CYBERNETICS AND PHYSICS, VOL. 11, NO. 1, 2022, 25–29

PROPERTIES OF THE ORDERED FEATURE VALUES AS A CLASSIFIER BASIS

Vladimir Shats
Saint Petersburg, Russia
vlshats@hotmail.com

Article history:
Received 13.08.2021, Accepted 25.05.2022

Abstract

The paper proposes a new classifier based on new concept closeness for objects finite set: feature values of the same class objects are close if the difference between these values is small enough. To pass to this concept, the combined sample data for each feature k were approximated by mapping onto a set of ordered pairs (k, m), where m is the interval number of the feature ordered values. The objects of each pair have close values of the considered feature. Number lists of training sample objects of the same class, forming ordered pairs, was called an information granule. The frequency of any granule is calculated from the length relation of corresponding subsets as a complex event. These frequencies allow us to calculate the frequencies of the object feature values in different classes, and then the object frequencies as a whole in a certain class, the maximum of which determines the object class. Simplicity, robustness and efficiency of the developed algorithm were confirmed experimentally on 9 databases.

Key words

New concept of object closeness, ordered feature values, information granulation, nearest neighbors.

1 Introduction

The concept of similarity is one of the fundamental in machine learning, since it allows you to compare data sets describe the objects of the training sample (TRS) in order to recognize objects of different classes and applying this knowledge for dividing the test sample (TS) into classes [Luger, 2016]. Usually, measure the similarity between two objects is the evaluation of their closeness by the distance between them in a metric space.This paper introduces a new closeness concept, according to which the feature values of the same class objects are close if the difference between these values is small enough. The developed classifier is based on this concept.

Obviously, this classifier operates mainly not on objects as whole and multidimensional quantities, but only on individual features of objects and one-dimensional quantities, and therefore its implementation requires the development of a completely new method.

Analysis of feature diagrams for several databases has led to a new concept of object closeness It showed the distributions of ordered feature values (OFV) differ for objects of each class. From this, it was concluded there is a fundamental possibility of classifying objects according to the frequency of the nearest neighbors based on OFV.

The computational procedure is based on approximation of features distributions of combined sample (CS) objects. It reduces to calculate the mapping of the entire data set of the CS onto a set of subsets whose elements are approximately equal to values of corresponding feature. This mapping, based on data granulation [Yao et al., 2013; Li et al., 2015] was computed as follows.

For each feature, the set of OFV of the CS objects is divided into the same number of intervals (it serves as a parameter) within which the difference between this feature values is standardized. Lists of the TRS objects of the same class falling within these intervals are called information granules. According to the new concept of closeness, objects falling into these intervals form a set of nearest neighbors. Therefore, it is approximately assumed granule and objects that form it have the same statistical characteristics.

The granule frequency is found as the frequency of the composite event of its occurrence for a certain ordered pair and class, which is calculated directly from the ration of the corresponding subsets lengths. Taking into account that the occurrence frequency of the corresponding feature value in a certain class is approximately equal to granule frequency, we find the frequency of the object in each class, as the average frequency for all features,

and then the object class corresponding to the maximum of these frequencies.

We were unable to find in the modern literature a similar conceptual approach, where the analysis of TRS is based on the existing concept of closeness and therefore aims to determine the object characteristics of various classes as a whole [Bishop, 2006; Hastie et al., 2009; Murphy, 2012]. However, the practical implementation of the proposed approach relies on existing research. It can be considered the article uses a new method for estimating nonparametric regression [Tsybakov, 2006], which combines the well-known methods of nearest neighbors and granulation.

The developed method shares common roots with the remaining unfinished general theory of pattern recognition [Grenander, 1976]. According to the theory, information at different hierarchical levels are divided into many non-overlapping blocks information at different hierarchical levels was break down into a set of non-overlapping blocks with the selection of the simplest standard blocks at the lower level. Here information granules play the role of the simplest blocks.

There is another "crossing" with existing methods. The composition and frequency of the granules depend on the above mentioned parameter. Therefore, for each of its values, we consider a different approximation of the TRS and actually analyze the properties of the sample ensemble, as is customary in the bagging method [Breiman, 1996].

