SOI: 1.1/TAS DOI: 10.15863/TAS

International Scientific Journal
Theoretical & Applied Science

p-ISSN: 2308-4944 (print) e-ISSN: 2409-0085 (online)

Year: 2018 Issue: 01 Volume: 57

Published: 30.01.2018 http://T-Science.org

Mirzayan Mirzaakhmedovich Kamilov
Akademic of Uzbek Academy of Sciences, The Scientific and Innovation Centre of Information and Communication (SIC IKT)

Mirzaakbar Khaqqulmirzaevich Hudayberdiev
Senior researcher, SIC IKT
mirzaakbarhh@gmail.com

SECTION 4. Computer science, computer engineering and automation

FORMATION OF A QUALITATIVE DESCRIPTION OF THE TRAINING SET IN SOLVING THE RECOGNITION PROBLEM

Abstract*: The paper discusses the possibilities of improving the quality of the recognition algorithm based on partial precedent, by the original pre-training procedures. The peculiarity of this algorithm is that as precedents only such "anchor points" of a pattern that ensuring the following conditions are left: the distance from any point on the training set of i-th pattern to their nearest precedent is less than the distance to the nearest precedent of another pattern. This set of precedents provides unmistakable recognition of all samples of the training set. Thus, the probability of correctly separating of classes increases significantly. The set of dedicated training samples gives a chance to improve the level of reliability of data mining. One of species of tulip has been chosen as object of research. This process is carried out via morphological features of tulip. The information about tulip is obtained from Central herbarium of institute of Botany of the Uzbek Academy of sciences.*

Key words*: data mining, pattern recognition, algorithm of partial precedents, precedent, training set, etalon objects, classification, clustering.*

Language*: English*

Citation: Kamilov MM, Hudayberdiev MK (2018) FORMATION OF A QUALITATIVE DESCRIPTION OF THE TRAINING SET IN SOLVING THE RECOGNITION PROBLEM. ISJ Theoretical & Applied Science, 01 (57): 33-37.

Soi: http://s-o-i.org/1.1/TAS-01-57-6 ***Doi***: crossref https://dx.doi.org/10.15863/TAS.2018.01.57.6

Introduction

It is known [1,10,11] that the classification and clustering are among the main tasks of data mining. They are included to the more general class of intellectual tasks in the problem of pattern recognition (PR). The difference between them lies in the problem (or rather, the problem of pattern recognition – “supervised learning”) that uses information contained in the so called "precedent or etalon table or training set" (table of reference objects whose belonging to a particular class is known). In other words, the assignment of a new (control) object to certain class (object classification) is based on identifying the extent of its "closeness" to the known precedent (pattern) of a training set, belonging to a particular class of which is known.

Since the 60’s of last century to the present various classification algorithms have been developed, studied and found its place in practice in solving a myriad of applications [1,10,11]. General fundamental step in the formation of the vast majority of these algorithms is the choice of task in one form or another of the function of the distance between objects, which is determined by the value of the degree of "similarity".

The aim of this article is not to develop a new algorithm of classification, but it considers a private matter involving the study of the possibility of applying a class of algorithms of pattern recognition – algorithms of partial precedents (APP) to the plant recognition [6-11].

Related works

Today there are a lot of developments analyzing morphology of the plants. For example, the information databases are widely used in the developing of programs. The following databases are available in the international network of the Internet [5]:

- IRIS-150 objects, 4 features;
- Mushroom-8124 objects, 22 features;
- Soybean-307 objects, 35 features;
- Plants-22632 objects, 70 features, etc.

In [4] combinations of features that can improve classification performance on a large dataset of similar classes were investigated. To this end they introduce a 103 class flower dataset. They compute four different features for the flowers, each describing different aspects, namely the local shape/texture, the shape of the boundary, the overall spatial distribution of petals, and colors. They combine the features using a multiple kernel framework with a SVM classifier. The weights for each class are learnt using the method of Varma and Ray [3,4], which has achieved state of the art performance on other large dataset, such as Caltech 101/256. Their dataset has a similar challenge in the number of classes, but with the added difficulty of large between class similarity and small within class similarity.

