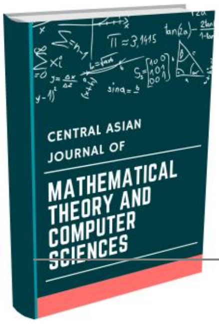


CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES

https://cajmtcs.centralasianstudies.org

Volume: 04 Issue: 12 | Dec 2023 **ISSN: 2660-5309**

DENSITY DISTRIBUTION OF DATA CONSIDERING OBJECT RELATIONSHIPS

Tursunmurotov D. X.

National University of Uzbekistan

Abstract

The problem of taking into account the influence of the distribution density of features on the structure of relationships between objects is considered. Analysis of the structure of relationships is necessary to find ways to increase the generalization ability of recognition algorithms. As a criterion for assessing the structure, it is proposed to use the values of the measure of compactness of class objects according to a given metric. A significant number of methods are based on the assumption of normal data distribution density. A technique is proposed for calculating the parameters of real density, the analytical form of which is initially unknown. The effectiveness of using real density relative to normal density is substantiated. When justifying, ordered sequences of values of differences between classes according to pairs of nominal characteristics are used. Data analysis using the proposed compactness measure is one means to solve the curse of dimensionality problem in BigData.

ARTICLEINFO

Article history:
Received 3 Oct 2023
Revised form 20 Nov 2023
Accepted 25 Dec 2023

Keywords: *compactness measure, data normalization, curse of dimensionality, distribution density.*

© 2023 Hosting by Central Asian Studies. All rights reserved.

---***---

1. Introduction

The combinatorial complexity of solving data mining problems is one of the reasons for using a large number of heuristics. Increasing the efficiency of a solution is associated with justifying the choice of heuristics[1]. A significant proportion of heuristics are based on assumptions about the density of data distribution both for individual characteristics and for their varieties [2]. There is no clear answer to the question of using criteria and performance assessments[3].

Convenient for analytical calculations and easy to interpret when analyzing data is the normal law of density distribution. The real “nature of the environment” may be far from the idealized representation according to the normal law.

Proof of the difference between real and normal distribution densities is proposed to be demonstrated through interval methods on recognition problems from two classes. The methods are designed to divide feature values into non-overlapping intervals according to special criteria. The optimal value of the criterion, provided that the number of intervals is equal to the number of classes, is interpreted as one of the compactness measures [4].

For the normal law, the lengths of the intervals are equal. In the case of real density, the value of the length or, similarly in meaning, the value of the boundaries of the intervals are determined algorithmically.

It is proposed to evaluate the differences through:

- transformation of quantitative characteristics into nominal ones, using the choice of gradations within the boundaries of non-overlapping intervals;
- ordering of features in relation to their information content [5].

In this work, the emphasis is on the problems of machine learning of metric recognition algorithms. Features of the implementation of metric algorithms are related to the relationships between class objects. One of the goals of studying the structure of relationships is to solve the problem of the curse of dimensionality [6-13]. The negative effect of the curse of dimensionality is expressed in the retraining of algorithms. Specific metrics are required to analyze the data and avoid overfitting.

As such an indicator, it is proposed to use a measure of compactness, calculated through the ratio of pairwise connectedness of class objects according to the hyperball system . The length of the hyperball radii depends on the choice of metric for calculating the distance.

It is convenient to interpret the set of admissible values of a measure in terms of fuzzy logic and trace their connection with estimates of the generalizing ability of algorithms. Among the factors influencing the value of the compactness measure under consideration, the dominant role is played by the choice of metrics.

There is a need to demonstrate the difference and relationship between compactness measures on the number line and in metric space. On the numerical axis, the measure of compactness is synonymous with the pattern of a characteristic, regardless of its origin (raw, latent) and the scale of its measurement. Therefore, the relations “more”, “less”, “equal” can be transformed into relations of a specific subject area and used to fill knowledge bases.

