Analysis of the structure of the relationship between the descriptions of objects of classes and evaluation of their compactness

E N Zguralskaya[1]

[1]Ulyanovsk Technical University. Institute of Aviation Technologies and Management, Sozidateley avenue, 13A, Ulyanovsk, Russia, 432072

e-mail: iatu@inbox.ru

Abstract. The study is conducted to assess the compactness of descriptions of objects of classes on the numerical axis and in the multidimensional attribute space. The computation of compactness is possible only in the defined boundaries of areas of the attribute space. In the one-dimensional case, the boundaries are calculated by the frequency of occurrence of the values of features of objects of classes in the interval. In the multidimensional case, a subset of the boundary objects of the classes is used for a given metric. A comparative analysis is given of the values of the compactness measure by latent attributes on the numerical axis and by the sets of initial features from which they are synthesized.

1. Introduction

In the pattern recognition theory objects are structured into classes based on the compactness hypothesis. Under this hypothesis, “close” objects shall belong to the same class. It is necessary to clarify (interpret) the terms “closeness” and “compactness” of objects.

No common determination of the “compactness” term has been adopted. [1] postulates a compactness measure of disjoint groups, set of admissible values of which is determined in (0, 1] and depends on structure of relations between objects. The following factors affecting values of compactness are pointed out:

- the choice of the metric to compute distances between objects;
- the dimension of the attribute space;
- the choice of the way to scale and normalize data;
- the usage of methods to select informative collections of attributes;
- conditions to select and remove noise objects from the sample;
- the number of standard objects of the minimal coverage of the learning sample;
- linear and nonlinear transformations of the attribute space for the description of the objects.

The aim of the searching for extremal values of compactness measures on the variety of parameters listed above is to improve generalizing ability of recognition algorithms. The method to obtain a quantitative estimate for the pattern compactness, described in [2], is based on the usage of the function of competitive similarity between objects (FRiS-functions). Using the FRiS-function, one can describe all distributions of classes by collections of standard objects. The collection of objects allows one to find the compactness measure of the whole sample or each separate object of the class and to clear the learning sample from objects adding negative contributions to the value compactness.

Implementation of machine learning algorithms becomes significantly more complicated when the dimension of the data is large. A geometric interpretation of origin of the effect of curse of dimensionality is given in [3]. The effect of curse of dimensionality arises from the fact that the number of possible sets of attributes in the description of objects significantly exceeds the number of training examples. Learning algorithm can only support correct generalization provided that the number of examples from learning sample is enough.

Compactness implies the existence of a boundary between areas of attribute space with a description of objects from different classes.

Numerical methods to obtain a quantitative estimate for compactness are differentiated as well. For one-dimensional cases the interval methods are used while for multidimensional cases – computation of measure of compactness of objects of classes and samples in a whole for a given metric. What both one-dimensional and multidimensional cases have in common is the existence of areas of attribute space on boundaries of which measure of compactness is computed.

For a one-dimensional case the objects can be compared on the numerical axis by values of its initial and latent attributes using relations “greater than”, “less than” or “equal to”.

When the measure of compactness is computed for a multidimensional case in [1] the property of connectedness of objects along the subset (spans) of boundary objects of disjoint groups is used. Based on this property the objects are decomposed into disjoint groups. Connectedness of objects S_i, S_j is treated as property of logical regularities in form of hyperballs with these objects being its centre. S_i and S_j objects are considered bound if their intersection contains spans objects. Any pair of objects (S_i,S_j) of one group can always be linked by a chain of connected objects. Ideally, all class objects shall represent one group of connected objects.

This paperwork reviews structure of relations between class objects on the numerical axis. It is suggested to use measures of compactness, computed through decomposition of either attributes values (initial and latent) or values of distance between the objects into intervals, as a research tool. Values of measure of compactness are used to detect latent patterns in data. Such patterns can be regarded as new knowledge obtained within the frames of information models of ill-structured subject areas.

2. Criteria for decomposition of attributes into intervals

Let us consider two computing algorithms put forward in [4, 5] to optimize criteria for decomposition of attributes values into intervals. For convenience let these criteria be referred to as CR1 and CR2.