Classification algorithm

In order to avoid difficulties in understanding the following material and, in particular, coverage of how to use a phased implementation APP form we investigate the procedure for calculating estimates (procedure "voting"), summary of the requisite laws of the theory class APP (for details-in these above sources [5,6]) is given below.

Practical problem solving of Data Mining (DM) established [6,8] that the initial information that must be processed usually has the form of numeric tables (matrices) consisting of *m* rows and *n* columns. The rows $s_1, s_2, \dots, s_m$ represent the information about the object under study and the columns $x_1, x_2, \dots, x_n$ reflect the properties (attributes, characteristics, features, signs) of these objects or phenomena. The intersection of the j row and i column indicates the value α_{ij} of i feature in the j object. Note that, in theory, and in clustering and classification so-called valid objects, etc. the features values of which are the elements of a certain set (the set of such elements forms a so-called alphabet feature), are only considered. The set of admissible *m* objects, each of which is described by a set of values of *n* features, reduces the so-called allowable table T_{nm}.

We introduce some concepts that we need in the future. Consider the set of Boolean vectors $\widetilde{\omega}$ of length n. All the individual coordinates $\widetilde{\omega}$ are selected. Let the numbers of these coordinates are $i_1, i_2, \dots, i_k$. All columns except the columns with numbers $i_1, i_2, \dots, i_k$ are removed from the table T_{nm}. The portion of the table T_{nm} corresponding to the coordinates of a single Boolean vectors $\widetilde{\omega}$, called a $\widetilde{\omega}$ part of the table T_{nm} is obtained. The strings $\widetilde{\omega}$ part of the table T_{nm}, denoted by $\widetilde{\omega}s_1, \widetilde{\omega}s_2, \dots, \widetilde{\omega}s_m$ parts called $\widetilde{\omega}$parts matching rows (represent $\widetilde{\omega}$ part descriptions of objects).

We note in passing that if the rows of the table T_{nm} are separated into groups (classes), we get a table with a given classification, denoted by (in the case of ℓ classes) through T_{nm}. With tables $T_{nm\ell}$ that are in particular, the reference set of objects of use cases, unless the partition of precedents in the table corresponds to objectively existing distribution by classes of objects in the population studied subject area. As noted above, such a table of reference objects is playing the role of "supervise learning" in solving classification problems.

We turn now to a brief description of the class of APP.

The basic model of algorithms of partial precedents presented below is defined by specifying the six main stages [2,7,11].

1. The system of support sets. Consider all non-empty $M_{\widetilde{\omega}}$ subsets of {1,2,...,n}. We denote the set of all subsets through Ω:

$$\Omega = \{\widetilde{\omega} | \widetilde{\omega} \subseteq \{1, \dots, n\}\}.$$

The first item is to set the definition of APP family of sets $\Omega_A \subseteq \Omega$, which is called a system of support sets of *A*. As these systems can be, for example, the set of all elements of the Ω with the same power (the power of elements characterized by parameter k, integer values, which can vary in the range from one to n) or the set itself Ω. There may be other examples of the system Ω_A [5].

2. Proximity function. Let s and s_q – be valid objects. The second step is to determine the APP task functions $r(\widetilde{\omega}s, \widetilde{\omega}s_q)$, whose values reflect the degree of "similarity" $\widetilde{\omega}$ part of two objects.

3. Estimates for the lines on a fixed support set. The third step in determining the algorithm *A* is to set numerical data - estimates for the line on the function value close to the lines $\widetilde{\omega}s, \widetilde{\omega}s_q$. In the simplest case, this estimate is denoted by

$$\widetilde{\omega}\Gamma(s, s_q) = r(\widetilde{\omega}s, \widetilde{\omega}s_q) \qquad (1)$$

4. Evaluation for the class on a fixed support set. In solving many problems of DM (including classification step), it is necessary to assess the degree of proximity of the object s to the class through the establishment of the degree of its proximity to all objects of a class separately. This estimate is given as follows.