2. Formulation of the problem

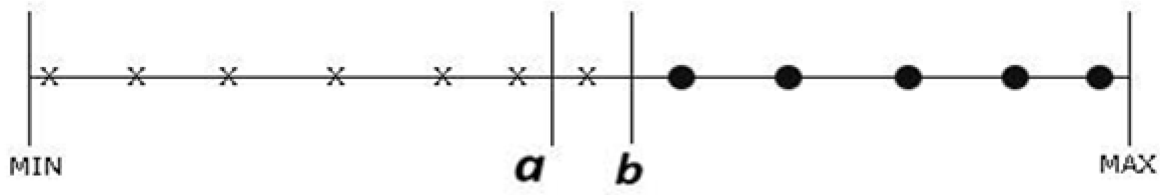
The recognition problem in the standard formulation is considered. It is considered that a set of $E_0 = \{S_1, ..., S_m\}$ objects is given, divided into two disjoint classes K_1 и K_2, $E_0 = K_1 \cup K_2$. Objects are described using set of n different types signs $X(n) = (x_1, ..., x_n)$. On many objects E_0 metric $\rho(x, y)$ is given. Defined $L(E_0, \rho)$ – a set of boundary objects of classes on E_0 by metric $\rho(x, y)$. Objects $S_i, S_j \in K_t$, $t = 1.2$ are considered related $S_i \leftrightarrow S_j$ if $\{S \in L(E_0, \rho) \mid \rho(S, S_i) < r_i \ \ u \ \ \rho(S, S_j) < r_j\} \neq \varnothing$, Where $\boldsymbol{r_i}$ (r_j) – distance to the nearest $S_i(S_j)$ object from K_{3-t} the metric $\rho(x, y)$.

It is believed that there are no E_0 procedures $\mu(x)$ defined for calculating , $\eta(E_0, X(\delta), \rho)$ respectively, the measure of compactness of a heterogeneous feature $x \in X(n)$ and $X(\delta) \subset X(n)$ the measure of compactness of a sample of objects E_0 on a set using $X(\delta)$ the metric $\rho(x, y)$.

You need to define the dependency:

- measures of compactness $\eta(E_0, X(\delta), \rho)$ from the values of the measure of compactness $\mu(x)$;
- the generalizing ability of the Bayesian decision rule on the values of the compactness measure $\eta(E_0, X(\delta), \rho)$.

An illustration of the placement of the boundary between two classes, obtained respectively for the normal and real distribution density p, is shown in Fig. 1



Rice. 1. Boundaries between classes according to normal (*a*) and real (*b*) distribution density

Meaning _ boundaries at normal density in Fig. 1 $a=\frac{MIN+MAX}{2}$. The limit for the real density was obtained using the criterion

$$\left(\frac{\sum_{p=1}^{2}\sum_{i=1}^{2}u_i^p\left(u_i^p-1\right)}{\sum_{t=1}^{2}|K_t|\left(|K_t|-1\right)}\right)\left(\frac{\sum_{p=1}^{2}u_1^p\left(|K_2|-u_2^p\right)+u_2^p\left(|K_1|-u_1^p\right)}{2|K_1||K_2|}\right)\to\max_{c_1<c_2<c_3}, \quad (1)$$

Where c_1 = MIN , c_3 = MAX , c_2 – optimal according to (1) interval boundary [c_1 ; c_2], (c_2 ; c_3] and $b=\frac{c_2+\tau}{2}$, where is τ the closest c_2 value from (c_2 ; c_3].

Algorithm *k* nearest neighbors (KNN) implements a Bayesian decision rule. The choice of a neighborhood of *k nearest neighbors* that is optimal according to the cross- validation criterion allows, when solving the problem of generalization ability, not to take into account the presence of noise objects in the training sample. Ignoring noise objects leads to a sharp decrease in accuracy on the control sample when recognizing using the nearest neighbor (NN) algorithm.

NN algorithm over KNN in terms of computational resource costs is:

- use of minimum coverage of a sample of objects to recognize standards;
- application of local metrics for each object - the standard of learning.

The measure of compactness $\eta(E_0, X(\delta), \rho)$ is determined by the binary relation of connectedness of objects. According to this relationship, the sample objects form disjoint groups $G_1,\ldots,G_t, t\geq 2$. The compactness of class objects K_i is defined as $\eta^i(K_i, X(\delta), \rho)=\frac{\sum_{G_t\in K_i}|G_t|}{|K_i|}$ and the selection as a whole as

$$\eta(E_0, X(\delta), \rho)=\frac{\eta^1(K_1, X(\delta), \rho)|K_1|+\eta^2(K_2, X(\delta), \rho)|K_2|}{m}. \quad (2)$$

It is proposed to select a space $X(\delta)\subset X(n)$ through a sequence of features ordered in relation to information content. It is argued that the order in the sequence depends on the assumption (hypothesis) about the normal or real law of distribution of features.

