Nonlinear transformation of signs and the search for patterns in the data of patients with chronic lymphocytic leukemia

Nikolay Ignatyev
National University of Uzbekistan
Tashkent, Uzbekistan
n_ignatev@rambler.ru

Ekaterina Zguralskaya
Ulyanovsk State Technical University
Ulyanovsk, Russia
iatu@inbox.ru

Maria Markovtseva
Ulyanovsk State University
Ulyanovsk, Russia
mmark7@yandex.ru

***Abstract*—The paper considers the search for logical patterns by descriptions of objects in rectifying space. The rules of hierarchical agglomerative grouping are applied for the synthesis of latent features of this space. A pair of characteristics for combining into a group is chosen according to the maximum of criterion for characteristic values decomposition into disjoint intervals. The analytical form of arithmetic expressions for calculating latent features used to detect hidden patterns in the data of patients with chronic lymphocytic leukemia (CLL) is given.**

Keywords—latent features, the search for logical patterns, hierarchical agglomerative grouping, chronic lymphocytic leukemia

I. INTRODUCTION

The choice of space to describe objects through non-linear transformations of features serves as a tool to detect latent patterns in data. When using such transformations, the structure of relations of objects in a new (latent) feature space changes. Quantitative estimators of the structure can be expressed in terms of compactness of objects of a class and the sample as a whole.

Several methods are proposed for evaluating compactness [1, 2]. In [2], connection between dimensionality of the feature space and the generalizing ability of recognition algorithms according to the nearest neighbor rule is shown through the measures of compactness of the objects of a class and the sample as a whole. The estimator of class compactness by a quantitative feature in [3] was calculated as an extremum of the criterion for decomposing the feature values into disjoint intervals.

For the data analysis, the numerical axis is considered as a universal scale with relations. The universal scale was used to study the relations between the objects of classes according to the results of the non-linear representation of their descriptions by the defined sets of features on the numerical axis [4]. Since the composition of each set is initially unknown, it was proposed to use the criterion of decomposing the feature values into disjoint intervals for its search.

Feature sets in [4] for further synthesis of latent features on their basis were sequentially formed according to the rules of hierarchical agglomerative grouping. The number of latent (groups) of features was determined upon the grouping results. A quality of informative orderliness of latent features makes them desirable for the analysis. This orderliness quality provides certain advantages for finding latent patterns in data.

The implementation of nonlinear transformations of features is considered as one of the stages of reducing the dimension of space [2]. When deciding whether to remove latent signs from the set, the property of their ordering is used. One of the goals of creating space through the removal of features is to solve the problem of retraining recognition algorithms. As a criterion for detecting the beginning of retraining in [2], a measure of compactness was used, calculated by solving the problem of minimal coverage of the sample with reference objects.

To detect hidden patterns in the data, you can use a combination of linear and nonlinear methods to reduce the dimension of the attribute space. The software implementation of many linear methods is presented in the Python language library [5]. As a linear display to the numerical axis, we consider the calculation of generalized object estimates [6] based on sets of heterogeneous attributes. Presentation of the results of data analysis on the numerical axis is a means for searching and recording logical patterns in the form of half-planes.

The issues of building information models in medicine are most often considered from the point of large or poorly structured systems [7]. Such systems require to take into account how various factors are connected and how they influence the processes in the body. The study explores the search for patterns for patients with CLL [8]. Latent features, synthesized according to the rules of a hierarchical agglomerative grouping, are considered to be factors affecting the duration of the patients actual survival. Here the results of the search for latent patterns in the data of CLL patients by latent features and the analytical form of arithmetic expressions (formulas) to calculate their values are presented. There are practically no publications describing other methods of forming an analytical representation (formulas) for calculating the values of latent features based on nonlinear transformations.

II. PROBLEM STATEMENT AND SOLUTION METHOD

Let a set of objects $E_0=\{S_1, \ldots, S_m\}$ be given containing representatives of two disjoint classes K_1 and K_2. Objects are described using a set of n quantitative features $X(n)=(x_1,\ldots,x_n)$. It is considered that the operator A is given on E_0 to transform the descriptions of objects from $X(n)$ to $Y(k)$, $k<n$.

It is required to determine:

- the number of latent features in $Y(k)$;
- analytical form (equations) of arithmetic expressions to calculate the latent feature $y_i \in Y(k)$, $i=1, \ldots, k$.

