THE ALGORITHM BASED ON METRIC REGULARITIES

Maria Dedovets, Oleg Senko

Abstract: *The new pattern recognition method is represented that is based on collective solutions by systems of metric regularities. Unlike previous methods based on voting by regularities discussed technique does not include any constraints on geometric shape of regularities. Metric regularities searching is reduced to connected components calculating in special contiguity graph. Methods incorporate statistical validation of found metric regularities with the help of permutation test.*

.***Keywords:*** *pattern recognition, metric regularities, permutation test*

ACM Classification Keywords*: I.5 Pattern Recognition; I.5.2 Design Methodology – Classifier design and evaluation*

Conference topic*: Pattern recognition and Forecasting, Machine Learning*

Introduction

Family of pattern recognition tools exist that are based on searching of regularities that are subregions of features space containing objects of one class only (complete regularity) or predominantly of one class (partial regularity). Usually in these methods subregions geometric shapes correspond to some sub predetermined model. The most frequently used model is the model of regularities, where the desired subregions of features space have the form of hyperparallelepiped. Logical regularities ([Yu. I. Zhuravlev et al.,2006], [V.V. Ryazanov,2007] , [A. A. Ivachnenko and K. V. Vorontsov]) and «syndromes» concept ([Yu. I. Zhuravlev et al.,2006], [V.A. Kuznetsov et al., 1996], [O.V. Senko and A. V. Kuznetsova]) may be mentioned. Besides approaches exists where regularities are formed with the help of hyperplanes in features space ([A.A. Dokukin and O.V.Senko], [O.V. Senko and A. V. Kuznetsova]). However a priori specification of the geometric shape is a significant limitation hampering detection of some regularities actually existing in data, but not satisfying constraints that are put on by used model. In this work a new type of regularities (metric regularities) is discussed. Metric-type regularities allow to find the feature space subregions containing representative groups of connected objects of some class with minimal inclusion of other classes. We use a concept of a generalized connectivity between the same class objects (contiguity relationship $\Diamond_k$) to define metric-type regularities. It must be noted that no suppositions about regularity shape are made.

Metric (generalized) regularities.

The standard problem of pattern recognition is considered. Let Ω is general set of recognized objects. Let $\tilde{S} = \{S_1, ..., S_q\} \subset \Omega$ - a training sample, $\tilde{K} = \{K_1, ..., K_l\}$ - a set of classes , the class $K_j, j = \overline{1,l}$ for each training set object $S_i, i \in \overline{1,q}$ is a priory known. Thus description of object $S_i \in \tilde{S}$ consists class indicator and feture description that is used for recognition.Our goal is to find a correct class $K_j, j = \overline{1,l}$ for every arbitrary object $S \in \Omega$. It is supposed that metric $\rho(s', s'')$ are defined at Cartesian product $\Omega \times \Omega$. Correct metric choice is a significant issue in problems of classification, clustering, and nonparametric regression. Metric

is a mathematical model of the similarity of objects, and its choice in many cases is not unique. In recent years increasingly used methods where the metric is adjusted to the training set. We use a weighted Minkowski metric for each fixed set of features: $\rho(S',S'') = \left(\sum_{j=1}^{n} w_j \left| f_j(S') - f_j(S'') \right|^p \right)^{\frac{1}{p}}$, where $f_j(S)$ is object *S* feature *j* value, w_j - feature *j* weight. The weights of features are specified by reasons of normalization: $M_j = \max_{i=1,..,q} | f(S_i) |, w_j = M_j^{-p}$

Definition 1. Two objects S_1 and S_2 are connected by class K or are in contiguity relation ship ($S_1 \Diamond_K S_2$), if there are not in the training set $\widetilde{S}$ such object $S \in CK$ (CK is addition of class K to *full train set*), that both following inequalities will be correct:

1) $\rho(S_1, S_2) < \rho(S, S_1)$

2) $\rho(S_1, S_2) < \rho(S, S_2)$, where ρ is a metric in feature space.

