_l} \left(\sum_{i \in I_k^{(l)}} c_i \right)^2 \geq 0$$

From this property, it follows that the averaged co-association matrix is a valid kernel matrix and can be used in kernel based classification methods.

Let us describe the main steps of the proposed algorithm KCCE (Kernel Classification with Cluster Ensemble).

Algorithm KCCE.

Input:

training data set $\mathbf{Z} = (\mathbf{X},\mathbf{Y}) = (x_i, y_i),\ i = 1,\dots,N$;

test data set $\mathbf{X_{test}}$;

L: number of runs for base clustering algorithm μ;

Ω: set of allowable parameters (working conditions) of μ.

Output:

decision function $y = f(x)$; class labels attributed to $\mathbf{X_{test}}$.

Steps:

1. Generate L variants of clustering partition of $\mathbf{X} \cup \mathbf{X_{test}}$ X using algorithm μ with randomly chosen working parameters; calculate evaluation functions and weights by formula (2);

2. For each pair $(x_i, x_j) \in \mathbf{X} \cup \mathbf{X_{test}}$ $(i \neq j)$

do

3. If the pair are assigned to the same group in lth variant, then

$$h_l(i,j) := 1, \text{ otherwise } h_l(i,j) := 0;$$

4. Using formula (1), calculate element $\bar{h}(i,j)$ of averaged co-association matrix **H**;

end;

5. Find decision function with the preset type of kernel classifier and matrix **H**;

6. Classify test sample $\mathbf{X_{test}}$ using the found decision function and matrix **H**;

end.

In this paper, we use K-means as base clustering algorithm, however it is possible to apply any other clustering technique. As the kernel classifier, we utilize soft margin version of SVM which aims at optimizing the following objective function:

$$\tfrac{1}{2}\|\omega\|^2 + C\sum_i \xi_i \to \min_{\omega,b,\xi} \quad \text{subject to: } y_i(\langle \omega, x_i \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1,\dots,N\ ,$$

where ω is normal vector to the separating hyperplane in the space induced by kernel, b is hyperplane's bias, ξ_i is a penalty imposed on ith example violating the separation margin, $C \geq 0$ is soft margin parameter. By solving for the Lagrangian dual, one obtains the quadratic optimization problem:

$$W(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \, K(x_i, x_j) \rightarrow max$$

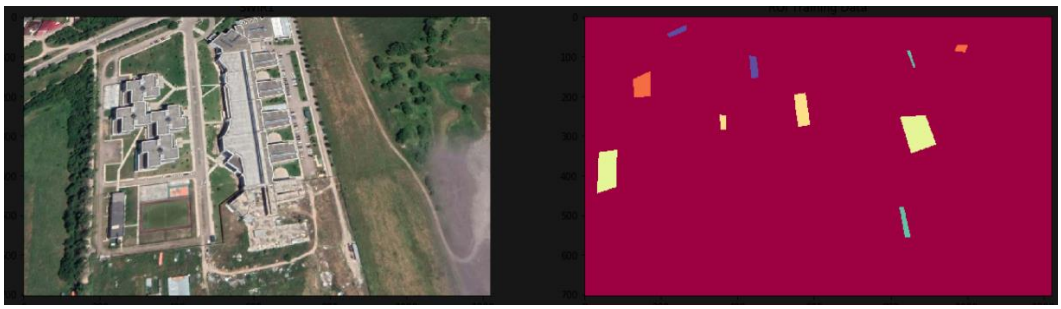
subject to: $\sum_i \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C,\ i = 1, \ldots, N$ where $K(\cdot,\cdot)$ is kernel function.

Research result

In an experiment we consider a real hyperspectral satellite image ”SDU”. The image size is 145 × 145 pixels; each pixel is characterized by the vector of 224 spectral intensities in 400- 2500 nm range. The image includes 16 classes describing different vegetation types, as one can see in Figure 1. There are unlabeled pixels not assigned to any of the classes.

These pixels are excluded from the analysis. To study the effect of noise on the performance of the algorithms, randomly selected $100r\%$ of the spectral intensity values have experienced a distorting effect: the corresponding value x is replaced by the quantity generated from the interval $[x(1-p), x(1+p)]$, where r, p are preset parameters. The dataset has been randomly divided on training and test sample in proportion 1:3.