We assume that class (for example, $\mathcal{K}_1$ forms a line (objects) $s_1, s_2, \dots, s_{m1}$of the table $T_{nm\ell}$, and for each of them in accordance with (1) values $\widetilde{\omega}\Gamma(s, s_1), \widetilde{\omega}\Gamma(s, s_2), \dots, \widetilde{\omega}\Gamma(s, s_{m1})$ are calculated. Estimate for the value of the class $\mathcal{K}_1$: $\Gamma_1(\widetilde{\omega}) = G[\widetilde{\omega}\Gamma(s, s_1), \widetilde{\omega}\Gamma(s, s_2), \dots, \widetilde{\omega}\Gamma(s, s_{m1})]$.

It can be defined as:

$$\Gamma_1(\widetilde{\omega}) = \sum_{q=1}^{m1} \widetilde{\omega}\Gamma(s, s_q) \qquad (2)$$

where, $\widetilde{\omega}$ as in stage 3, corresponds to the selected training set.

5. Estimate for the class system of support sets. Let according to (2), in claim 4 for each item $M_{\widetilde{\omega}} \in \Omega_A$ is based assessment $\Gamma_u(\widetilde{\omega})$, $u = \overline{1, \ell}$ (assuming there ℓ classes). Then the estimate $\tilde{A}_u(s)$ for the system class training set can be determined, for example, as follows:

$$\Gamma_u(s) = \sum_{M_{\widetilde{\omega}} \in \Omega_A} \Gamma_u(\widetilde{\omega}) \qquad (3)$$

6. The decision rule for the algorithm ***A***. The decision rule algorithm is function of the values $\Gamma_u(s)$, $u = \overline{1, \ell}$ calculated in the previous step. The

range of values of this function F is $0,1,2,\dots,\ell$. If $\mathrm{F}[\Gamma_1(s),\Gamma_2(s),\dots,\Gamma_\ell(s)] = u, u = \overline{1,\ell}$, then object of s as the most similar to the class C_u, is considered to belong to this class. If it is $\mathrm{F}[\Gamma_1(s),\Gamma_2(s),\dots,\Gamma_\ell(s)] = 0$, considered that class for the s is not determined.

Generation Dataset

The recognition quality is defined as a functional $\varphi_A = \alpha_1\xi_1 + \alpha_2\xi_2$, here ξ_1 – number of incorrectly recognized objects, ξ_2 – number of objects that the system refused to recognize, α_1 & α_2 – respectively the coefficients that determine the quality and reliability requirements of recognition after each state. An analysis of the values obtained with the help of φ_A, must satisfy the recognition coefficients $\psi_A \geq 70\%$, proposed by experts. Removing one or more columns in a table after each iteration is done by analyzing values of the functional φ_A. This process continues for as long as inequality $\varphi_A \leq g$ is satisfied. For this purpose the following condition should be satisfied:

$$\mathbb{Z} = \begin{cases} \varphi_A \to min, \varphi_A \leq g, \\ \quad \psi_A \to max. \end{cases} \tag{4}$$

here $\mathbb{Z}$ – task of pattern recognition, ψ_A-average recognition accuracy, $g -$ threshold.

Decreasing values of ξ_1, ξ_2 or using them as constants provide high efficiency recognition procedures. It should be noted that $\xi_1 = 0$, $\xi_2 = 0$ is accepted as the accuracy of recognition $\mathbb{Z}$. In this case reducing of training set and defining of the active fragment objects carried out by the following procedure: influence of the objects on the accuracy of recognition is determined by the values satisfying$\mathbb{Z}$. During the training process, training set separates all objects into classes. Every class шы possessed them as etalon of objects:

$$S_j^E = \{s_1^E, s_2^E, \dots, s_\ell^E\}, \quad j = \overline{1,\ell},$$

$$\Gamma_\Omega\left(S_j^E, S\right)\gamma_j\left(\textstyle\sum_{i:\omega_i=1} p_i\right)B_\Omega\left(S_j^E, S\right),\ i = \overline{1,n}. \tag{5}$$

here $s_j^E = \frac{1}{\mathrm{m}}\sum_{j=1}^m s_{ij}$, j-number of etalon object, $i -$number of object, γ_j-parameter is characterized degree of the importance object and p_i-value of importance of the feature.