The relation $\Diamond_K$ defines an indirected contiguity graph $G_{K\tilde{S}}$ such that there is one to one correspondence between vertices of $G_{K\tilde{S}}$ and objects from $\tilde{S} \cap K$. There is an edge between to vertices v' and v'' if an only if corresponding them objects S' and S''.

Definition 2. We call a set $F \subseteq K \cap \widetilde{S}$ a regularity by metric ρ (or connectivity field) of class *K*, if: $\forall S', S'' \in F \;\; \exists S_1,...,S_n \in F : S' \Diamond_K S_1 \Diamond_K ... \Diamond_K S_n \Diamond_K S''$.

In other words any regularity by metric ρ correspond connected component of graph $G_{K\tilde{S}}$. Intuitively regularity by metric regularity by metric ρ corresponds compact in terms of metric regularity by metric ρ region of feature space with minimal inclusion of objects that do not belong to class K

Definition 3 (Regularity quality definition)

. A set *F'* is called a closure by training set $\widetilde{S}$ of a class *K* regularity *F* if: $F' = F \cup \{ S \in CK \cap \widetilde{S} \mid \exists S' \in F : S' \Diamond_K S \}$.

Thus closure *F'* is union of *F* and objects from $\widetilde{S}$ that do not belong to class *K* but are connected to class objects from *F*. It is naturally to define regularity *F* quality as fraction of objects from class *K* in closure *F'*: $val(F) = \frac{|F|}{|F'|}$.

Closure *F'* may be calculated using reference objects concept.

Reference objects

A minimal set $\tilde{S}_{ref}(F)$ of objects from F such that any object from F is connected to at least one object from $\tilde{S}_{ref}(F)$ will call reference patterns:

Let F is a regularity. Then $\tilde{S}_{ref}(F) = \arg\min_{A \subseteq F} (| A \subseteq F \; || \; \forall S' \in F, \exists S \in A : S \Diamond S')$ is a reference objects set.

Closure *F' now* is defined union of *F* and objects from $\widetilde{S}$ that do not belong to class *K* but are connected to class objects from $\tilde{S}_{ref}(F)$. At that regularity quality is define in the same way. Using reference objects concept allows to decrease significantly amount of information that is need for regularity description. Instead of storage of description of all objects from F it is sufficient to store descriptions of objects from $\tilde{S}_{ref}(F)$.

Regularities validation

Parameter $val(F)$ describe regularity quality but it does not allow to answer a question if discovered regularity is statistically valid? However statistical validity of regularity may be evaluated with the help of permutation test that is based on generating of set $\{\tilde{S}_{rand}\}$ of randomized data sets. To receive randomized data set $\tilde{S}_{rand}$ positions of class indicators in initial data set $\tilde{S}$ are randomly permuted. Then they are newly put in correspondence to fixed positions of features descriptions. The metrics regularities are found for each $\tilde{S}_{rand}$. Let $val_{\max}(\tilde{S}_{rand})$ is maximal value of quality parameter $val(F)$ for regularities revealed at $\tilde{S}_{rand}$. Measure of statistical validity of regularity F found at true initial data set (p-value) is defined as ratio $p = \frac{|\{\tilde{S}_{rand} \mid val_{\max}(\tilde{S}_{rand}) > val(F)\}|}{|\{\tilde{S}_{rand}\}|}$. It is evident that statistical validity is greater if p-value is less.

Possibility of statistical validation with the help of permutation test allows to use metric regularities searching as a tool for evaluating if there is an effect of explanatory variables $X_1,\ldots,X_n$ on outcome categorical variable. It is sufficiently to define metrics in X space, to find regularities and to evaluate their validity.