For this purpose, a matrix of pairwise differences is constructed $B=\{b_{ij}\}_{n\times n}$ based on nominal characteristics

$$b_{ij} = \begin{cases} \dfrac{\sum\limits_{u=1}^{m}\sum\limits_{v=1}^{m}\alpha(u,v)g(u,v,i,j)}{2\sum\limits_{p=1}^{l}|K_p|(m-|K_p|)}, i \neq j, \\ 0,\ i = j, \end{cases} \quad (3)$$

Where

$$g(u,v,i,j) = \begin{cases} 2, x_{ui} \neq x_{vi}\ u \quad x_{uj} \neq x_{vj}, \\ 1, x_{ui} = x_{vi}\ или \quad x_{uj} = x_{vj}, \\ 0, x_{ui} = x_{vi}\ u \quad x_{uj} = x_{vj}. \end{cases} \text{ And } \alpha(u,v) = \begin{cases} 0, S_u, S_v \in K_i, i = 1,2 \\ 1, S_u \in K_i, S_v \in K_j, i \neq j. \end{cases}$$

To convert quantitative characteristics into nominal ones, depending on the assumption of the distribution law, the boundary *a* or *b is used (see Fig. 1)* . Ranking by pairs of features using (3) is invariant to the scale of measurement of quantitative features. As a way to transform different types of characteristics into quantitative ones, the method of calculating generalized assessments of objects can be considered.

The subject of the study is to establish a connection between the accuracy of the KNN algorithm and the measure of compactness. When implementing KNN [14] and calculating the compactness measure (2), the same basic metric is used. The implementation of the NN algorithm is based on local metrics. For this reason, instead of compactness (2), it is necessary to use other indicators to evaluate accuracy.

3. Computational experiment

The purpose of the computational experiment is to establish the relationship between the accuracy indicators of the KNN algorithm at the optimal value of k and the compactness measure (2) on sets obtained $X(\delta) \subset X(n)$ by ordering features by $B = \{b_{ij}\}_{n \times n}$. The aritmiya sample [15] was used as data for the experiment . Sample objects are described by 150 characteristics, 149 of which are quantitative, 1 nominal, $|K_1|$ = 92, $|K_2|$= 76. The sequence of features according to (3) when ranking by normal density was as follows:

138,146,9,77,48,51,140,148,41,43,66,130,76,93,65,91,71,94,26,80,101,102,118,131,58,87,16,31,5,107,0,105,125,132,61,141,2,68 , 18,33,111,115,67,139,116,123,45,136,20,29,3,39,75,100,36,113,83,89,103,108,24,79,21,97,17,25,11,60,44, 46,124,147,23,59,121, 128,99,119, 14,54,42,53,37,149,13,56,28,32,106,114,6,64,34,120,19,127,81,92,110,117,22,73,72,144,27,88,129,137,4,84,70,96,10, 63,55,143,12,50,1,126,15,62,104,112,78,90,30,52,82,109,57,122,74,98,47,49,133,135,40,69,35,38,134, 142,8,145,7,86, 85,95,150.

KNN experiment, features were removed from the end (from right to left) from an ordered sequence.

Tab. 1 Results of the experiment according to (3) when ranking by normal density

Power dial	*k* values	Accuracy %	Compactness
150	1	59 , 52	0.3456
130	1	59.52	0.2360
110	1	60.11	0.2677
90	eleven	63.69	0.1956
70	9	63.09	0.2786
50	1	61.90	0.1807
thirty	5	64.88	0.2339

10	33	57.73	0.0600

Tab. 2 Experiment results using (3) when ranking normal density and normalizing in [0;1]

Power dial	*k* values	Accuracy %	Compactness
150	3	62.5 0	0.3158
130	5	61.90	0.3690
110	3	62.50	0.3238
90	3	60.11	0.4041
70	5	62.50	0.4290
50	1	58.33	0.2282
thirty	7	62.50	0.2318
10	1	57.14	0.0542

The sequence of features according to (3) when ranking by real density was as follows:

21,72,32,98,9,77,25,80,11,54,38,40,66,84,27,88,19,70,35,37,17,64,133,141,29,90,51,144,137,145, 7,92,138,146,45,48,15,62,115,122,41,43,20,71,13,56,74,99,26,85,34,104,6,28,128,136,8,91,52,142, 111,119,16,31,18,33,2,5,44,47,0,1,57,83,75,100,58,108,42,121,24,79,93,94,61,78,46,49,106,109, 14,30,36,39,105,113,69,95,126,130,3,10,23,101,12,103,107,129,87,89,125,132,110,117,123,149,67,102,120 ,124,86,118,55,135,127,143,63,97,50,59,131,139,4,116,68,140,60,114,53,148,81,96,65,147,22,73, 76,134,82,112,150.

