ABOUT NEW PATTERN RECOGNITION METHOD FOR THE UNIVERSAL PROGRAM SYSTEM “RECOGNITION”

Alexander Dokukin and Oleg Senko

Abstract: In this work the new pattern recognition method based on the unification of algebraic and statistical approaches is described. The main point of the method is the voting procedure upon the statistically weighted regularities, which are linear separators in two-dimensional projections of feature space. The report contains brief description of the theoretical foundations of the method, description of its software realization and the results of series of experiments proving its usefulness in practical tasks.

Keywords: pattern recognition, statistically weighted regularities, voting procedure.

Introduction

Nowadays there are a great number of effective pattern recognition methods based on voting procedure upon some kind of regularities in the data, as well as different approaches for searching these regularities. The term “regularity” is interpreted as some sub-region in space of prognostic variables where fraction of at least one of the classes differs significantly from its fraction in neighbor regions. For example, there are the method of voting upon the sets of irreducible tests [1] or representative tests [13], the method of voting upon statistically weighted syndromes (further in the text it is referred as SWS) [2], method of voting upon sets of logical regularities [3] and etc. Results of hands-on testing show the higher steadiness of voting procedure to the minor changes in training and testing samples, which leads to the significant increase of quality of voting-based methods. This advantage is especially important in relatively high-dimensional tasks with limited number of cases in data sets. The theoretical substantiation of this fact [4] exceeds the bounds of this report, but the detailed proof by the means of mathematical statistics is now being prepared for publication.

All these methods have one strong restriction and maybe even disadvantage. It is the fact that all regularities are some kind of hyper parallelepipeds in feature space with planes orthogonal to the datum lines. However in many tasks the essentially multidimensional regularities may arise which are separated from neighborhood by multivariate linear boundaries. So the method of two-dimensional linear separators (further referred as TLS) presents an attempt to complicate shape of elementary regularities preserving all advantages of voting procedure.

The Method of Two-Dimensional Linear Separators (TLS)

Further following notation will be used. Let’s consider the set of permissible objects M , let’s also consider that it presents Cartesian product of n sets of permissible values of features $M = M_1 \times ... \times M_n$. It is presumed that there is the unknown subdivision of the set M into l classes $K_1,..., K_l$. This subdivision is described by means of training sample $S_1,..., S_m$ of objects $S_i = a_{i1},..., a_{in}, i = 1,..., m$, for which the classification is known: $\alpha(S) = \alpha_1,...,\alpha_l$, where $\alpha_j =< S \in K_j >, j = \overline{1,l}$. It is necessary to restore the unknown classification of the testing sample $S^1,..., S^q$.

The main point of the method is the successive examination of different pairs of features and construction of the linear separator for every one of them and for every class. These separators must divide two-dimensional projections of objects of selected class and its additive inversion. For every class K_i and for every pair of features (u,v) the found line $L^i_{(u,v)}$ is called elementary regularity. Moreover the weight of the regularity $w^i_{(u,v)}$ is calculated due to the separating ability of the line.

The recognition is based on the weighted voting procedure by the set of elementary regularities. Let’s consider the estimation for k-th class, the rest ones are calculated in much the same way. Each regularity $L^k_{(u,v)}$ refers the new object S to the k-th class or to its additive inversion, so the part of training objects of k-th class in the half plain there the object was referred to can be calculated for each pair of features $\upsilon^{k}_{(u,v)}(S)$. The final estimation is calculated according to the following formula:

$$\Gamma(S, K_1) = \frac{\sum\limits_{(u,v)} \upsilon^{1}_{(u,v)}(S) w^{1}_{(u,v)}}{\sum\limits_{(u,v)} w^{1}_{(u,v)}}.$$

The weights of regularities are calculated in much the same manner as in the method of statistically weighted syndromes (SWS) and depend on the quality of separating of training sample. If there are any errors in separation, i.e. some objects from the class are referred to its additive inversion or objects from inversion are referred to the class, the weight is set to be inversely proportional to the variance of the error

$$\frac{1}{p(1-p)},$$

there p is the part of errors. If this is not the case and all training objects are separated correctly than the variance of error is replaced with its Bayesian estimation

$$\frac{\int\limits_0^1 (1-p)^n \, dp}{\int\limits_0^1 (1-p)^n \, p dp}$$

In the TLS method the linear separators are sought by means of pattern recognition method called Linear Machine [5]. Its main point is that the task of finding separating line is replaced with the task of finding the maximal simultaneous subsystem of the system of linear inequalities and its subsequent solving by means of relaxation algorithm.

In conclusion of this paragraph let's consider the results of some hand-on testing. In the table 1 there are some tasks that clearly demonstrate the advantages of unification of voting procedure and complex elementary regularities. It contains the results of comparison of Linear Machine, SWS and TLS methods. The method with best performance is marked with gray color.