When computing with respect to CR1 number of intervals on the ordered sequence of attribute values equals to number of disjoint classes. Values of interval boundaries are determined via the maximum of product of intraclass similarity and interclass difference. Ideally every interval shall be represented by all attribute values of objects of one class.

For the CR2 criterion the number of classes is 2, the number of intervals is equal to or greater than 2. When computing boundaries of disjoint intervals, number of which is initially unknown, the absolute difference in frequency of occurrence of attribute values (both initial and latent) in the description of objects of two classes is used. The values of attributes on the numerical axis form a sequence of clusters (intervals). There should not be two neighboring clusters in which representatives of one class would dominate (in terms of frequency of occurrence). Those decompositions are considered ideal in the sense of consistency, for which values of (not necessarily all) objects of only one class are contained within the boundaries of each interval.

The set of admissible values by the CR1 criterion and the consistent decomposition of attributes into intervals over CR2 are contained in the segment [0; 1] and are further considered as a measure of their compactness. The value 1 corresponds to a perfect decomposition with respect to CR1 and CR2. The degree of deviation from the ideal can be inferred by values less than 1.

Combined use of CR1 and CR2 criteria is necessary to detect latent patterns in data. The search for patterns is based on results of a computational experiment. To interpret the results of the experiment, known forms of logical regularities are used (hyperball, half-plane, parallelepiped).

Let a set of objects $E_0=\{S_1,\ldots,S_m\}$ be given, containing representatives d of disjoint classes $K_1,...,K_d$. The objects are described using a set of n different types of $X\ (n)$ attributes, $\delta\ (\delta <n)$ of which are measured in nominal, $n-\delta$ in interval scales. Gaps and duplicate values in data are allowed.

Search for latent, that is, hidden attributes is of great interest as those can be very informative in the classification being one of the objectives of the present study. It is believed that the CR1 and CR2 criteria are used to decompose values of a quantitative attribute (both initial and latent) into disjoint intervals. Latent attributes can represent combinations of nominal and quantitative attributes [6, 7]. Required to determine:

- method to compute latent attributes;
- boundaries of intervals and CR1 criterion values on initial and latent attributes;
- number of intervals, values of their boundaries and consistency of decomposition of initial attributes by the CR2 criterion.

A variety of methods to form latent attributes and criteria for decomposition of their values into disjoint intervals is essential to detect hidden patterns in databases of subject areas. We will obtain latent attributes from a set $X\ (n)$ in form of combinations of $x_i{*}x_j$ and x_i/x_j. If number of gradations of a nominal attribute is equal to the number of disjoint classes of objects, then they can always be associated with a set of integers $a_1,\ldots,a_d$, where $a_i\neq 0,\ i=1,\ldots,d$ and $a_{j+1}\text{-}a_j=const$, where $j=1,\ldots,d\text{-}1$. Each disjoint interval in respect with CR1 will be represented by one value. For example, if number of gradations equals to 2, the choice of values from [-1, 1] would constitute a representation form convenient for calculating.

3. Selection of information attribute sets by compactness of objects

Let the metric $\rho(x,\ y)$ be defined on the set of attributes $X(h)\subset X(n),\ 1\leq h\leq n$. The object $S\in E_0\cap K_p$, p=1,2 is considered as a centre of a hyperball from which, according to the ordered set of objects $\{S,S^1,\ldots,S^{m-1}\}=E_0$, a sequence of hyperballs nested in each other is formed having radii

$$\rho(S,S)\leq\rho(S,S^1)\leq\rho(S,S^2)\leq\ldots\leq\rho(S,S^{m-1}). \quad (1)$$

Values of boundaries of intervals of each object $S\in E_0\cap K_p$, p=1, 2 to E_0, computed by the CR1 criterion at (1), are used as a means to select an informative set of diverse attributes from $X\ (n)$. The geometric interpretation of formation of the ordered set $\{S,\ S^1,\ldots,\ S^{m-1}\}$ is shown at "Figure 1".