In the author's papers [Shats, 2019; Shats, 2020], classification problems that were considered used a similar solution method, when the OFV properties were not yet known. The data set of problem was considered as a hierarchically organized system, where the relationships between features, objects and classes were established. In addition, those studies concerned objects with only quantitative or categorical features, used data randomization [Granichin and Polyak, 2003] by introducing sufficiently small additive components into feature values in the form of random variables evenly distributed, and used other dependencies to estimate the feature frequencies (see below).

It was shown in [Shats, 2018] that on the basis the method used in those articles it is possible to study the perception process in the sensory systems of an animal. A model for processing information stored in the brain of an animal was considered for recognizing classes of environmental objects by searching for their prototypes. Since the algorithms of the previous and the methods proposed here are basically the same, we can count the classifier is based on a bio-inspired approach.

The algorithm simplicity is the most important distinguishing feature of the new classifier. It is a direct consequence of application of the new concept of closeness, according to which all objects of the same granule are close, the object class is determined by linear functions of object feature frequency in individual classes, and these frequencies are calculated from the simplest dependencies of probability theory. Another advantage of the algorithm is its high robustness, which is provided, in accordance with [Jaynes, 2003], by grouping data into information granules.

The rest of the paper is organized as follows. The properties of OFV are discussed in Section 2. Section 3 discusses problem setting and data preparation. The calculation algorithm is discussed in Section 4. The calculation results for 9 databases are presented in Section 5. Comments on the results of the work and its development are given in Section 6.

2 Diagrams of the ordered feature values

In the paper, diagram OFV of the CS were developed, which are built for any feature as follows. Let's arrange the TRS objects in order of non-decreasing values of some feature and assign them numbers $s = 1, 2, \ldots, M$, where M is the TRS length. On segments of horizontal straight lines $i = const$, going at equal distance from each other, we mark points (t, i) corresponding to the value $t = s/M$ for objects of each class i. As a result, we will get the diagram of OFV distribution. We can build the same diagrams for others features. Obviously, the diagram set visualizes all the information contained in the TRS.

An example of such diagram for two features of the "Glass" database [Asuncion and Newman, 2007] is shown in Fig. 1. Analysis of diagrams for this and other databases have shown that the distribution OFV is substantially different for each class and feature. This property of OFV served as the basis for a new concept closeness and corresponding new classifier.

3 Problem statement and data preparation

Consider the following classification problem. Let the TRS be represented by the set $\{(\mathbf{x}_s, y_s) \mid s \in (1, M)\}$ of objects $\mathbf{x}_s = (x_{s1}, \ldots, x_{sN})^T$, which belong to disjoint classes $y_s \in (1, C)$ and have features $k \in (1, N)$. The problem is to construct an algorithm and check its quality on the TRS and TS, belonging to a single sample.

Note that the TRS features are not order statistics, since the values of any feature $\mathbf{x} = (x_1^k, \ldots, x_M^k)^T$ have different distribution functions.

Some peculiarities of the algorithm require consideration at the stage of data preparation. Here, the feature values themselves act as labels because the method operates primarily with the frequencies of the feature values. Therefore, the values of categorical features will be described by a continuous sequence of integers. In addition, the widely used normalization of features is redundant here.

If the lengths of the TRS classes differ by several times, then we are dealing with the problem of classifying unbalanced classes, the solution of which has been addressed in the literature [Lopez et al., 2013] The most popular methods simply reduce the design of a new

