Development of Classifiers which Based on Tuned Model of the Identification

M.Tatur, D.Adzinets, V.Ostrouski, D.Lavnikevich

Byelorussian State University of Informatics and Radioelectronics

Introduction

Effectiveness of recognition systems and decision making directly depends on classification technique and learning algorithm which are intelligence core of whole system. During classifier development for any applied problem researcher must carry out the following actions:

1) To choose the classification technique (classifier structure)
2) To create the representative learning data sets
3) To choose and apply the effective method of learning
4) To estimate the effectiveness of the technical solution
5) To implement the classifier as a soft and hardware

In modern theory of pattern recognition the lot of classification methods with appropriate learning algorithms and also multiple examples of their applications is known [1-5]. They are classification on minimal distance, artificial neural networks, support vector machine and etc. Usually classification problem is described as dividing of patterns in hyperspace of informative features and so learning algorithms as dividing hypersurface synthesis. The most "simple" problems are linearly dividable, so surface described by linear polynomial is sufficient for their solving, but "difficult" problems need polynomial of higher ranks. And here the multilayer neural networks with good math apparatus are once approach which wide known and really used. However with the known popularity the neural network have a number of restrictions and weak spots. Already they are marked as convergence learning problem and interpretable of classification and learning processes. With the attempts of hardware realizations of neural networks the researcher meet the problem when each technical solution has unique topology of neural structure. As a result the architectures of modern neural computers are not effectiveness on productivity/cost criterion.

The creating of representative data sets and estimation of learning effectiveness also are the part of classifiers development problem. To provide the correct comparative estimations the general (standard) data sets from application fields are used, for example: data sets of face images, of medical diagnoses, texture images and etc. [6-9]Often to provide the testing and estimation of formal classifiers the abstract numerical data sets are needed. Currently the conventional data sets with universally recognized methodology for experimental estimation of classification models don't exist.

1. General approach

To explain for presented approach let introduce the definition on *ideal* classifier. The function for decision making Z will be ideal classifier,

$$Z = F(X, Y)$$

when functional F and tuning parameters Y are known a priori and where X - input vector of informative features.

In the classical formulation of the task on classifier development it is necessary to choose the functional Ψ with so tuning parameters U that to provide the minimal error (mean square error d) on learning or testing data sets.

$$Z^* = \Psi (X, U),$$

$$d = \sqrt{\frac{1}{(m-1)}\left(\frac{\sum (z_i - z^*_i)^2}{\sum z_i^2}\right)}$$, where m – number of test vectors

In significant number of applied problem the classifier function can be written with known functional F and approximately given tuning parameters Y `.

$$Z` = F (X, Y `),$$

Then classifier development will not include the choose of math method or structure design (for example, structure of multilayer neural network). Learning of classifier will be present as tuning of parameters Y `.

In the given work the original model of identification is proposed. It is similar to model of formal neuron and combines from general methodical positions the basic properties of classical and fuzzy prototypes. By using of the present model the synthesis of classifiers according to applied problems and generation of representative learning data sets and corrective comparative estimation for complete technical solution will be provided.

2. Mathematical model of identification

Given model of identification contains as logical L and arithmetical Z components. In simplest variant the logical part is presented as arithmetical minimum of informative features on determined set N'. Following development of model can execute in direction of fuzzy conclusion implementation. (In present version it doesn't show).

$$Z^L = L\ Z$$

$$L = \min_{i=1}^{n} \varphi' (x_i,a_i,b_i,c_i,d_i) \quad ,$$

$$\varphi' (x_i,a_i,b_i,c_i,d_i) = \begin{cases} \varphi (x_i,a_i,b_i,c_i,d_i), & i \in N' \\ 1, & i \notin N' \end{cases}$$

Functions of informative features parameterization $\varphi\ (x_i,a_i,b_i,c_i,d_i\ )$, and also activation function Z are linearly approximated with aim for simple hardware realization.

$$\varphi\ (x_i,a_i,b_i,c_i,d_i\ ) = \begin{cases} 0, & a_i < x_i\ ,\ x_i > d_i \\ \dfrac{d_i - x_i}{d_i - c_i}, & c_i \le x_i \le d_i \\ 1, & b_i < x_i < c_i \\ \dfrac{x_i - a_i}{b_i - a_i}, & a_i \le x_i \le b_i \end{cases}$$

$$Z = \begin{cases} 1, & S(X) > p_2 \\ \dfrac{S(X) - p_1}{p_2 - p_1}, & p_1 \le S(X) \le p_2 \\ 0, & S(X) < p_1 \end{cases}$$

$$S(X) = \sum_{i=1}^{n} w_i\ \varphi\ (x_i,a_i,b_i,c_i,d_i\ )\ ,$$

Fundamental significance of proposed model consists of following. It combines as discrete and fuzzy, as arithmetical and logical components, which can use as separately and jointly. Whole model connects two directions known as data processing and knowledge processing. By means of composition of tuning parameters $W, A, B, C, D\ , N', p_1\ , p_2$ model demonstrates a smooth transformation of qualitative difference and allows to produce the typical functions, which adequately corresponds to applied recognition problems.

