Pattern Recognition and Forecasting

"AVO-POLYNOM" RECOGNITION ALGORITHM

Alexander Dokukin

Abstract: Estimates Calculating Algorithms have a long story of application to recognition problems. Furthermore they have formed a basis for algebraic recognition theory. Yet use of ECA polynomials was limited to theoretical reasoning because of complexity of their construction and optimization. The new recognition method "AVO-polynom" based upon ECA polynomial of simple structure is described.

Keywords: pattern recognition, estimates calculating algorithms, algebraic approach, recognition polynomials.

ACM Classification Keywords: I.5.2 [Pattern Recognition]: Design Methodology – Classifier design and evaluation

Conference: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009

Introduction

ECA or Estimates Calculating Algorithms [1] are a parametrical family of methods for pattern recognition developed in Computing Centre about thirty years ago. The idea of method is simple. Training sample is divided into two parts: actual training and check ones. Closeness to each object of training sample as well as remoteness from it is stimulated, i.e. the estimation of object S belonging to class K is increased if S is close to some representative of K or is far from a representative of K's addition. The value of increasing is determined by the representative's weight.

ECA was widely used for solving applied tasks. In addition, a number of theoretical results have been achieved for its algebraic closure. The most important of them proved existence of correct polynomial over ECA [2]. Yet there was a huge distance between theoretical reasoning and application, since former was based on polynomial constructions over ECA family, while latter on optimization of single ECA by its weights [7].

The major step in applying polynomials to the real world problems was made by reducing correct polynomial's complexity both in number of items and power. The approach was based on maximizing ECA's height, i.e. difference between minimal estimation of regular pair (object, class) and maximal estimation of irregular one [4]. A number of algorithms for minimization of ECA height have been suggested and tested, both precise [5] and approximate [8]. Either of them had a major drawbacks: precise ones being too slow for polynomial construction [6] while approximate ones not precise enough.

Nevertheless during the analysis of different combinations of methods a regularity has been noticed. ECA's of maximal height tend to have good recognition quality in some areas close to their so called center. This fact has been assumed as a basis for a novel recognition method named "AVO-polynom" that is Russian for ECA-polynomial.

Definitions

The following recognition problem is referred to as a standard problem. We consider two samples of vectors from the n-dimensional feature space: a learning sample and a check one. For definiteness, we assume that the former sample contains m objects: $S_1, \ldots, S_m$, while the latter one contains q objects: $S^1, \ldots, S^q$. We also assume that the set of admissible objects is divided into l classes, which may intersect in general case. The classification of each object in the learning sample is known; it is necessary to reconstruct classification of the check sample.

The family of ECAs is defined as follows.

1. Each feature is ascribed a certain weight p_i, $i = \overline{1,n}$.

2. Certain subsets of the set of features, which are referred to as supporting subsets, are singled out. The aggregate of these subsets is denoted by Ω_A. Each supporting set $\omega \in \Omega_A$ has a weight.

3. The proximity function $B_\omega(S, S')$ for two objects in the supporting set is introduced. We will use the threshold proximity function unless specially announced; i.e., two objects $S = (a_1,...,a_n)$ and $S' = (b_1,...,b_n)$ will be regarded close if the following inequalities hold for all supporting features:

$$\rho_i(a_i, b_i) < \varepsilon_i, \quad \forall i \in \omega.$$

Here ε_i, $i = \overline{1,n}$ are called the proximity function thresholds.

4. Each S_j of the learning sample is ascribed its own weight $\gamma(S_j)$, $j = \overline{1,m}$.

5. The estimate of an object class is calculated by the formula

$$\Gamma_j(S^t) = x_0 \cdot \Gamma_0^j(S^t) + x_1 \cdot \Gamma_1^j(S^t),$$
$$\Gamma_0^j(S^t) = \sum_{S_i \in C\tilde{K}_j} \gamma(S_i) \sum_{\omega \in \Omega_A} p(\tilde{\omega}) \cdot \overline{B_\omega}(S_i, S^t),$$
$$\Gamma_1^j(S^t) = \sum_{S_i \in \tilde{K}_j} \gamma(S_i) \sum_{\omega \in \Omega_A} p(\tilde{\omega}) \cdot B_\omega(S_i, S^t).$$

Here, the following variables and notation are used: $x_1, x_0 \in \{0,1\}$, $\tilde{K}_j = K_j \bigcap \{S_1,...,S_m\}$, $C\tilde{K}_j = \{S_1,...,S_m\} \setminus \tilde{K}_j$, $\overline{B_\omega}(S_i, S^t) = 1 - B_\omega(S_i, S^t)$.

The height of the ECA is defined as the difference between the minimal estimate of a regular pair (object, class) (i.e., the pair whose object belongs to the corresponding class) and the maximal estimate of an irregular pair [4].