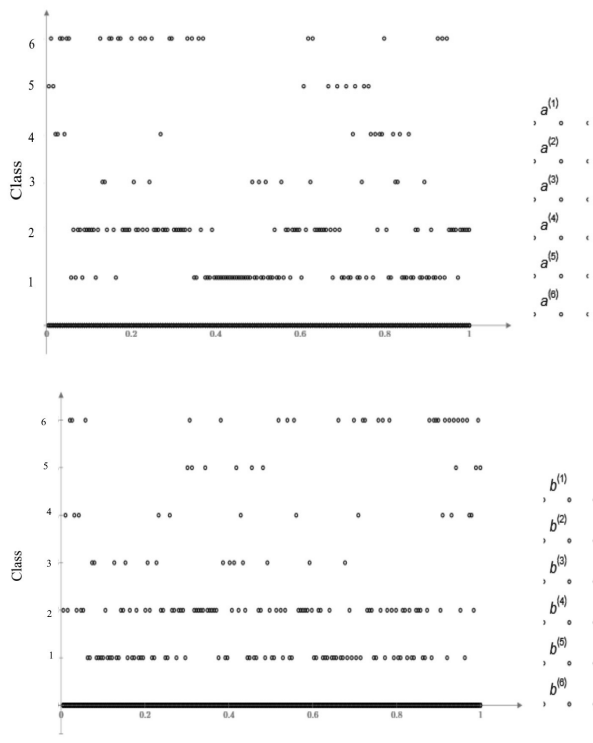

Figure 1. Diagrams of feature 1 (top) and feature 5 (lower) for the TRS of "Glass" database.

sample, where the class lengths do not differ markedly. Methods founded on information granulation are also being developed [Chen et al., 2008], where granulation is performed in several stages and begins by clustering objects in accordance with the k-means method. However, models of this type cannot be applied within the framework of the proposed concept.

When choosing a balancing scheme, we consider that a minimum number of excluded objects will preserve the information contained in the sample as much as possible. As a result of numerical experiments, the following method was developed for forming a balanced training sample (BS) from the TRS objects. The BS contains all objects of "long" classes, the length of which l_i exceeds the threshold value L. In the remaining classes each of the objects included in the BS of not less than L/l_i times.

The parameter L is taken to be equal to the half-sum of the maximum and minimum length of the TRS classes. If the quality of the classification turns out to be unsatisfactory, then it is advisable to change L and exclude objects in extremely "short" classes from the TRS. Thus, the BS contains all objects of the TRS, as well as many duplicates of objects of "short" classes, and its length can significantly exceed the length of the TRS.

Because balancing of the TRS is not used in all problems to simplify we also denote the BS by the TRS.

4 Problem solving algorithm

With the new concept of object closeness, value subsets of each feature for all objects of individual classes become the central element of calculations. To implement this change, a mapping of data onto a set of subsets was found, whose elements are roughly equal to the corresponding feature values in individual class.

The following procedure was used to calculate the mapping of an arbitrary feature k of the OBV. Let's order the feature values and then divide the entire range of its values into n equal intervals $h_k = (x_k^{max} - x_k^{min})/(n-1)$, where n is the closeness norm, parameter, x_k^{min} and x_k^{max} are the minimum and maximum values of the feature, respectively. Let us denote the boundaries of the intervals by the range of numbers $1, 2, \ldots, n$. Then, we determine the value x_k in the scale of indices according to the following definition: the index x is equal to m if $W(x/h_k) = m$, where $W(\cdot)$ is the integer part of the number. Obviously the value x_k falls within the interval $[m, m+1)$.

As a result, for each n, we get a mapping of data set of the TSR to the ordered pairs set $\{(k,m) \mid k \in (1,N), m \in (1,n)\}$. Each element of this set is a list of objects with close values of features, some lists may be empty. Any object $\mathbf{x}_s = (x_{s1}, \ldots, x_{sN})^T$ will be approximated by the vector of indices $\mathbf{d}_s = (d_{s1}, \ldots, d_{sN})^T$ with an error not exceeding step h_k for feature k, where d_{sk} is the index x_{sk}. Thereby, we have found the matrix of indices, which is an approximation of the OBV data matrix: $\|x_{sk}\| \to \|d_{mk}\|$, where $s \in (1,K), k \in (1,N), m \in (1,n)$, K is length of the CS.

Let us break down the subset of the ordered pairs set, related to the TRS objects, into subsets $\omega_i = \{((k,m),i)\}$, which we will call information granules (k,m) of class i. A granule is a list of the TRS objects of a certain class, which for any list object are considered as nearest neighbors by the value of corresponding feature. It is obvious that the set $\{\omega_i \mid i \in (1,C\}$ approximates the entire set of the TSR data.