For example:

with $p_1 = p_2$ - discrete decision making is realized;

with $p_1 \neq p_2$ -fuzzy decision making is realized;

with $w_i \neq 1$ - weighted adding of features is realized;

with $a_i = 0,\ b_i = c_i = d_i = 1$ –absence of parameterization of informative features is realized;

with $a_i = b_i,\ c_i = d_i$ -discrete borders of parameters for each informative feature are realized;

with $a_i \neq b_i,\ c_i \neq d_i$ – fuzzy borders of parameters for each informative feature are realized;

with $a_i = b_i = c_i = d_i$ – arithmetical or logical comparison of informative feature with standard parameter is realized;

if $N' \in \varnothing$ –logical conditions are not used;

if $N' \notin \varnothing$ - logical conditions on key features depending on parameterization with each informative feature are;

with $p_1 = p_2 = 0,\ N' \notin \varnothing$ - logical conditions are realized only.

3. Ideal test data sets generation

Proposed approach allows to generate the numerical data sets $X \rightarrow Z$ for typical classification problem. These data sets can be used for testing and verifying of classifier effectiveness and learning algorithms, un depended on mathematical method of realization. Term as "Ideal Data Set" is used because accurate values of function $Z=F(X,Y)$ are calculated with given tuning parameters of model – Y and generated sequence of input data – X. But usually data sets are produced be means of expert estimations from concrete applied fields (medicine, image processing and etc.).

Due to data sets creation the possibility to put the following main parameters is present:

- the number of informative features -n;
- the number of test vectors-m;
- format and accuracy of representation of input and output data;
- kind and level of noises.

Created data sets can be used in two different applications

In first application

-for verifying and estimation of effectiveness of concrete classifier and learning algorithm, developed for concrete recognition system;

- for testing of stability of concrete classifier and learning algorithm to "expert" noise.

In this application it is necessary to create nest data set for the concrete typical recognition problem. Due to experiments providing the detail techniques of comparative analysis can be used.

In second application:

-for creating common objective estimation and providing competitive selection of classifiers and learning algorithms;

-for creating classifier library and learning algorithm which reflecting of modern level of knowledge in given area.

According to this application the prototypes of test data sets for three typical tasks has been generated [10]. Each of them includes as discrete and fuzzy modes of decision making. So six variants of tasks ranked on level if complexity (intelligence) increasing. Test data set is the table, which contains input vectors of informative features and values of classification functions $X \rightarrow Z$ on six task. In given variant the data sets don't contain noise and aren't intended to process of experiments for stability estimation to noise.

Other characteristics of data sets are:

-number of informative features $n = 5$;

-number of test vectors $m = 100$;

-accuracy of input and output data representation $0,01$.

To carry out testing the researcher must adapt, turn own classifier to quantitative parameters of *common* data sets and execute testing own programs according to *common* technique [11]. Obviously to providing the global research measurement the data set must be verified on representation and as organizer the known scientific organization can issue. So given data sets will be proposed for UCI repository [6] and "Polygon of algorithms" [11].

4. Soft and Hard implementation of classifiers

The proposed identification model combines the model of formal neuron and one rule of fuzzy conclusion as composed components. To develop the classifiers for multi class problems the trivial typing of identification model is sufficiently. Functional complexity (ability to nonlinear differentiation of classes) is provided by each module of identification.

In difference from neural networks which have unique (multilevel as normal) topology of neurons interconnections the classifiers based on proposed model will have standard topology as full connected two level graph with number of input equal number of informative features and number of output equal number of classes. This fact provides to get the significant advantages on hardware implementation of classifier core as co-processor.