Task	LM	SWS	TLS
Breast	94.9	94.1	95.2
Ionosphere	85.2	90.1	90.1
Iris	97.5	95	97.5
Mel	50	65.6	68.8
Patomorphosis	76.5	85.3	91.2

Table 1. Comparison of LM, SWS ans TLS methods

The following tasks were considered during the test series:

- Breast – the breast cancer recognition, 9 features, 2 classes, 344 training examples, 355 testing ones (Breast cancer databases was obtained from Dr. William H. Wolberg from the University of Wisconsin Hospitals, Madison [6]);
- Ionosphere – the recognition of structural peculiarities in ionosphere, 34 features, 2 classes, 170 training examples, 181 testing ones (data from Johns Hopkins University Ionosphere database);
- Iris – Iris recognition, 4 features, 3 classes, 71 training examples, 81 testing ones (data from Michal Marshall's Iris Plants Database);

- Mel – Recognition of melanoma by the set of geometrical and radiological features, 33 features, 3 classes, 48 training examples, 32 testing ones [12];
- Patomorphosis – forecast of destruction level of malignant growth after chemotherapy by the set of parameters characterizing optical behavior of its cell nucleus, 7 features, 2 classes, 43 training examples and 31 testing ones (the data has been received from Dr. Matchak from Cancer Research Center of the Russian Academy of Medical Sciences).

In the table 2 the results of comparison of TLS with some other methods build-in to the Recognition software system are shown. The methods ate tested with two tasks which features are small number of objects in comparison with dimension of task. It is important that the suggested method has shown the significant increase of quality in this class of tasks.

Method	Mel	Patomorphosis
TLS	68.8	91.2
LM	50	76.5
SWS	65.6	85.3
LDF	59.4	76.5
AVO	62.5	76.5
IT	62.5	85.3
QNN	62.5	70.6
Perceptron	65.6	79.4
SVM	56.3	76.5

Table 2. Comparison with other methods on the tasks with short samples.

Following methods were used: TLS – Two-dimensional Linear Separators, LM – the mentioned above Linear Machine[5], SWS [2], LDF - Fisher's Linear Discriminant [7], AVO or ECA – Estimates Calculating Algorithm [8], IT – voting upon Irreducible tests [1], QNN – q Nearest Neighbors [7], Perceptron – Multilayer Perceptron [7,9], SVM – Support Vector Machine [10].

Software Realization

Software system “Recognition” has been developed in Dorodnicyn Computing Centre of Russian Academy of Sciences in cooperation with Solutions Ltd. The system's detailed description can be found, for example, in the proceedings of the Open German-Russian Workshop [11] or at the developer's Internet sight http://www.solutions-center.ru. In this article only the brief description of its basic principles will be given, because these principles have been being considered throughout the whole TLS' development process. They are universality, uniformity, modularity and intellectuality.

The universality of the system is understood as a wide coverage of different approaches to pattern recognition and classification including so-called classifier fusion, which are realized in the system's library of methods. The methods have been developed as separate interchangeable modules of uniform structure. On the software level each module is a single dynamic-link library with standardized interface.

While solving a wide variety of different practical task the initial assumption that each recognition method has its advantages and there is no single best one for every kind of tasks has been proven. So the main accent was made on using of different kinds of classifier fusions. And the uniformity of the methods allows combining the results of every subset of developed methods into one classifier, providing results more accurate than average and even close to maximal ones in automatic training mode. This fact allows claiming some kind of intellectuality of the developed system.

Passing on to software realization of Two-dimensional Linear Separators method itself, it is important to take note of two facts:

First of all, developing of TLS method’s software realization was significantly simplified due to availability of software system “Recognition”, since it was taking care of all the chores including preparation of methods environment, quality control and etc. So the developers in the person of the authors of this paper were able to get concentrated on the method itself.

Secondly, the fact that TLS method has significantly increased the quality of recognition for some kind of tasks has been already mentioned in section 2. Thus the addition of this method to classifier fusions allows increasing their quality greatly for these tasks. The experimental proof of this fact is shown in table 3. One of the simplest ways of constructing classifier fusion has been considered. The simple majority voting procedure has been applied firstly to the set of LM, SWS, LDF, AVO, IT, QNN, Perceptron and SVM, and secondly to the same set of algorithms in addition with TLS.

Task	Without TLS	With TLS
Mel	59.4	65.6
Patomorphosis	76.5	82.4

Table 3. Quality of majority voting

Conclusion

In conclusion we can claim that the developed method has justified our hopes. The combination of voting procedure and linear separators has increased recognition quality in some class of practically important tasks. Thus the developed software realization of TLS method can serve as a great support for researchers.

Acknowledgment

The research has been supported by RFBR grants NN 02-07-90137, 02-07-90134, 02-01-00558, 03-01-00580, 02-01-08007 and state program НШ-1721.2003.1 (scientific schools).

Also thanks goes to Dr. William H. Wolberg from the University of Wisconsin Hospitals, Madison; Vince Sigillito from Johns Hopkins University (vgs@aplcen.apl.jhu.edu); Michal Marshall (MARSHALL%PLU@io.arc.nasa.gov); Dr. Ganster from Linz University and Dr. Matchak from Cancer Research Center of Russian Academy of Medical Sciences for the granted data.