Some changes have been made to a classical ECA optimization. First of all, optimization by objects' weights was replaced with optimization by similarity functions thresholds for better flexibility. Secondly, the optimization criterion has been changed too. Instead of recognition quality over whole check sample the height on its subset is considered. The optimization problem is reduced to the search for the values ε^* of the ε-thresholds of the proximity function, which maximize the functional:

$$\varepsilon^* = \arg\max_{\varepsilon \in (0,\infty)^n} \left(\min_{(i,j) \in M_1} \Gamma_j(S^i) - \min_{(u,v) \in M_0} \Gamma_v(S^u) \right).$$

Here M_1 denotes set of regular pares and M_0 of irregular ones.

"AVO-polynom"

The method has been designed to be a part of software system RECOGNITION [3] that applies some restrictions on training sequence. First of all, the input sample has to be divided into training and checking parts. By default the division is made randomly in proportion 2 to 1. This parameter is a single one which can be adjusted by user, and its default value covers most part of tested cases.

Second and the most time consuming part is devoted to finding a set of simple ECAs with better recognition quality. The input sample divided into two parts is further divided to q smaller overlapping ones. Each checking object in combination with all training ones forms a set for training simple ECA. The checking object used is referred to as central object of the ECA. The method of fastest descent [8] is then used to find ECA of maximal height. If positive height can't be achieved the central object is considered as outlier and corresponding ECA is dropped out.

The local nature of each recognition operator achieved is taken in account by dividing its contribution by distance to the central object. I.e. final estimations are calculated by formula

$$\Gamma_j(S) = \sum_{i=1,n} \frac{\Gamma_j^i(S)}{d(S,S^i)}.$$

The second multiplier can be expressed in terms of ECA with use of specific distance functional. Thus, the whole construction represents second degree polynomial over ECA.

In the next section "AVO-polynom" will be compared to some over recognition methods. They are simple ECA [7], logical regularities and linear machine [3].This choice is not accidental. Simple ECA shows advantages of using polynomial instead of single item. Logical regularities have similar nature since it founds some typical hyper parallelepipeds in feature space. Linear machine shows results of completely different approach.

Testing results

The testing was performed with the set of seven real world tasks from the UCI Repository of Machine Learning Databases. All samples have been pre-divided into training and testing ones. The latter was used only for quality estimation. Here is the list of used samples: Abalone, Breast-canser, Ionosphere, Echocardiogram, Hepatitis,

Image, Credit. Testing results are described in following table:

Task	Simple ECA	Logical regularities	Linear Machine	AVO-polynom
Abalone	57.3	-	**65.5**	62.3
Breast canser	**96.3**	94.1	95.5	96.1
Ionosphere	81.9	89.6	85.2	**98.7**
Echocardiogram	76.1	59.2	70.4	**77.4**
Hepatitis	79.5	83.1	78.3	**88.0**
Image	89.0	**93.2**	93.7	89.4
Credit	**86.2**	77.9	85.9	**86.2**

In general "AVO-polynom" performed on the same level with best methods, but some results deserve to be mentioned specially. For example in Abalone task the best result has been achieved with Linear Machine, but AVO-polynome has far surpassed Simple ECA and Logical regularities. In some other tasks AVO-polynom have shown simply the best results.

Acknowledgements

The work is presented with financial support of RFBR (Projects 08-01-00636-a, 08-07-00437-a) and grant of the President of Russian Federation "Scientific School – 5294.2008.1".

Bibliography

[1] Yu.I. Zhuravlev, Well-Posed Algebras over a Set of Ill-Posed (Heuristic) Algorithms I, Kibernetica, No. 4, 14–21 (1977).

[2] Yu.I. Zhuravlev, Well-Posed Algebras over a Set of Ill-Posed (Heuristic) Algorithms II, Kibernetika No. 6, 21–27 (1977).

[3] Yu.I. Zhuravlev, V.V. Ryazanov, O.V. Senko, RECOGNITION. Mathematical methods. Software System. Practical Solutions. (in Russian), Moscow, Phasis, 2006, ISBN 5-7036-0106-8.

[4] Yu.I. Zhuravlev, I.V. Isaev, Construction of Recognition Algorithms Correct for a Given Control Sample, Zh. Vych. Mat. Mat. Fiz. 19 (3), 726–738 (1979).

[5] A.A. Dokukin, Generalization of the Method for Constructing Maximum-Height Estimate-Calculating Algorithms to Recognition Problems, Pattern Recognition and Image Analysis, 2006, Vol. 16, No. 4, pp. 689–694.

[6] A.A. Dokukin, On complexity of searching the optimal ECA (in Russian), Reports to All-Russia Conference MMPO-12, 2006.

[7] V.V. Ryazanov, Optimization of estimates calculating algorithms by representativeness parameters of precedents (in Russian), Zh. Vych. Mat. Mat. Fiz. 16 (6), 1559--1570 (1976).

[8] A.A. Dokukin, On construction of samples for testing approximate methods for optimization of estimates calculating algorithms (in Russian), Zh. Vych. Mat. Mat. Fiz. 46 (5), 978--983 (2006).

Authors' Information

Alexander Dokukin – Researcher; Dorodnicyn Computing Centre of Russian Academy of Sciences, 40, Vavilova St., Moscow, Russian Federation; e-mail: dalex@ccas.ru