Tab. 3 Results of the experiment according to (3) when ranking by real density

Power dial	*k* values	Accuracy %	Compactness
150	1	59.52	0.3456
130	7	59.52	0.2532
110	25	60.11	0.1362
90	99	63.69	0.1469
70	21	60.11	0.0978
50	25	60.11	0.1075
thirty	1	62.50	0.0733
10	5	57.73	0.0734

Table 4 Results of the experiment according to (3) when ranking by real density and normalizing in [0;1].

Power dial	*k* values	Accuracy %	Compactness
150	3	62.50	0.3158
130	9	63.09	0.1943
110	9	60.71	0.1947
90	85	61.90	0.1537
70	5	59.52	0.1603
50	13	59.52	0.1518
thirty	5	61.90	0.1383
10	3	66.07	0.0591

For a comparative analysis of the ordered sequence of features at different distribution densities, value (3) was used. The first 5 pairs of features from the ordered sequences are given in table. 5.

Tab. 5. Ordering by values (3) at different distribution densities

Density according to law			
To normal		real	
pair	Difference	pair	difference
138 – 146	0.6869 _ _	21 - 72	0.7462
9 – 77	0.6707	32 – 98	0.7392
48 -51	0.6582	9 – 77	0.7242
140 – 148	0.6575	25 - 80	0.7232
41 – 43	0.6497	11 - 54	0.7062

The value of the differences according to (3) (see Table 5) for the real density is greater than for the normal one. This proves the effectiveness of using criterion (1) for data analysis.

4. Conclusion

The experimental results show that there is no strong correlation between the accuracy of KNN and the values of compactness measures for sets of features. New opportunities are opening up for the selection of informative sets of features associated with taking into account the density of data distribution.

Literature

1. Zhuravlev Yu.I. On an algebraic approach to solving problems of recognition or classification // Problems of Cybernetics. 1978. T. 33. P. 5–68.
2. Vorontsov K.V. Mathematical methods of teaching using precedents. MIPT course of lectures, 2006.
3. K.V. Rudakov. On some Factorizations of semimetric cones and estimates quality of heuristic metrics in data analysis tasks.
4. Zagoruiko N. G. Hypotheses of compactness and λ -compactness in data analysis methods // Sib . magazine industrial _ mathematics _ 1998. T.1, No. 1. pp. 114-126.
5. Zinoviev A.Yu., Visualization of multidimensional data, Krasnoyarsk, Publishing House KSTU, 2000.180
6. Saidov D.Yu. Information models based on nonlinear transformations of feature space in recognition problems: Diss . ... Doctor of Philosophy (PhD) in physical and mathematical sciences. Tashkent, 2017.- 93 p .
7. Ignatyev NA, Structure Choice for Relations between Objects in Metric Classification Algorithms // Pattern Recognition and Image Analysis. 2018. V . 28. No. 4. P. 590–597.
8. Zagoruiko N.G., Kutnenko O.A., Zyryanov A.O., Levanov YES . Learning pattern recognition without retraining // Machine learning and data analysis, 2014. Vol. 1 . No. 7. pp. 891–901.
9. Mirzaev 2021 – Mirzaev A.I. On the choice of space for describing objects in machine learning on large data samples // Problems of Computational and Applied Mathematics. No. 6 (36) 2021, pp. 120 – 127.
10. Ignatyev NA On Nonlinear Transformations of Features Based on the Functions of Objects Belonging to Classes // Pattern Recognition and Image Analysis. 2021. V. 31. No. 2. P. 197-204.
11. Adilova FT, Ignat'ev NA, Madrakhimov Sh.F. _ The Approach to Individualized Teleconsultations of Patients with Arterial Hypertension // Global Telemedicine and eHealth Updates: Knowledge Resources, Vol. 3, 2010. –P.372-375.
12. Zhamby M. Hierarchical cluster analysis and correspondence: Transl. from fr. - M.: Finance and Statistics, 1988. 342 With .

13. Gyamfi , KS, Brusey , J, Hunt, A & Gaura , E 2018, 'Linear dimensionality reduction for classification via a sequential Bayes error minimization with an application to flow meter diagnostics' Expert Systems with Applications, vol 91, pp. 252-262 https://dx.doi.org/10.1016/j.eswa.2017.09.010
14. http s :/ scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
15. https://archive.ics.uci.edu/ml/datasets/arrhythmia