ACADEMICIA
An International Multidisciplinary Research Journal
(Double Blind Refereed & Peer Reviewed Journal)

ACADEMICIA
ISSN (online) : 2249 -7137
ACADEMICIA
An International Multidisciplinary Research Journal
published by:
South Asian Academic Research Journals.
A Publication of CDL College of Education
(Affiliated to Kurukshetra University
Kurukshetra,India).

DOI: 10.5958/2249-7137.2021.01318.5

SEARCH FOR REGULARITIES BY INTERVALS OF VALUES OF QUANTITATIVE FEATURES

Shavkat Fayzullaevich Madrakhimov*; Alisher Khasanovich Khurramov**

*Associate Professor,
Doctor of Physical and Mathematical Sciences,
Head of the Department of “Algorithms and Programming Technologies”,
Faculty of Mathematics, National University of UZBEKISTAN
Email id: mshavkat@yandex.ru

**Lecturer,
Department of Applied Mathematics,
Karshi State University, UZBEKISTAN
Email id: xurramov2009@yandex.ru

ABSTRACT

The problem of searching for hidden patterns in a sample of objects described by the initial and combined features is considered. The knowledge obtained on the basis of dividing the values of quantitative attributes into intervals is presented in the form of a set of fuzzy inference rules. The methodology of using these rules for decision making is described.

KEYWORDS: *Hidden Patterns, Latent Features, Domination Intervals, Partition Stability, Fuzzy Inference Rules.*

INTRODUCTION

Regularities are relationships (explicit and hidden) between properties (features) of objects. One of the ways to identify patterns is to divide the values of quantitative features into intervals and search for patterns by them [1, p. 19; 2, pp. 217-225].

In [3, p.310], a method is described for the preliminary analysis of training information, based on finding an informative zone in the training material (such sub-descriptions (or fragments of descriptions) are considered informative if they allow one to distinguish objects from different classes or distinguish a given object from all objects that do not belong to the same class. as considered) and typical for their classes of objects (the most representative).

THE MAIN FINDINGS AND RESULTS

The search for informative zones is based on the use of the apparatus of discrete mathematics, in particular, Boolean algebra, the theory of disjunctive normal forms, the theory of coverings of Boolean and integer matrices. So, in the estimation algorithms developed by Y.I. Juravlev and his students, estimates of ensembles of features are found considered in [4, p. 916; 6, pp. 32-40], which are generalizations of the coefficients of information content.

It is possible to introduce taxonomy into the division of features into intervals - statistical methods for dividing features into equal or unequal intervals. The methods of partitioning into equal intervals include histograms, dicylic partitioning and so forth. In [7, p. 447], the criterion for partitioning into intervals is based on the analysis of the probability distribution density.

In many procedures of statistical data analysis, the percentage distribution for a certain characteristic is used relative to another indicator. The distribution of the ordered characteristic values into intervals is made on the basis of certain criteria. For example, the Gini coefficient [8, p. 403] was originally defined as a statistical indicator of the degree of stratification of society in a given country or region in relation to any studied attribute. The value of the coefficient changes from 0 to 1. The closer this value is to zero, the more evenly the indicator is distributed.

Most often in modern economic calculations as in [9, pp. 67-70], the Gini index is used as an indicator of the discriminating ability of the classifier in solving the problem of classifying bank customers.

In [7,p. 447], an algorithm for extremal division of attribute values into gradations is presented. The principle that the algorithm implements is as follows: it is necessary to split the parameter values into a finite number of gradations so that the uncertainty (entropy) estimate when classifying using this feature is minimal (or close to minimal). At the beginning, the interval of values of the feature is divided into a sufficiently large number (τ) of gradations, and then by “gluing” adjacent gradations (thereby reducing the number of gradations τ) to achieve the minimization of entropy with respect to τ . Then, among the remaining gradations $\tau - 1$, two adjacent ones are “glued together” in order to minimize entropy, and so on. The disadvantage of this algorithm is the absence of a criterion for constructing the boundaries of intervals, the initial partitioning is performed in an intuitive way and the final partition is locally optimal.