Training

At initial stage contiguity graph $G_{K\tilde{S}}$ is constructed for each class *K* and connected components of these graphs are calculated. Closures are found for corresponding regularities and quality parameters $val(F)$ are calculated. At the second stage permutation test is used to evaluate statistical validity of regularities found at the first stage. At the third stage final set $\tilde{F}_{final}$ of regularities is formed. At that only regularities with high statistical validity or with evaluated p-values less some threshold are joined $\tilde{F}_{final}$. Usually threshold equal 0.1 was used.

Often there are outlying objects in data sets that deviate significantly from main part of data set. These outliers may significantly affect relationships $\Diamond_k$ between objects of the same class and decrease method efficiency. So procedure was suggested that allow to reveal outliers. It is assumed that the fraction of outliers in a class can be no more than X% of objects. Outlying objects $\tilde{S}_{outliers}$ are stored, analyzed and removed or not removed at the end of the learning.

Recognition

Let *S* - arbitrary regognized legitimate object. Estimates for class *K* are calculated as fraction of reference sets of class *K* metric regularities, connected to *S*, amoung full set of class K metric regularities. Object *S* is put into the class with maximal estimate.

Experiments

The program version of algorithm based on metric regularities was realized. In this version regularities are searched for each pair of features and estimates are calculated by set of all found regularities. Performance of the version was evaluated in variety of tasks including tasks from UCI Machine Learning Repository . At that its recognition ability was compared with recognition abilities of standard statistical methods, neural networks, support vectors machine and algorithm based on voting by systems of regularities. Experiments demonstrated sufficiently high generalization ability of method. Inclusion of algorithm in RECOGNITION system ([Yu. I. Zhuravlev et al.,2006])and in the set of algorithms for resource Poligon.MachineLearning.ru for further study is planned.

Acknowledgment

The authors would like to thank ITHEA -XXI for support of these researches:

Bibliography

[Yu. I. Zhuravlev et al.,2006] Yu. I. Zhuravlev, V.V. Ryazanov, O.V. Senko. *RECOGNITION. Mathematical methods. Program system. Applications* (in Russian). - Fuzis, Moscow, 2006. p.176.

[V.V. Ryazanov,2007] Ryazanov V.V. Logical regularities in recognition tasks (nonparametrical approach). ЖВМ и МФ, №10, 2007, с.1793-1808.

[V.A. Kuznetsov et al., 1996] V.A. Kuznetsov, O.V. Senko, A.V. Kuznetsova et all. Recognition of fuzzy systems by method of statistically weighed syndromes and its use for immune and hematologic norm and chronical pathology. *Chemical Physics*, 15 (1) (1996), p.81-100

[Yu. I. Zhuravlev et al.,2008] Zhuravlev Yu.I., Kuznetsova A.V., Ryazanov V.V., Senko O.V., Botvin M.A. The Use of PatternRecognition Methods in Tasks of Biomedical Diagnostics and Forecasting // Pattern Recognition and Image Analysis, MAIK Nauka/Interperiodica. 2008, Vol. 18, No. 2, pp. 195-200.

[O.V. Senko and A. V. Kuznetsova] O. Senko, A. Kuznetsova A recognition method based on collective decision making using systems of regularities of various types.

Pattern Recognition and Image Analysis, Vol. 20, No. 2. (1 June 2010), pp. 152-162.

[A. A. Ivachnenko and K. V. Vorontsov] Ivachnenko A. A., Vorontsov K. V. Upper boundaries of overfitting and variety profiles for logical // Mathematical methods of pattern recognition-13. — M.: MAKS Press, 2007. — p. 33–37.

[A.A. Dokukin and O.V.Senko] Dokukin A.A. Senko O.V. About new pattern recognition method for the universal program system Recognition. Proceedings of the International Conference I.Tech-2004, , Varna (Bulgaria), 14-24 June 2004,pp. 54-58.

Authors' Information

Oleg Senko *– Leading researcher in Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119991, Moscow, Vavilova, 40, e-mail: senkoov@mail.ru*

Dedovets Maria *– postgraduate student in Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119991, Moscow, Vavilova, 40, e-mail: dedovets_m@mail.ru*