Now let's find a set of comparative frequencies of granules that differ for ordered pairs of different classes of the TRS $\{f_{k,m}^i \mid k \in (1,N), m \in (1,n), i \in (1,C)\}$. Various estimates of $f_{k,m}^i$ are possible. In [Shats, 2019; Shats, 2020], this frequency was equal to the ratio of the number of granules ω_i to the number of objects of class i or to the total number of pairs (k,m) of all classes. In this paper f_{km}^i is taken to be equal to the frequency of the nearest neighbor of object s by feature k in class i:

$$f_{km}^i = \frac{l_{km}^i}{L_i l_{km}}, \tag{1}$$

where l_{km}^i and l_{km} are the number of granules ω_i and the total number of pairs (k,m), respectively, L_i is the number of objects of class i.

Let us denote by $p_i(d_{sk})$ the probability that the k-th feature of the object s in class i has index m. Since $p_i(d_{sk}) = p(s \in \omega_i \mid d_{sk} = m)$, then $p_i(d_{sk}) =$

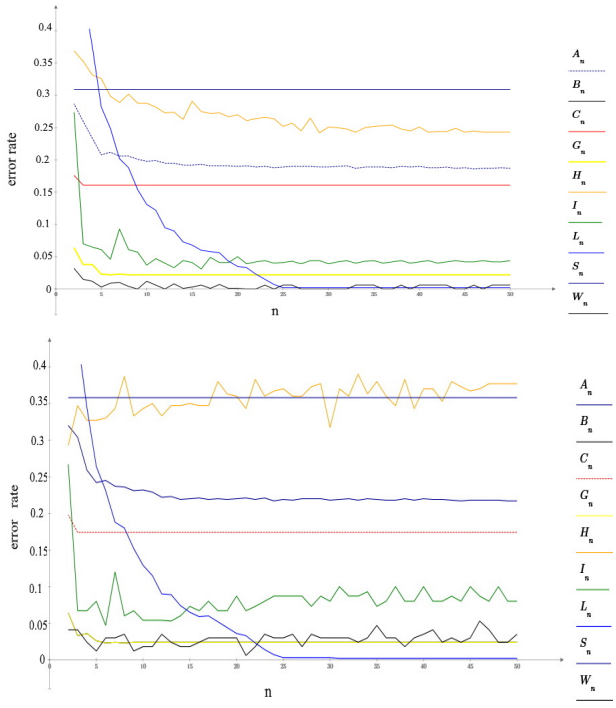

Figure 2. Panels of curves of error rates $\mu(n)$ for the TRS (top) and TS (bottom).

f^i_{km}/M. For objects of class i, the occurrence event $\mathbf{d}_s$ consists of a complete group of independent events $d_{s1}, \ldots, d_{sN}$. Then, by the total probability formula we obtain

$$p_i(\mathbf{d}_s) = \frac{1}{MN} \sum_{k=1}^{N} f^i_{k,m}. \qquad (2)$$

where the value $\frac{1}{N} \sum_{k=1}^{N} f^i_{k,m}$ is the average frequency of granules ω_i. The design class of the object s is equal to

$$I(s) = arg\, max_{1 \le i \le C} p_i(\mathbf{d}_s). \qquad (3)$$

Formulas (1)-(3) are valid for both the TRS and TS objects, since they belong to a single sample. The quality of the solution will be estimated by the error rate $\nu = \nu(n)$ for training and classification that occurs in the case of $I(s) \neq y_s$.

So, the problem is reduced to determining values range of closeness norm n, within which the solution will be acceptable with respect to the error rates for training and for classification in this problem, calculated in accordance with the cross-validation procedure. Therefore, the value of ν are calculated on some set $n \in \{2, 3, \ldots, J\}$, a subset of which, as is assumed, to contain the closeness norms for the given problem.

5 Effectiveness of the classifier

The effectiveness of the classifier was studied with nine databases from the UCI repository [Asuncion and Newman, 2007]: Adult, Breast Cancer, Car evaluation, Glass, Haberman's Survival, Iris, Letter Image, Spect and Wine. The characteristics of the bases cover a fairly wide range of values by the number of objects (267-20,000), features $(3 - 22)$ and classes $(2 - 26)$; and the database objects have quantitative, categorical, or mixed attribute types.

For all databases, with the exception of Adult and Spect database, the calculations were performed for 10 variants of splitting the CS into training and test samples according to the 10-fold cross-validation procedure. Since the obtained distributions of error rates could not were considered normal, the mean values for the considered options were taken as ν. For the marked databases, this procedure was not applied, since the composition of the TRS and TS was fixed.