In [10, p. 33], the main problems of methods of searching for logical patterns in data were identified and analyzed. A common problem for traditional methods is the “first step” problem (feature segmentation). The well-known algorithms for finding *if-then* rules make a mistake at the very beginning of their work, using heuristic assumptions during segmentation to limit further enumeration. In the author’s abstract of V. A.Dyuk’s doctoral dissertation [10, p. 33], the thesis is substantiated that the first step in the operation of an algorithm claiming a “high result” should be the smallest possible (taking into account the available computing power) partition of the initial features into intervals.

In [11], it is proposed to search for informative regularities through the generation of a family of predicates by an arbitrary feature.

Formulation of the problem

A sample of objects $E = \{S_1, \ldots, S_m\}$ is given that contains representatives of 2 disjoint classes K_1 and K_2. The description of objects is made using a set of features of different types $X(n) = (x_1, \ldots, x_n)$, of n diverse traits, ξ of which measured in interval scales, $n - \xi$ – measured in nominal.

To search for patterns by sets of latent features in the description of sample objects, you need to perform the following action:

- the sequence of values of the initial and combined quantitative characteristics, in the description of classified objects of the sample, split into intervals;
- calculate the stability of the partitioning into intervals;
- Presentation of the revealed patterns by the intervals of partitioning of the initial and combined features in the form of fuzzy inference rules.

Latent (clearly unmeasured) features for describing the objects of the sample can be obtained through the use of linear arithmetic operations on the original quantitative features, or in the form of generalized estimates of objects for subsets of features [12, pp. 1-10].

Consider a variant of constructing latent features using arithmetic operations. The latent feature $x^*$is calculated as:

$$x^* = x_i \odot x_j,$$

where$i \neq j, 1 \leq i, j \leq n$и$i, j \in I$, $\odot = \{*, /\}$.

New knowledge on a latent feature is revealed when comparing its properties with respect to the original features from which it is synthesized.

Division into intervals according to the criterion of dominance of class representatives. A method is proposed for calculating disjoint intervals of quantitative features, within the boundaries of which the values of certain classes dominate [13, pp. 35-40; 15, p. 72].

Ordered set of values of a quantitative characteristic x_j:

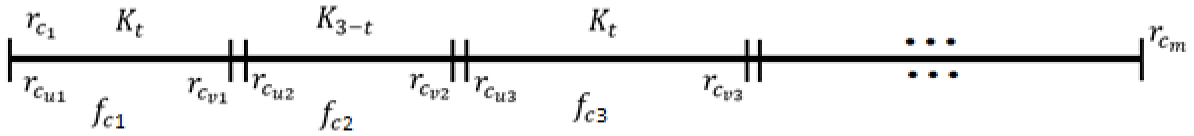
$$r_j = r_{j_1}, \ldots, r_{j_p}, \ldots, r_{j_m} \qquad (1)$$

According to the criterion defined below, an ordered sequence of the form (1) is divided into disjoint intervals$\left[r_{c_u}, r_{c_v}\right]^i$, $1 \leq u, u \leq v \leq m$, $i = \overline{1, \tau_c}$.

$d_t^i(u,v)$ - number of class K_t representatives, in interval $\left[r_{c_u}, r_{c_v}\right]^i$. For the recursive procedure for selecting values r_{c_u}, r_{c_v} the criterion is used:

$$\left|\frac{d_1^i(u,v)}{|K_1|}-\frac{d_2^i(u,v)}{|K_2|}\right|\to\max. \qquad (2)$$

Let's denote by $\eta_{1i}=\frac{d_1^i(u,v)}{|K_1|}$, $\eta_{2i}=\frac{d_2^i(u,v)}{|K_2|}$ results of optimal partitioning according to (3) for each interval $\left[r_{c_u},r_{c_v}\right]^i, i=\overline{1,\tau_c}$.