The recognition is carried out as follows. There is input object in the system, belonging to a particular class. Distances from the object to the etalons of all patterns are measured, s system belongs to the class, the distance to which is minimum. The distance is measured with the metric, which is introduced to solve a specific problem of recognition.

The first stage in the training set "cover" all the objects of each class hyper sphere as a smaller radius as possible. Calculate the distance from a etalon to all objects of this class, included in the training set. The hyper sphere is selected to cover the maximum area $B_\Omega(S_\tau, S_\nu), \tau \neq \nu$[10,11]. The hyper sphere is constructed with the center in the etalon with distance:

$$B_\Omega(S_\tau, S_\nu) = \begin{cases} 1, if\ d\left(S_j^E, S\right), \\ 0, \quad otherwise. \end{cases} \tag{6}$$

here $d\left(S, S_j^E\right) = |\{\nu: \left|x_\nu(S) - x_\nu\left(S_j^E\right)\right| \leq \varepsilon_\nu, \nu = \overline{1,n}\}$, set of $u = \left\{x_{i_1}(S_\nu), x_{i_2}(S_\nu), \dots, x_{i_k}(S_\nu)\right\}$ is called a representative set for the class $S_\nu \in C_j$.

It covers all objects of this class. This procedure is carried out for all classes.

Experimental evaluation

For the experiments we use of the flower dataset of Central herbarium Institute of the gene pool of flora and fauna of the Uzbek academy of sciences [6]. Specialists on the subject proposed genus of Tulip, because it is good investigated subject area by Uzbek specialists. It consists of 34 types of flower (Tulip), 780 objects in data base and 16 features [6,7,8]. We are called types of tulip with classes or precedents. We separated four classes, which have sufficient information on the dataset. The dataset is split into a training set and a test set. In this case, recognition of objects with informative features system, since the source data TulipaL tulips family consists of four classes, every class consists of 20 objects and all 80 objects over the four classes. It means that objects are given to their class more than 70% or less of their voice and the satisfy classes ordered the descending $\mathrm{K}_1 \geq \mathrm{K}_2 \geq \dots \geq \mathrm{K}_\ell, (here\ \ell = 34)$, and we used three groups of training sets and four classes, that shown in the table 1.

The content group will not change, if it satisfy task of the pattern recognition other cases will changing content of group and here E_1, E_2, E_3-groups of etalons.

Table 1

Steps of reproduce actual results are expected quantitative etalon objects

Classes	Etalon groups and ψ_A (%)		
	E_1	E_2	E_3
K1	94	94	94
K2	85	90	91
K3	71	79	80
K4	64	75	75

The analysis voice of objects given to their class into the distance $d(S, S_j^E)$, that shown the following figure 1. As it can be seen, selection of the etalon objects for the training set to determine the threshold.

$$\Gamma_u(S) = \begin{cases} 1, if \sum_{\Omega_A} \Gamma_u(\widetilde{\omega}) > g, \\ 0, if \sum_{\Omega_A} \Gamma_u(\widetilde{\omega}) < g. \end{cases} \qquad (7)$$