Fig 2 presents errors curves ν for the TRS (top) and TS (bottom) databases that are identified by the first letters in their names. They are constructed for $J = 50$, because, as follows from the calculations, when assessing the solution quality, one can restrict oneself to the values $n \in (5.50)$. The graphs showed: values $\nu \in (0.002, 0.38)$; for each database, the classification curves have a similar shape to training curves, but go above them. For any database with a closeness norm of $25 < n \leq 50$, the values of ν for training or for classification approach some constants, and for half of the databases, the corresponding curves degenerate into horizontal straight lines $\nu = const \in (0.002, 0.38)$.

The fact of these constants existence and their nearness for the TRS and TS of the same database indicates new concept of object nearness reveals the regularities of the class distribution, and the method considered is robust. Apparently, values of specified constants are proportional to the error level with which the hypothesis the classes differ in the features distribution is fulfilled.

6 Conclusion

This paper proposes a classifier based on new concept closeness for objects finite set, according to which proximity is evaluated for each feature of objects of the same class, and not for the object as a whole.

The computational procedure is based on approximation of features distributions of the CS objects. At first, the data set of the CS is mapped onto a set pairs (k,m), and then information granules are found. They are lists of object numbers of individual classes of the TSR, which are closest neighbors in the value of corresponding feature.

The granule frequencies and any object belonging to them are defined as a complex event from the ratio of the lengths of corresponding subsets. Using these frequencies, we calculate the object classes of based on formula of total probability.

The proposed classifier differs from most existing classifiers by the simplicity of its algorithm, since it operates mainly with one-dimensional values, which are individual features values, and not multidimensional values that describe objects as a whole.

Calculations for 9 databases testify to the efficiency and robustness of the algorithm. For four databases, errors for classification and for training were less than 5%. Sufficiently high quality of the obtained results corresponds to the conclusion that the new classifier is based on a bio-inspired approach.

The results of this study appear perspective with respect to application of new concept of object closeness as well as new classifier. In particular, by considering the changing of numbers erroneously qualified objects on sequence values of closeness norms as a random process.

References

Asuncion, A. and Newman, D. (2007). *UCI Machine Learning Repository*. Irvine University of California, Irvine.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer, Berlin.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, pp. 123–140.

Chen, M., Chen, L., Hsu, C., Zeng, W., and Herrera, F. (2008). An information granulation based data mining approach for classifying imbalanced data. *Information Sciences*, **178**, pp. 3214–3227.

Granichin, O. and Polyak, B. (2003). *Randomized Algorithms of an Estimation and Optimization Under Almost Arbitrary Noises*. Nauka, Moskow.

Grenander, U. (1976). *Lectures on pattern theory*. Springer-Verlag, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, Berlin.

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, New York.

Li, J., Mei, C., Xu, W., and Quian, Y. (2015). Concept learning via granular computing: A cognitive viewpoint. *Information sciences*, **298**, pp. 447–467.

Lopez, V., Fernandez, A., Garcia, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, **250**, pp. 113–141.

Luger, G. (2016). *Artificial intelligence: structures and strategies for complex problem solving*. Addison-WeslaySpringer, Berlin.

Murphy, K. (2012). *Machine Learning. A Probabilistic Perspective*. MIT Press, Cambridge.

Shats, V. (2018). The classification of objects based on a model of perception. In *Advances in Neural Computation, Machine Learning, and Cognitive Research, Studies in Computational Intelligence*, Springer, Moscow, pp. 125–131.

Shats, V. (2019). Error-free training via information structuring in the classification problem. *Journal of Intelligent Learning Systems and Applications*, **10**, pp. 81–92.

Shats, V. (2020). Two simple classification algorithms based on information granulation. In *Proceeding XXII International Conference Neuroinformatics*, Moskow, pp. 127–133.

Tsybakov, A. (2006). *Pattern Recognition and Machine Learning*. Springer, Berlin.

Yao, J., Vasiliacos, V., and Pedrycz, W. (2013). Granular computing: Perspective and challenges. *IEEE Trans. Cibernetics*, **43** (6), pp. 1977–1989.