The analytical form of arithmetic expressions for calculating latent features is formed on the basis of the algorithm from [2]. The set of feature numbers in the description of the objects E_0 will be identified as $I = \{1,\ldots,n\}$. The rules of hierarchical agglomerative grouping

are applied to calculate the values of latent features. Latent features obtained at the *p*-th step of the grouping are denoted as x_j^p, $j \in I$, $p \geq 0$. For $p=0$, $|I|=n$. We will divide the ordered set of x_j^p feature values of objects from E_0 into two intervals $[c_1^{jp}, c_2^{jp}]$ and $(c_2^{jp}, c_3^{jp}]$, each of which is considered as a nominal feature gradation.

Lets u_i^1, u_i^2 be the number of values of the feature x_j^p, $j \in I$ of class K_i, $i = 1,2$, respectively in the intervals $[c_1^{jp}, c_2^{jp}]$ and $(c_2^{jp}, c_3^{jp}]$, $|K_i|>1$, v is the serial number of the element in an ascending order

$$r_{j_1}, \dots, r_{j_v}, \dots, r_{j_m} \quad (1)$$

of values x_j^p of objects from E_0, which defines the boundaries of the intervals as $c_1^{jp} = r_{j_1}, c_2^{jp} = r_{j_v}, c_3^{jp} = r_{j_m}$.

Criterion

$$\left(\frac{\sum_{i=1}^{2} u_i^1(u_i^1-1) + u_i^2(u_i^2-1)}{\sum_{i=1}^{2}|K_i|(|K_i|-1)}\right)$$

$$\left(\frac{\sum_{d=1}^{2}\sum_{i=1}^{2} u_i^d\left(|K_{3-i}| - u_{3-i}^d\right)}{2|K_1||K_2|}\right) \to \max_{c_1^{jp}<c_2^{jp}<c_3^{jp}} \quad (2)$$

allows to calculate the optimal value of the boundary c_2^{jp} for the intervals $[c_1^{jp}, c_2^{jp}]$ and $(c_2^{jp}, c_3^{jp}]$.

The extremum of criterion (2) is used as the weight w_j^p ($0 \leq w_j^p \leq 1$) of the feature x_j^p. When $w_j^p = 1$, values of the feature x_j^p for objects from classes K_1 and K_2 are mutually disjoint.

The value of the combination b_{rij}^p by a pair of features (x_i^p, x_j^p), $0 \leq p < n$, $i, j \in I$, $i \neq j$ of the object $S_r = \{a_{ru}^p\}_{u \in I}$, $S_r \in E_0$ is calculated as

$$b_{rij}^p = \eta_{ij}(t_i w_i^p (a_{ri}^p - c_2^{ip})/(c_3^{ip} - c_1^{ip}) + t_j w_j^p (a_{rj}^p - -c_2^{jp})/(c_3^{jp} - c_1^{jp})) + (1-\eta_{ij}) t_{ij} w_{ij}^p \left(\lambda_{rij}^p - c_2^{ijp}\right)/(c_3^{ijp} - -c_1^{ijp}), i, j \in I, t_{ij}, t_i, t_j \in \{-1,1\}, \eta_{ij} \in [0;1] \quad (3)$$

where w_i^p, w_j^p, w_{ij}^p are the weights of the features determined by (2), respectively from the set of values x_i^p, x_j^p and their product $\lambda_{rij}^p = \left(\frac{a_{ri}^p - c_2^{ip}}{c_3^{ip} - c_1^{ip}}\right)\left(\frac{a_{rj}^p - c_2^{jp}}{c_3^{jp} - c_1^{jp}}\right)$ on E_0; the values t_{ij}, t_i, $t_j \in \{-1,1\}$, $\eta_{ij} \in [0;1]$ are selected by the extremum of the functional

$$\varphi(p,i,j) = \frac{\min_{S_r \in K_1} b_{rij}^p - \max_{S_r \in K_2} b_{rij}^p}{\max_{S_r \in E_0} b_{rij}^p - \min_{S_r \in E_0} b_{rij}^p} = \max_{t_{ij}, t_i, t_j \in \{-1,1\}, \eta_{ij} \in [0,1]} \quad (4)$$

The extremum of functional (4) is interpreted as the space between objects of classes K_1 and $K2$ by the set of values b_{rij}^p for a pair of features (x_i^p, x_j^p), $0 \leq p < n$, $i, j \in I$, $i \neq j$.