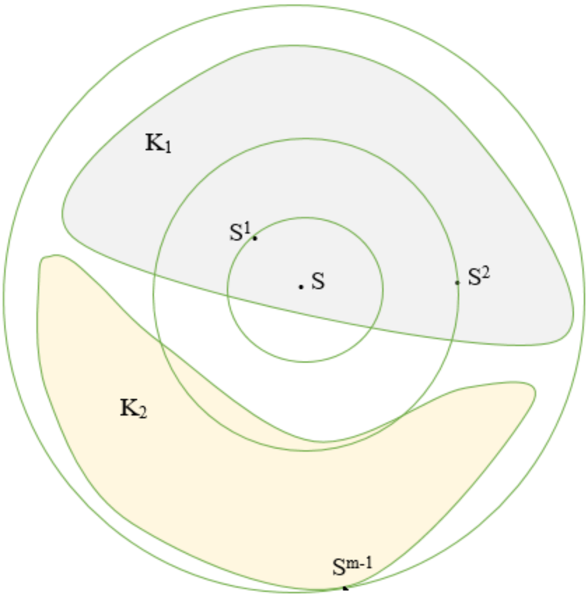

Figure 1. Sequence of nested hyperballs

Let the boundaries of the intervals $[c_1,c_2]$, $(c_2,c_3]$, c_1=ρ(S,S)=0 be defined at (1) in respect with CR1. Estimate of compactness of the object $S \in K_p$ by the set of attributes $X(h) \subset X(n)$ is calculated as

$$\varphi(S,X(h))=\theta_1(1-\theta_2), \qquad (2)$$

where

$$\theta_1=\frac{\left|\left\{S^i \in K_p \middle| \rho\left(S,S^i\right)\in[c_1,c_2]\right\}\right|}{\left|K_p\right|}, \qquad \theta_2=\frac{\left|\left\{S^i \in K_{3-p} \middle| \rho\left(S,S^i\right)\in[c_1,c_2]\right\}\right|}{\left|K_{3-p}\right|}.$$

As a means to reduce combinatorial complexity when searching for logical regularities in [8], it is offered to use hierarchical grouping methods. It is noted how important a choice of the first step for such search for patterns is.

How these hierarchical agglomerative grouping methods are implemented depends on the rules used in them. Let $\varphi(X(h),S_i)$ be the compactness estimate (2) of the object $S_i \in E_0$ на $X(h)$. When forming the set *X(h+1) of X(h)* it is necessary to calculate

$$R\left(X\left(h+1\right)\right)=\frac{1}{m}\sum_{i=1}^{m}\begin{cases}1, \varphi\left(X\left(h+1\right),S_i\right)\geq\varphi\left(X\left(h\right),S_i\right),\\ 0, \varphi\left(X\left(h+1\right),S_i\right)<\varphi\left(X\left(h\right),S_i\right).\end{cases} \qquad (3)$$

The condition (rule) for adding a attribute $x_j \in X(n) \backslash X(h)$ in $X(h+1)$ is:

$$R\left(X\left(h\right)\cup\left\{x_j\right\}\right)>\frac{1}{2} \quad u \quad R\left(X\left(h\right)\cup\left\{x_j\right\}\right)=\max_{x_i\in X\left(n\right)\backslash X\left(h\right)}R\left(X\left(h\right)\cup\left\{x_i\right\}\right). \qquad (4)$$

The group (set) $X(h)$ is considered as formed if the attribute $x_j \in X(n) \backslash X(h)$, for which (4) holds true, does not exist.

One peculiarity of Bigdata methods is analysis of data samples in which the number of attributes is greater than or equal to the number of objects. The number of groups into which the set of attributes *X (n)* is decomposed by (4) is initially unknown. It has been experimentally proved in [9] that, when majority rules in hierarchical agglomerative grouping are used, the informativeness of each subsequent group of attributes is less than that of a preceding one. Group formation sequence is determined by the principle of dynamic programming. For this reason, composition of attributes of the first group is considered as an informative set.