Pic. 1. Splitting into intervals of dominance of the values of the *c* feature

Dominance is quantitatively expressed through the $f_{ci}\in(0.5,1]$ class of $K_l, l=1,2$ membership function. The value of the membership function of the *c* attribute to the K_1 interval $\left[r_{c_u},r_{c_v}\right]^i$ is defined as:

$$f_{ci}=\frac{\eta_{1i}}{\eta_{1i}+\eta_{2i}}. \, (3)$$

An additional alternative for ranking by $K_t, t=1,2$ with an equal number of intervals is the stability of the feature values by the membership function, which is calculated as:

$$U(c)=\frac{1}{m}\sum_{\left\{\left[r_{c_u},r_{c_v}\right]^i\right\}}\begin{cases} f_{ci}(v-u+1), & f_{ci}>0.5,\\ (1-f_{ci})(v-u+1), & f_{ci}<0.5,\end{cases} \quad (4)$$

expresses the degree of homogeneity (non-displacement) of the values of the *c* attribute of objects within the boundaries of the intervals of dominance determined by (2). If (ideally) the values of the attribute of one class lie within the boundaries of the intervals, then it is $U(c)=1$.

Computational experiment

For the experiment, a sample of *"Echocardiogram"* data from [16] was taken, describing the condition of patients who had a heart attack. The sample consists of 108 objects, of which 74 belong to the class (the patient died within 1 year), 34 to K_2 (the patient survived or died after 1 year).

Each object is described by sets *X(10)=(x1,…,x10)*, containing 8 quantitative features (*x1=«survival»*, *x2=«wall-motion-score»*, *x3=«epss»*, *x4=«age-at-heart-attack»*, *x5=«lvdd»*, *x6*=«*wall-motion-index*», *x7*=«*fractional-shortening*», *x8*=«*mult*») and 2 nominal (*x9*=«*alive-at-1*», *x10*=«*pericardial-effusion*»).

Dividing into intervals of dominance of the values of quantitative features in the description of objects is a way to reveal hidden patterns. Table 1 shows the number of intervals of dominance of quantitative features and their boundaries for the sample.

TABLE 1. CHARACTERISTICS OF THE INTERVALS OF DOMINANCE OF THE VALUES OF THE INITIAL QUANTITATIVE CHARACTERISTICS

Attribute	**Boundaries of intervals [u, v] by (3)**	**The value of the membership function (4)**	**Stability of the trait by intervals (5)**
x1	[0.25, 10]	0.05	0.86
	[12, 57]	0.82	
x2	[7.5, 8]	0.13	0.69
	[9, 14.5]	0.68	
	[15, 39]	0.33	
x3	[11, 40]	0.36	0.67
	[0, 10.3]	0,69	
x4	[35, 64]	0.63	0.65
	[65, 86]	0.3	
x5	[2.32, 2.32]	1	0.68
	[3, 3]	0	
	[3.1, 4.55]	0.72	
	[4.56, 6.73]	0.37	
	[6.74, 6.74]	1	
x6	[1, 1.3]	0.71	0.7
	[1.31, 3]	0.31	
x7	[0.01, 0.24]	0.37	0.7
	[0.25, 0.61]	0.81	
x8	[0.28, 0.57]	0.28	0,68
	[0.59, 0.81]	0.6	
	[0.86, 0.93]	0.31	
	[0.93, 1]	0.87	

From Table 1, one can distinguish relatively strongly pronounced patterns in the belonging of objects to classes (in the particular case to K_1), for signs *x1* in the interval [12.57] with a membership function value of 0.82 and *x7* in the interval [0.25, 0.61] with a membership function value of 0.81 s stability of objects in the intervals of 0.86 and 0.7, respectively. If the boundaries of the intervals coincide (for example, as for *x2* - [5.5, 5.5]), then objects with such a value require additional checking for anomalies.

