The problem of stripe classification of numbers and letters of the Ukrainian alphabet

I. Solomianiuk

The problem of distorted digital images recognition using a stripe classification algorithm based on pseudoinverse methods is analysed. The technique of digital image recognition and its computer implementation in the software environment engineering and mathematical package Matlab is suggested. This method is tested on numbers and letters of the Ukrainian alphabet. The comparative characteristic of stripe method of classification, method of neural networks, and support vector method was made.

Introduction

Theory and practice of classification systems have significant been developed in numerous works, see [1] - [6]. Nevertheless, there is a topical problem to improve the quality of recognition in the study of face images, speech signals, textual information with some formulae, and other various data processes, images and events that occur in our environment.

We consider the problem similar to the recognition using the group accounting of arguments method and support vector machines method (SVM). But in comparison with them, we use theory of perturbation of pseudoinverse and projection operations to analyse the classification systems. We defined necessary and sufficient conditions for the existence of robastic dichotomic linear separability of sets in space of features. The synthesis of the classification systems is reduced to find the best components of the vector or optimal formed linear combinations of the components. We simplify the computation using perturbation formulae for pseudoinverse matrices by replacing rows of vector in elementary matrices new rows of vector. In this paper, we consider classification algorithm using neural networks and the synthesis of linear systems by perturbation facilities of pseudoinverse and projection operations. Also we suggest its computer implementation in the software environment engineering and mathematical package Matlab.

Methods of solving the problem of classification

There are many approaches that helps us to solve different complex problems of classification. One of such approach is SVM that was proposed by V.N. Vapnik. The classificated function $F(x)$ is written in the form:

$$F(x) = sign(\langle w, x\rangle + b),$$

where $x \in \mathbb{R}^n, \langle\cdot,\cdot\rangle$ is the dot product in $\mathbb{R}^n$, w is the normal vector to the separating hyperplane, b is a parameter.

If $F(x) = 1$ we put the objects to one class while those with $F(x) = -1$ is in another class. Then we need to choose w and b to maximize the distance to each class. In this way we formulate the problem of quadratic programming to solve it using the Lagrange multipliers.

Another approach of solving this problem is neural network (NN). We built two-layer neural network, which includes 35 inputs and 43 outputs (letters 33 +10 digits). The activation function can be log-sigmoid function, which is useful because the values of output vectors are in the range [0,1]. On the hidden level we select 10 neurons. For training we use the procedure of reverse dissemination - the dissemination of the error signals from outputs of NN to its inputs, in the direction converse to direct signal dissemination.

In [4], [5] a method of a stripe classification (SC) based on pseudoinverse operations is proposed. In terms of pseudoinverse operations are presented the necessary and sufficient conditions of the existence of robastic dichotomic linear separability of sets in space of features. We need to find the best nonlinear transformations or component of vector of features or optimal formed linear combinations of components. In turn, this search is facilitated in the computation aspect by using perturbation formulae of pseudoreverse and projection matrices by replacing rows of vector in elementary matrices new rows of vector. We use the direct and inverse Grevil formulae for it. These operations naturally allow their use in superposition and also, that is especially important, in an applied sense - in the form of cascade-robastic dichotomic classification of points in the space of features.

The results of numerical experiments. Comparative characteristic of the classification methods: SVM, NN and SC

Experiments were conducted on the numbers and letters of the Ukrainian alphabet. The input images were with varying degrees of distortions. Initially, trainings were only on ideal data. For clarity we can see the graph(fig.1) that illustrates the accordance between the input and the corect recognition in output.

There is the question: how many neurons should be in the hidden layer? To answer this question we examine the graph and see the correlation between the amount of neurons in the hidden layer and the percentage of erroneous recognition.

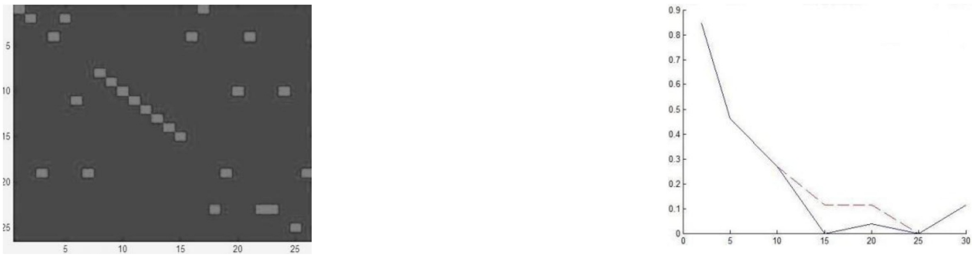

Figure 1. The accordance between the input and the corect recognition in output and the correlation between the amount of neurons in the hidden layer and the percentage of erroneous recognition (- - - error of testing, – - error of learning).

Analyzing figure 1, we can see if we increase the number of neurons, the data will be 100 percentage recognited. Similarly, re-training of the previously trained NN on large number of data will increase the probability of correct recognition. After this stage we can see the increasing percentage of correct recognited letters or digits. But it works good to a certain point. When the amount of neurons more than 25 neuronal network starts making mistakes.