As the first step in selecting an informative set of attributes, it is proposed to choose a subset $Y \subset X(n)$ consisting of one or two attributes. The subset Y shall satisfy the following requirement:

$$B(Y)=\max_{\{i,j\}\in X\,(n)}\sum_{d=1}^{m}\left|\left\{S_j\in K_p\middle|\rho\left(S_j,S_d\right)<R, \qquad R=\min_{S_c\in K_{3-p}}\rho\left(S_c,S_d\right)\right\}\right|. \qquad (5)$$

4. Computation experiment

Let us review the results of decomposition of quantitative attributes into disjoint intervals with respect of the CR1 and CR2 criteria on a data sample from [10].

The sample consists of two classes - K_1 and K_2 - and contains data on cardiovascular diseases. The description of objects is given by the following set of attributes $X(13)=(x_1,\ldots,x_{13})$. Number of objects of class K_1 is 150, ones of class $K_2$is 120. x_1, x_4, x_5, x_8, x_{10}, x_{11}, x_{12} are quantitative attributes, while x_2, x_3, x_6, x_7, x_9, x_{13} are nominal. The nominal attributes x_2, x_6, x_9 have two gradations (i.e., number of gradations of a attribute is equal to number of classes).

The compactness of the quantitative attributes from *X(13)* and the limits of the CR1 intervals are given in Table 1.

The product of intraclass similarity and interclass difference is used to compute nominal attribute weights (as well as quantitative attributes compactness) by CR1. If the values of gradations in the description of objects of each class do not intersect each other, then the weight of the nominal attribute equals to 1. Table 2 shows the values of all six nominal attributes.

Table 1. Interval boundaries and values of compactness by CR1.

Attribute	Attribute Information	Interval boundaries	Compactness
x_1	Age	[29..54] (54..77]	0.2871
x_4	Resting blood pressure	[94..135] (135..200]	0.2548
x_5	Serum cholestoral in mg/dl	[126..252] (252..564]	0.2684
x_8	maximum heart rate achieved	[71..147] (147..202]	0.3413
x_{10}	Oldpeak = ST depression induced by exercise relative to rest	[0..1.6] (1.6..6.2]	0.3177
x_{11}	The slope of the peak exercise ST segment	(1..2] (2..3]	0.3246
x_{12}	Number of major vessels (0-3) colored by flourosopy	[0..1] (1..3]	0.3772

Table 2. Nominal attribute weights.

Attribute	Attribute Information	Weight
x_2	Sex	0.2727
x_3	Chest pain type (4 values)	0.3203
x_6	Fasting blood sugar > 120 mg/dl	0.1873
x_7	Resting electrocardiographic results	0.2762
x_9	Exercise induced angina	0.3453
x_{13}	Thal: 3 = normal; 6 = fixed defect; 7 = reversable defect	0.4193

As is seen from the Table 1 and Table 2, the compactness of quantitative (attributes) by CR1 and the values of nominal attribute weights are very different from the ideal ones. The value of quantitative attribute within the disjoint interval by CR1 can be considered as a gradation (interval number) in the nominal measurement scale. In such a description of objects the attribute weight in nominal scale will coincide with the value of compactness in respect with the CR1 criterion.

The number of disjoint intervals and the stability of the decomposition by the CR2 criterion are given in the Table 3 below.

Table 3. Attribute stability and interval boundaries by CR2.

Attribute	Interval boundaries	Stability
x_1	[29..54], [55..70], [71..76], [77..77]	0.6571
x_4	[94..122], [123..200]	0.5585
x_5	[126..160], [164..174], [175..245], [246..353], [354..394], [407..409], [417..564]	0.6309
x_8	[71..147], [148..194], [195..195], [202..202]	0.7030
x_{10}	[0..0.8], [0.9..6.2]	0.6957
x_{12}	[0.. 0], [1..3]	0.7316

As is seen from the Table 1 and Table 3, relatively high values of compactness are obtained for the attribute x_{12}.

Table 4 gives information about decomposition of latent attributes into two intervals obtained from the operations of multiplication and division of values of initial attributes.

Analysis of the results from the Table 4 and Table 1 shows that it is in fact feasible to search hidden patterns by latent attributes, the compactness of which is higher than each of the initial attributes that make up their composition.