Bibliography

[1] А.Н.Дмитриев, Ю.И.Журавлев, Ф.П.Кренделев, О математических принципах классификации предметов и явлений. Сб. "Дискретный анализ". Вып. 7. Новосибирск, ИМ СО АН СССР. 1966. С. 3-11.

[2] Сенько О.В. Использование процедуры взвешенного голосования по системе базовых множеств в задачах прогнозирования// М. Наука, Ж. вычисл. матем. и матем. физ. 1995, т. 35, № 10, С. 1552-1563.

[3] V.V.Ryazanov, Recognition Algorithms Based on Local Optimality Criteria, Pattern Recognition and Image Analysis. 1994. Vol.4. no.2. pp. 98-109.

[4] Сенько О.В., Кузнецова А.В. Метод предварительной селекции признаков. // Докл. Всеросс. Конф. «Матем. методы распознавания образов - 11». М.: Регион-Холдинг, 2003, С. 171-173.

[5] Обухов А.С., Рязанов В.В. Применение релаксационных алгоритмов при оптимизации линейных решающих правил. Доклады 10-й Всероссийской конференции "Математические методы распознавания образов (ММРО-10)", Москва, 2001, 102-104.

[6] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp. 9193-9196.

[7] R.O. Duda and P.E. Hart. Pattern Recognition and Scene Analysis. A Wiley-Interscience Publication. John Wiley and Sons. New York, 1973.

[8] Ю.И.Журавлев, Об алгебраическом подходе к решению задач распознавания или классификации, Проблемы кибернетики. М.: Наука, 1978. Вып.33. С.5-68.

[9] Рязанов В.В., Челноков Ф.Б., О склеивании нейросетевых и комбинаторно-логических подходов в задачах распознавания, Доклады 10-й Всероссийской конференции "Математические методы распознавания образов (ММРО-10)", Москва, 2001,115-118.

[10] Christopher J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition, Appeared in: Data Mining and Knowledge Discovery 2, 1998, 121-167.

[11] Yu.I. Zhuravlev, V.V. Ryazanov, O.V. Sen'ko, A.S. Biryukov, D.P. Vetrov, A.A. Dokukin, N.N. Katerinochkina, A.S. Obukhov, M.Yu. Romanov, I.V. Ryazanov, F.B. Chelnokov. The Program System for Data Analysis "Recognition" (LOREG). The 6th German-Russian Workshop "Pattern Recognition and Image Understanding". Proceedings. Novosibirsk, Russia, 2003, pp. 255-258.

[12] Ganster H., Gelautz M., Pinz A., Binder M., Pehamberger H., Bammer M., Krocza J. Initial Results of Automated Melanoma Recognition //Proceedings of the 9th Scandinavian Conference on Image Analysis, Uppsala, Sweden, June 1995, Vol.1, pp. 209-218.

[13] Dyukova Ye.V., Ryazanov V.V., The Solution of Applied Problems by Recognition Algorithms Based on the Voting Principle. VTs Akad. Nauk S.S.S.R., Moscow, 1986.

Authors' Information

Alexander A. Dokukin – Dorodnicyn Computing Centre of the Russian Academy of Sciences, Vavilov st., 40, Moscow GSP-1, 119991, Russia; e-mail: dalex@ccas.ru

Oleg V. Senko – Dorodnicyn Computing Centre of the Russian Academy of Sciences, Vavilov st., 40, Moscow GSP-1, 119991, Russia; e-mail: senkoov@mail.ru

ANALYSIS OF SECURITY IN ARCHIVING

Dimitrina Polimirova–Nickolova

Abstract: Some basic types of archiving programs are described in the paper in addition to their advantages and disadvantages with respect to the analysis of security in archiving. Analysis and appraisal are performed on the results obtained during the described experiments.

Keywords: Web Security, Mail Security, Information Security, Archive Programs, Compressed Objects, Methods Of Encryption.

The Present Situation

In the development of the computer science the creation and the use of archived objects is a classical research problem, which has found different resolutions for decades past. Nowadays the availability of several dozens of methods and their varieties represent an excellent demonstration of the ambitions of the information systems' programmers and designers for a real high-speed and high-effective compression of information flows.

The following basic types of archiving programs could be defined with respect to the information security of compressed objects, obtained after examination of more than 320 archiving programs, known by now:

1) E-mail archiving programs – in this kind of archiving programs the relative homogeneity of the information flow (e-mail traffic) is used and the most suitable methods of compression are selected. There are some differences among the basic existing e-mail clients (MS Outlook, MS Outlook Express, Netscape Mail, Opera Mail, Eudora Mail, Pegasus Mail etc.), which make possible the applying of different realizations of the compressing process. The advantages consist in the multiple reduction of the saved e-mail folders' volume and in the high degree of security against unauthorized access (viruses, worms, spyware, malware etc.). The disadvantages above all are related to the consumption of computing resources to realize the right and the reverse transformation.