Also NN was trained using distorted data. The results were compared. We noticed that NN which was traned on distorted data works much better - the percentage of erroneous recognition is lower.

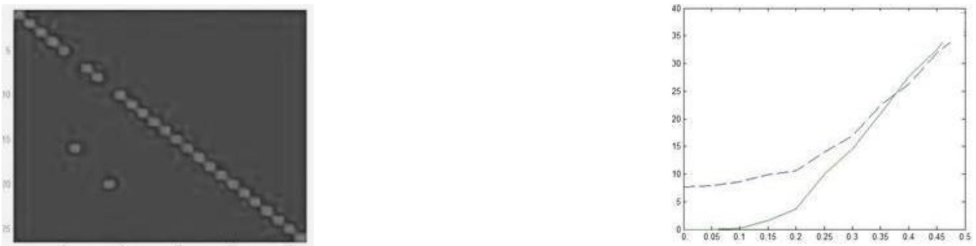

Figure 2. The accordance between the input and the corect recognition in output and the percentage of recognition errors (- - - learning without distortion, – - training with distortion).

These results are shown us to improve the correctness of recognition we need more time to study NN and increase the number of neurons in the hidden layer. We should increase the size of the input vectors, if it is possible. The most important fact is that in learning we should use more sets of input data, for example, a great deal of distorted of important information.

Vector-features of the numbers or letters of the Ukrainian alphabet were also classified using support vector machines method. It allows to solve instead of the multiextremal problems the problem of quadratic programming, which has a unique solution. Automatically this method determined the number of neurons in the hidden layer, which equals the number of support vectors. The principle of optimal separating hyperplane leads to maximize the width of separating stripes between classes, namely is more efficient classification. But in gradient neural network methods the position of separating hyperplanes is arbitrarily chosen. But if the training set contains a large amount of distorted information, it significantly affect and is considered in constructing the separating hyperplanes. There are rather small number of parameters to adjust in this method and therefore trainings are very slowly on distorted data, even which can lead to significant errors of method. To solve this problem we need to do some complex and sometimes unknown nonlinear transformation that greatly complicate the classification of data.

Taking into acount the disadvantage of support vector machines method and neural network we propose a classifier, which uses the theory of pseudoinverse operations. He was tested on distorted and ideal data.

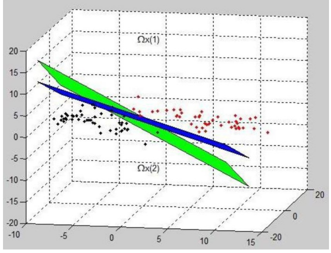

Figure 3. Example of classification of the data

Classifier optimally divided the numbers and letters of the Ukrainian alphabet by hyperplanes. We defined the width of the stripes which is equal to 0.026 units of distance. This method also allowed to expand the width between hyperplanes by using some transformations on the data (such as changing the impact of the most distorted components of the vector-features of letters (numbers) by certain transformations).

Conclusion

The theory of perturbation of pseudoinverse and projection matrices can be used to build constructive and explicit scheme to allocate from the finite set of discrete points separable subsets. It also allows to optimize the quality of such process by brute-force search elements from the subset.

The optimal synthesis of classification linear systems algorithms allow to solve the problem of classification and to stay in the class of the linear models.

In this work we present the neural network algorithm, the support vector machines method and a stripe classification method, which is in a software environment engineering mathematical package Matlab suggested. Many experiments on destorted data were done. They show us the importance of using the idea of synthesis of neurofunctional transformations (a stripe classification) that help us solve more complex problems.

References

[1] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks, 13(2): 415-425, March 2002.

[2] Kohonen T. Self-Organizing Maps, -3-d ed. - Tokyo: Springer, 2001.-501 p.

[3] Vapnik, V.N. Statistical Learning Theory. New York: Wiley. 1998

[4] Kirichenko N.F., Krivonos Y.G., Lepekha N.P. Optimization of the systems by hyperplanious clasters and neurofunctional transformations in systems of classification of the signals: Cybernetics and Systems Analysis. - 2008. - 6th ed. P. 107-124 (in russian).

[5] Kirichenko N.F., Krivonos Y.G. Lepekha N.P. Synthesis of the systems of neurofunctional transformations in the problems of classification. / / Cybernetics and Systems Analysis. - 2007. - 3-d ed. P. 47-57(in russian)

[6] Kussul N.M., Chelestov A.U., Lavrynyuk A.M. Intelligent computing : Textbook. - Kyiv: Naukova Dumka, 2006.-186 p.(in ukrainian)

Authors

Inga Hryhorivna Solomianiuk — the 4th year student, faculty of cybernetics, Taras Shevchenko national university of Kyiv, Kyiv, Ukraine; E-mail: *Solomeniukinga@gmail.com*