Let a set of indices of respective quantitative and nominal attributes in the set $X(n)$ be designated by I, J . In order to unify the measurement scales, we will display the values of quantitative attributes by a linear fractional transformation in [0; 1]. The Zhuravlev metric will be used as a measure of the distance between the objects S_u, $S_v \in E_0$ ($S_c=(a_{c1},\dots,a_{cn})$, $c=1,\dots,m$) for selection of informative attributes

$$\rho\left(S_u, S_v\right) = \sum_{i \in I} \left|a_{ui} - a_{vi}\right| + \sum_{i \in J} \begin{cases} 1, a_{ui} \neq a_{vi}, \\ 0, a_{ui} = a_{vi}. \end{cases}$$

Table 4. Interval boundaries for latent attributes and compactness values by CR1.

Latent attribute	Interval boundaries	Compactness
x_4*x_9	[-192..105] (105..200]	0.3552
x_8*x_9	[-202..-115] (-115..186]	0.3718
$x_{10}*x_{11}$	[1..3.3] (3.3..21.6]	0.3684
x_2/x_8	[-0.0104..0.0067] (0.0067..0.0140]	0.3597
x_8/x_{10}	[16.8182..66.6667] (66.6667..202]	0.3726
x_8/x_{11}	[32..75] (75..202]	0.3555
x_9/x_4	[-0.0106..-0.0062] (-0.0062..0.0106]	0.3504
x_9/x_8	[-0.0140..0.0061] (0.0061..0.0113]	0.3523
x_{10}/x_8	[0.0049..0.0149] (0.0149..0.0594]	0.3726
x_{11}/x_8	[0.0049..0.0132] (0.0132..0.0312]	0.3555

The first pair of attributes added to the informative set in (5) is (x_8, x_{13}). The process of stepwise selection of attributes according to (4) is shown in the Table 5.

Table 5. Stepwise selection of informative attributes according to (4).

Number of attributes *h* in set	Attribute added in *X*(*h*-1)	Value *R*(*h*) acc. to (3)
3	x_{10}	0.6926
4	x_{12}	0.6815
5	x_5	0.6852
6	x_9	0.6185
7	x_4	0.6222
8	x_3	0.6111

As a result of stepwise selection (see the Table 5) the informative set of attributes $X(8)=(x_3,x_4,x_5,x_8,x_9,x_{10},x_{12},x_{13})$ is obtained.

5. Conclusions

Two interval methods have been offered to analyze the structure of relations between objects of disjoint classes by quantitative initial and latent attributes. Numerical estimates of the structure of relations by these methods differ in that the number of disjoint intervals can be initially known or determined by the algorithm. The rule of hierarchical agglomerative grouping for the formation of an informative attribute set has been described.

The considered methods are recommended to be used to search for hidden patterns in data when developing information models based on knowledge.

6. References

[1] Ignatyev N A 2018 Structure Choice for Relations between Objects in Metric Classification Algorithms *Pattern Recognition and Image Analysis* **28** 590-597

[2] Zagoruiko N G, Kutnenko O A, Zyryanov A O and Levanov D A 2014 Learning to recognize patterns without retraining *Machine Learning and Data Analysis* **1** 891-901

[3] Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge: MIT Press) p 652

[4] Zguralskaya E N 2018 Stability of data partitioning into intervals in the tasks of recognition and the search for hidden patterns *Proceedings of the Samara Scientific Center of the Russian Academy of Sciences* **4** 826-829

[5] Zguralskaya E N 2012 Selection of informative features for solving classification problems using artificial neural networks *Neurocomputers: development, application* 20-27

[6] Ignatyev N A 2011 Calculation of generalized indicators and data mining *Automation and Remote Control* 183-190

[7] Saidov D Y 2017 Data visualization and its proof by compactness criterion of objects of classes *International Journal of Intelligent Systems and Applications (IJISA)* **9** 51-58

[8] Duke V A 2005 *Methodology of the search for logical laws in the subject area with fuzzy systemology: an example of clinical and experimental studies* URL: https://dlib.rsl.ru/viewer/01002930373#?page=1

[9] Madrakhimov Sh 2018 Calculation of the Generalized Estimations in Sets of Features and their Interpretation *International Journal of Software Engineering and Its Applications* **12** 29-38

[10] UCI repository of machine learning databases URL: http://archive.ics.uci.edu