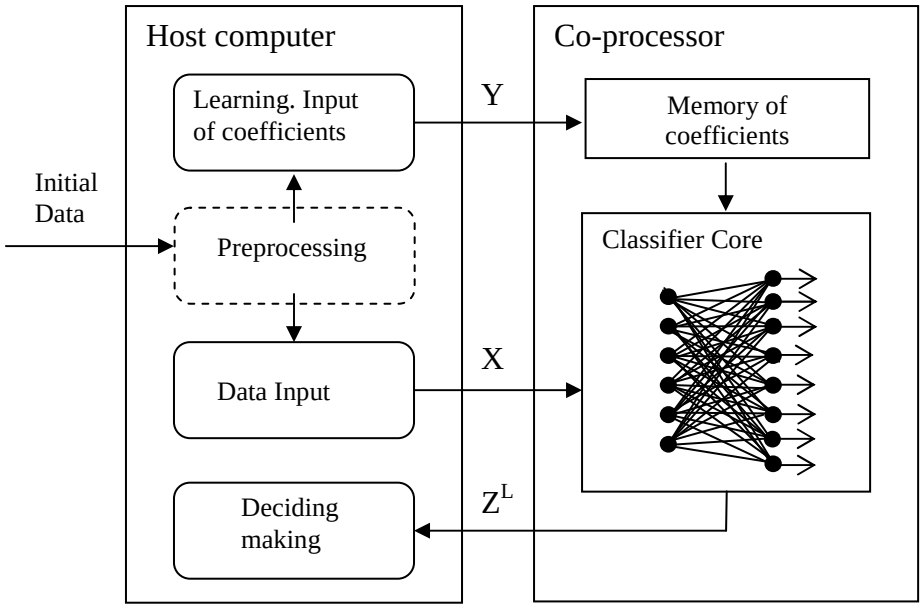

Fig.1. Soft and Hard implementation of classifiers

In fig.1 the general scheme of Soft and Hard implementation of classifiers based on proposed model is shown. The functions of preprocessing, data input/output, learning, decision making and global control are put on Host computer, which can be usual PC. Co-processor is aimed fast to process the K task of identification. Increasing of productivity will be rich by means of hardware paralleling and will be higher when more informative features and more classes are present in applied classification problem. Co-processor is realized with FPGA and connected to Host computer by means of standard interface.

Architecture of such computer complex can increase the calculating productivity of classifier core in depend on solving task and currently stayed simple for application by end user. Cost of complex will be compared with cost of serial produced neural computers with DSP.

Conclusion

The general tuned model of identification is presented. Function of model includes as operation of weighted adding similar formal neuron as and elements of fuzzy conclusion similar fuzzy neural networks. It is proposed to consider this model as basic component with development cascade classifiers for various number of classes. Then structure and topology of processor network will be ones level and standard with un depend on implemented function.

In the model the possibility flexible to choose the function according to applied problem is archived. If to set arguments X , W, A, B, C, D , N' ,p_1 , p_2, then it is possible calculate ideal volumes of function $F(X,Y)$. So, model provides the generation of test numerical data sets for typical classification problem. With using such "ideal" data sets the corrective comparing analyzes of developed classifiers is possible.

The property of flexible tuning of model and standard architecture are important aspects to choose the model as mathematical basis for implementation of multi repeated operation due to development of neural liked computers.

References

1. J.Yu, Y.Ou, Ch.Zhang, Sh. Zhang "Identifying Interesting Visitors through Web Log Classification" // Intelligent Systems , 2005, Vol.20, No3, p.55-59.
2. B.Krisnapuram, A.J.Hartemink, L.Carin, M.A.Figneiredo "A Bayesian Approach to Joint Feature Selection and Classifier Design" // Pattern Analysis and Machine Intelligence, 2004, Vol. 26, No9, p.1105-1111.
3. O.Melnik, Y.Vardi, C.-H.Zhang "Mixed Group Ranks: Preference and Confidence in Classifier Combination" // Pattern Analysis and Machine Intelligence, 2004, Vol. 26, No8, p.973-981.
4. L.Chen, Y.Pan, X.Xu "Scalable and Efficient Parallel Algorithms for Euclidean Distance Transform on the LARPBS Model" // Parallel and Distributed Systems, 2004, Vol.25, No11, p.975-983.
5. S.Theodoridis, K.Koutroumbas Pattern Recognition. Second Edition. Academic Press an imprint of Elsevier(USA), 2003.
6. UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

7. http://www.grappa.univ-lille3.fr/~torre/Recherche/Experiments/Datasets/

8. http://face.nist.gov/frvt/feret/feretdocuments.htm

9. http://www.cbsr.ia.ac.cn/IrisDatabase.htm
The Center for Biometrics and Security Research (CBSR) is founded by the Institute of Automation, Chinese Academy of Sciences (CASIA).
10. www.i-proc.com
11. www.machinelearning.ru