TABLE 2. INTERVALS OF DOMINANCE OF VALUES OF LATENT QUANTITATIVE FEATURES (FRAGMENT)

Combination of signs	**Boundaries of intervals [u, v] by (3)**	**The value of the membership function (4)**	**Stability of the trait by intervals (5)**
x1 / x4	[0, 0.13]	0.02	0.85
	[0.15, 1]	0.83	
x4 / x1	[1, 6.6]	0.83	0.87
	[7.7, 344]	0	
*x1 * x8*	[0.18, 7.14]	0.02	0.87
	[8.12, 52.16]	0.83	
*x1 * x5*	[1.21, 42.3]	0.03	0.88
	[52.6, 280.9]	0.83	
x1 / x5	[0.05, 4]	0.13	0.84
	[4.07, 16.38]	0.93	
x5/ x1	[0.06, 0.25]	0.94	0.82
	[0.252, 20.8]	0.14	
x2 / x7	[21.31, 56]	0.77	0.64
	[60.53, 3900]	0.38	
*x7*x8*	[0.01, 0.04]	0.28	0.64
	[0.05, 0.17]	0.68	
	[0.19, 0.35]	0.76	
	[0.38, 0.38]	0	
x7 / x2	[0, 0.02]	0.38	0.67
	[0.021, 0.05]	0.77	
x8 / x7	[3.49, 92.8]	0.38	0.56
	[0.94, 3.45]	0.67	

Table 2. A subset of combined (latent) features with high values according to the criterion of the type:

$$\frac{U(i)}{P(\text{i})} \to \max,$$

where $U(i)$ is the stability of objects in the intervals of dominance by the *i*attribute, $P(i)$ is the number of intervals when dividing the *i* attribute.

In table 2 there are examples of combined features with both *"improved"* property indicators and *"deteriorated"* ones. For example, *x1* in 2 intervals, had a stability of 0.86 and *x8* in 3 intervals, had a stability of 0.65 when their combination of the form x1 * x8, the latent feature is divided into 2 intervals of dominance with object stability 0.87. On the other hand, the combination of features *x7* with 2 intervals of dominance with a stability of 0.7 and *x8* with 4 intervals of dominance, with a stability of 0.68 with their combination of the form *x7* * *x8* is divided into 4 intervals of dominance with a stability of 0.64.

The patterns revealed during the experiment (see Tables 1-2) can be described using fuzzy inference formalisms [17, pp. 338-353; 18] and linguistic variables. For example, according to the tables 1 and 2, it is possible to compose a set of fuzzy inference rules about the belonging of an object to a class of the form

If A then B ('coefficient of confidence'),

where*' coefficient of confidence'* is the coefficient of confidence to the conclusion B if condition A.

Fragment of the list of rules for displaying table 1:

P_{11}: *If* $x1 \in \left[0.25,10\right]$ *then* $S \in K_1$ *(0.05);*

P12: If $x1 \in \left[12,\ 57\right]$ *then* $S \in K_1$*(0.82);*

P_{13}: *If* $x2 \in \left[7.5,\ 8\right]$ *then* $S \in K_1$*(0.13);*

P_{14}: *If* $x2 \in [9,\ 14.5]$ *then* $S \in K_1$*(0.68);*

P_{15}: *If* $x2 \in [15,\ 39]$ *then* $S \in K_1$*(0.33);*

P_{16}: *If* $x3 \in \left[11,\ 40\right]$ *then* $S \in K_1$ *(0.36);*

P_{17}: *If* $x3 \in [0, 10.3]$ *then* $S \in K_1$*(0.69);*

…

Fragment of the list of inference rules according to Table 2:

P21: If $x1/x4 \in \left[0,\ 0.13\right]$then (0.02);

P22: If $x1/x4 \in \left[0.15,\ 1\right]$then (0.83);