We use multiclass SVM following ”one-against-one” strategy. Cluster ensemble size is $L = 200$. For the construction of each variant, three hyperspectral channels are randomly chosen. To obtain more diverse results of K-means, the number of its iterations is limited to 1, and the initial centroids are randomly sampled from data. In the ensemble generation, data matrix $\mathbf{X_{test}}$ is not used. The number of clusters in each variant equals $\lceil\sqrt{N}\rceil$. The weights of clusterings are constant values.



a) *b)*

Figure 1. SDU image: (a) Composite image of hyperspectral data; (b) labeled data

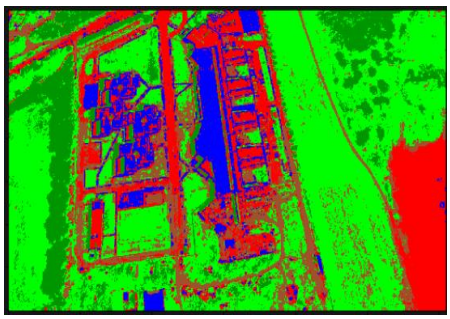


Figure 2. SDU image: classification result: Ground-truth map

We compare the proposed algorithm with SVM, Random Forest, XGBoost under similar conditions (the parameters are chosen as recommended default values in Matlab environment; RBF kernel with $\sigma = 10$ gives the best results). Table 1 shows the accuracy of classification (rate of correctly predicted class labels) on test sample for some of the noise parameters. The running time on a dual-core Intel Core i5 processor with a clock frequency of 2.8 GHz and 4 GB RAM is about 50 sec in average for KCCE and 14 sec for SVM (note that an unoptimized code is used in KCCE implementation, in contrast with efficient implementation of SVM). One can see that KCCE has revealed itself as more noise resistant than SVM, especially for large distortion rates.

Conclusion

In this work, we have introduced a supervised classification algorithm using a combination of ensemble clustering and kernel based classification. In the clustering ensemble, we used a scheme of a single clustering algorithm that constructs base partitions with parameters taken at random. It was verified that the weighted co-association matrix obtained with a clustering ensemble is a valid kernel matrix. Noise parameters r= 0.05, KCCE accuracy 0.8, Random Forest accuracy 0,659, XGBoost accuracy 0,792, SVM accuracy 0.767. The proposed combined approach experimentally has been proven to be successful when comparing with Support Vector Machine and Kernel Fisher Discriminant, Random Forest, XGBoost. The experiment with a real hyperspectral satellite image has shown that the suggested algorithm is more accurate than SVM, Random Forest, XGBoost under noise distortion.

Acknowledgment

This research has is funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (BR10965172).

References:

1. Zhuravlev, Y.I.: Principles of construction of justification of algorithms for the solution of badly formalized problems // Mathematical Notes of the Academy of Sciences of the USSR. – 1978. – Vol. 6. – P. 493–501.

2. Ajdarkhanov, M.B., Amirgaliev, E.N., La, L.L. Correctness of algebraic extensions of models of classification algorithms // Kibernetika i Sistemnyj Analiz. -2001. – Vol. 5. – P. 180–186.

3. Jain, A.K. Data clustering: 50 years beyond k-means // Pattern Recognition Letters. –Vol. 8. – P. 651–666.

4. Fred, A. and Jain, A. Combining multiple clusterings using evidence accumulation // IEEE Transaction on Pattern Analysis and Machine Intelligence. – 2005. –Vol. 27. – P. 835–850.

5. Gray, R.M. Vector Quantization // IEEE ASSP Magazine. – 1984. – Vol. 2. – P. 4–29.

6. Berikov, V. Cluster Ensemble with Averaged Co-Association Matrix Maximizing the Expected Margin // 9th International Conference on Discrete Optimization and Operations Research and Scientific School (DOOR 2016). – 2016. – P. 489–500.

7. Berikov, V., Pestunov, I. Ensemble clustering based on weighted co-association matrices. Error bound and convergence properties // Pattern Recognition. – 2017. – Vol. 63. – P. 427– 436.

8. Rahman, A., Verma, B. Cluster-based ensemble of classifiers // Expert Systems. – 2013. – Vol. 30. – P. 270–282.

9. Chapelle, O., Zien, A., Scholkopf, B. Semi-supervised learning // MIT Press. – 2006. –Vol. 27. – P. 235–250.

10. Chapelle, O., Weston, J., Scholkopf, B.: Cluster kernels for semi-supervised learning. Adv. Neural Inf. Process. Syst. 15, 601–608 (2002)

11. Ng, A. Y., Jordan, M. I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In Advances in Neural Information Processing Systems 14 (2001)

12. Iam-On, N., Boongoen, T.: Diversity-driven generation of link-based cluster ensemble and application to data classification. Expert Systems with Applications 42(21), 8259–8273 (2015)

13. Dietterich, T., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2, 263–282 (1995).