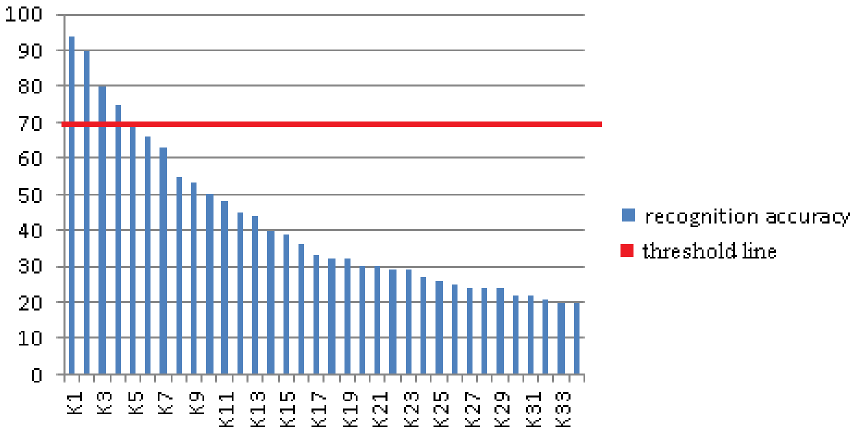

Figure 1. Voice of objects given to their class

As seen from the results in figure 2, for some of classes the number of objects is not enough to be the etalon. In this case, training set is required to supply additional objects. We defined them as minimum half parts of training set is contained etalon objects. And we obtained next results by the realization these requirements. The training set consists of 20 objects per classes and is used to learn the 16 features. Here is a list of data varieties: $\mathcal{K}_1$ – Tulipa korolkowii Regel; $\mathcal{K}_2$ – Tulipa lehmanniana Mercklin; $\mathcal{K}_3$ – Tulipa scharipovii Tojibaev; $\mathcal{K}_4$ – Tulipa sogdiana Bunge.

Conclusion

We selected etalon objects for training set. The giving of different weights for every class enables us to use an optimum features combination for each classification. This allows improving intellectual data analysis results. The investigation realized by means of program-recognition complex “PRASC-2M”, which is based on algorithms of partial precedents.

References:

1. Aleksey A., Pastukhov A., Aleksander A., Prokofiev A. (2017) Clastering algorithms application to forming a representative sample in the training of multilayer perceptron//St.Petrsburg Polytechnical University Journal: Physics and Mathematics 3(2017), p. 127-134. http://dx.doi.org/10.1016/j.spjpm.2017.05.004
2. Hudayberdiev M. Kh., Akhatov A. R., Hamroev A. Sh. (2011) "On a model of forming the optimal parameters of the recognition algorithms". International journal of KIMICS, Vol.9, #0.5, October 2011, p. 95-97
3. Nilsback M.E., Zisserman A. (2008) Automated flower classification over a large number of classes. in: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, 2008, p.722-729.
4. Varma M., Ray D. (2007) Learning the discriminative power-invariance trade-off. In Proc. ICCV, 2007
5. (2018) UCI. Centre Machine Learning and Intelligent Systems. Available: https://archive.ics.uci.edu/ml/datasets.html
6. Kamilov M.M., Hudayberdiev M.Kh., Khamroev A.Sh. (2016) About the approach to the solution of the problem of formation of active features for recognition of one family of the objects of the plant world. Scientific journal Problems of Computational and Applied Mathematics. Tashkent, 2016, #3(5). p. 44-49.
7. Khamroev Alisher (2015) The solution of problem of parameterization of the proximity function in ACE using genetic algorithm. IJRET:International Journal of Research in Engineering and Technology, India, Bangalor. Volume: 04, Issue: 12, December-2015, 100-104 p.
8. Khamroev Alisher (2017) An algorithm for constructing feature relations between the classes in the training set. Original Research Article Procedia Computer Science, Volume 103, 2017, Pages 244-247. http://dx.doi.org/10.1016/j.procs.2017.01.094.
9. Zagoruiko N.G. (2013) Cognitive Analysis of Data. Novosibirsk, Academic Publishing GEO press. 185 p.
10. Zhuravlev Yu.I., Ryazanov V.V., Senko O.V. (2005) Recognition (Mathematical methods. Software system. Practical applications). Moscow:FAZIC Press. 2005.