P23: If $x1*x8 \in \left[0.18,\ 7.14\right]$then $S \in K_1$ (0.02);

P24: If $x1*x8 \in \left[8.12,\ 52.16\right]$then $S \in K_1$ (0.83);

P25: If $x7/x2 \in \left[0,\ 0.02\right]$ then $S \in K_1$ (0.38);

P26: If $x7/x2 \in \left[0.021,\ 0.05\right]$ then $S \in K_1$(0.77);

You can use a measure of confidence of the form

$$МД\left[h:e1,e2\right]=МД\left[h:e1\right]+МД\left[h:e2\right]\left(1-МД\left[h:e1\right]\right), \qquad (6)$$

which is interpreted as “a measure of confidence in an event h with evidence $e1$and equal to the measure of confidence in an event h with evidence added to a mixed me $e2$ asure of confidence in an event h with evidence $e2$” [19]. This measure is used to assess the credibility of an event with an increase in the number of evidence. This can be demonstrated with the following example. If, in the process of inference, the sending of the rules P_{11} and P_{17} are executed (see the list above), then the measure of confidence in the event takes the form

$$МД\left[«S\in K_1»:x1\in\left[0.25,10\right],x3\in[0,10.3]\right]=МД\left[«S\in K_1»:x1\in\left[0.25,10\right]\right]$$

$$+МД[«S\in K_1»:x3\in[0,10.3]](1-МД[«S\in K_1»:x1\in\left[0.25,10\right]]).$$

Substituting the confidence coefficients from the rules, we get

$$МД\left[«S\in K_1»:x1\in\left[0.25,10\right],x3\in[0,10.3]\right]=0.05+0.69*\left(1-0.05\right)=0.706.$$

If there are more than two evidences in favor of an event, the measure of confidence is computed in a cascade manner.

To interpret the results of inference in natural language, it is possible to assign a linguistic variable and a membership function to each feature [12, pp. 1-10; 13, pp. 35-40]. As an example, let us take the sign *x1*, the value of which is the number of months that have passed since the patient has a heart attack. The linguistic variable for attribute *x1* can be defined as follows:

β_1 – “Number of months after a heart attack”; T_1 = {*“Big enough”, “Very small”*}; X_1 = [0.25, 57]; G_1 - β_1 –expanded set of new values of the linguistic variable {*“Very small”*, *“Small”*, *“Normal”*, *“Good”*, *“Large”*, *“Large enough”*} (see Table 6).

TABLE 6. THE SET OF VALUES OF THE LINGUISTIC VARIABLE β_1

Survival interval by x_1 (*survival*)	$\boldsymbol{T_1\cup G_1(T_1)}$
[0,25..1)	*Very small*
[1..3)	*Small*
[3..7,5)	*Normal*
[7,5..12)	*Good*
[12..24)	*Large*
[24..57]	*Large engough*

The tabular form of setting the values of the linguistic variable and thememembership functions $f_c(i)$ to the K_1 class presented in Table. 7.

TABLE 7. TABULAR REPRESENTATION OF THE LINGUISTIC VARIABLE

The value of the membership function $f_c(i)$	**Linguistic variable values**
[0..0,2)	*Practically absent*
[0,2..0,35)	*Small*
[0,35..0,5)	*Tangible*
[0,5..0,7)	*Sensitive*
[0,7..0,85)	*Strong*
[0,85..1]	*Very strong*

CONCLUSION

A method is proposed for revealing hidden patterns from databases based on dividing feature values into non-intersecting intervals. The revealed knowledge on making decisions about the belonging of the sampled objects to the classes is presented in the form of a set of fuzzy inference rules. A method for calculating the membership function in fuzzy rules is described. A decision-making mechanism is proposed by calculating the confidence measure of an object's belonging to a class according to fuzzy inference rules.

A methodology for mapping the values of the membership function to the values of a linguistic variable is proposed for the purpose of interpreting knowledge.

Decision-making on the classification of an arbitrary admissible object can be made if it does not have measured values for some of the features.

REFERENCES

1. Nikolaev A.B., Fominykh I.B. (2003). *Intelligent analysis and data processing*. Tutorial. - Moscow: MADI (STU), – p. 119.

2. Orlov A.I. (2012). *Measurement theory and methods of data analysis* // Modern sociology of modern Russia. Digestofarticles. Moscow: NRUHSE, – pp. 217-225

3. JuravlevY.I., GurevichI.B. (2000).*Pattern recognition and image analysis* // Artificial Intelligence: Models and Methods. T. 1. - Moscow: Radio and communication, –p. 310.

4. Juravlev Y.I. (1989)*On algebraic methods in problems of recognition and classification //* Recognition, classification, forecasting. Mathematical methods and their application. Moscow: Nauka,. Issue 1. – pp. 9-16.

5. JuravlevY.I. (1978). *On an algebraic approach to solving problems of recognition and classification //* Problems of Cybernetics. Mosciw: Science. – pp. 5-68

6. Krivenko M.P. (2016)*Significance criteria for selection of classification signs* // Informatics and its application, Vol 10, Issue 3, – pp. 32–40

7. Vapnik V.N. (1979). *Recovering dependencies from empirical data*. – Moscow: Nauka, –p. 447.

8. Riplay B.D. (2005).*Pattern Recognition and Neural networks* // Cambridge university press, – 403 p. (Riplay B.D. Pattern Recognition and Neural networks// Cambridge university press, 2005.-403 p.)

9. Shunina Y.S., Alekseeva V.A., Klyachkin V.N. (2015).*Performance criteria for classifiers*. - Bulletin of the Ulyanovsk State Technical University, No. 2 (70), – pp. 67–70.

10. DukeV. A. (2005). *Methodology of searching for logical patterns in the subject area with fuzzy systemology:* Author's Abstract of Doctoral Diss. Of Technical Sciences, St. Petersburg, – p. 33

11. VorontsovK.V. Mathematicalteachingmethodsbyprecedents (machinelearningtheory) Retrievedfrom: http://www.ccas.ru/voron

12. Ignatyev N.A., MadrakhimovSh.F.,Saidov D.Y . (2017). *Stability of object classes and selection of the latent features* // International journal of engineering technology and sciences, Malaysia, Vol. 7, – pp. 1–10. (Ignatyev N.A., MadrakhimovSh.F.,Saidov D.Y.. Stability of object classes and selection of the latent features // International journal of engineering technology and sciences, 2017, Malaysia, Vol. 7, pp. 1-10.)

13. Ignatiev N. A., Saidov D. Y. (2014)*Computing the complexity of effective algorithms for choosing the optimal boundaries of intervals* // Problems of Informatics and Energy, - Tashkent, no. 6, – pp. 35–40.

14. Ignatiev N.A. (2011).*Calculation of generalized indicators and data mining* // Automation and tele-mechanics, No. 5, – pp. 183–190.

15. Ignatiev N.A. (2014) *Generalized estimates and local metrics of objects in data mining*. Monograph. - Tashkent: Publishing House "University", – p. 72.

16. Retrieved from online source: http://archive.ics.uci.edu/ml/machine-learning-databases/echocardiogram

17. Zadeh, L. (1965*). Fuzzy sets. Information and Control*, 8 (3): pp. 338 – 353. (Zadeh, L. (1965). Fuzzy sets. Information and Control, 8(3):338 – 353.)

18. Shtovba S.D. Introduction to the theory of fuzzy sets and fuzzy logic / S. D. Shtovba. Retrieved from: http: // matlab.exponenta.ru/fuzzy-logic/bookl/index.php.

19. Buchanan, B. E. (1984). Rule based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley. (Buchanan, B. E. Rule based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley, 1984.)