In (3), valuation of features along the boundaries of intervals calculated by (2) is applied. Due to the valuation of features, the values b_{rij}^p are scale independent.

Let $\{z_{ij}^p\}_{i,j \in I}$, $p \geq 0$ denote the square matrix of size $(n-p)(n-p)$, whose element z_{ij}^p value at $p = 0$ is defined as

$$z_{ij}^p = \begin{cases} w_i^p, i = j, \\ value\ (2)\ by\ \{b_{rij}^p\}_{r=1}^m, i \neq j, \end{cases} \quad (5)$$

by Γ_η, $\eta>0$ being the subset of feature numbers from X (n).

A step-by-step implementation of the hierarchical agglomerative grouping algorithm will be as follows:

Step 1: $p=0$, $\lambda c=0$, $\eta = 1$. Execute $\Gamma_\eta = \{\eta\}$, $Margin_\eta = -2$, $\eta = \eta + 1$ while $\eta \leq n$;

Step 2: Calculate the values of the elements of the matrix $\{z_{ij}^p\}_{i,j \in I}$ by (5);

Step 3: Select $\Phi = \{z_{uv}^p | z_{uv}^p \geq \max(w_u^p, w_v^p)$ and $u \neq v$, $u, v \in I\}$. If $\Phi = \varnothing$, then go 9;

Step 4: Calculate $\lambda n = \max_{z_{uv}^p \in \Phi} z_{uv}^p$. Select $\Delta = \{(s,t), s, t \in I \mid z_{st}^p = \lambda n$ and $s<t\}$. Define a pair $\{i, j\}$, $i<j$ as

$$\{i,j\} = \begin{cases} \Delta, |\Delta| = 1, \\ \{s,t\}, (s,t) \in \Delta\ and\ \varphi(p,s,t) > \max_{(u,v) \in \Delta \setminus (s,t)} \varphi(p,u,v); \end{cases}$$

Step 5: If $\lambda n > \lambda c$ or $\lambda n = \lambda c$ and $Margin_i < \varphi(p, i, j)$, then $\Gamma_i = \Gamma_i \cup \Gamma_j$, $\Gamma_j = \varnothing$, $Margin_i = \varphi(p, i, j)$, go to 7;

Step 6: Display feature numbers from Γ_i, $\Gamma_i = \varnothing$, $I = I \setminus \{i\}$, go to 3;

Step 7: $p = p+1$, $I = I \setminus \max(i,j)$, $k = \min(i,j)$, $\lambda c = \lambda n$. Replace the values of the features in the description of the object $S_r = \{a_{ru}^{p-1}\}_{u \in I}$, $r = 1, \dots, m$ with

$$a_{ru}^p = \begin{cases} a_{ru}^{p-1}, u \in I \setminus \{k\}, \\ b_{rij}^p, u = k; \end{cases}$$

Step 8: For each pair (u, v), $u, v \in I$ determine the value

$$z_{uv}^p = \begin{cases} z_{uv}^{p-1}, u \in I\{k\}, v \in I, \\ value\ (9) on\ \{a_{rv}^p\}_{r=1}^m, u = k, v \in I. \end{cases}$$

If $n-p > 1$, then go to 3;

Step 9: The end.

When implementing the algorithm described above, in order to form the analytical representation of arithmetic expressions the parameter values were used.

For example, here is the value $\eta \in \{0, 1\}$ in (3). Then, in the analytical representation of arithmetic expressions, a linear or nonlinear parts are written while calculating the latent feature by (3).

If two or more nominal features are used to describe admissible objects, then latent quantitative features can be obtained based on combinations of their gradation [6]. The values of such latent features are calculated as generalized estimator of objects. The purpose of using generalized estimator is to abandon the point estimator of objects obtained on the basis of expert subjective criteria.

Let us consider one of the methods for calculating generalized estimates of objects using a set of latent attributes $Y(k)=(y_1,\dots,y_k)$, $k < n$. Let $V=\{v_i\}_{i \in \{1,\dots,k\}}$ and $F=\{[c_{i1},c_{i2}], (c_{i2},c_{i3}]\}_{i \in \{1,\dots,k\}}$, respectively, the set of weights

and the set of boundaries of the intervals calculated from (2) on *Y(k)*. We will form a new feature space for describing objects in the nominal scale of measurements by gradations from *{1, 2}*. Depending on the values of latent features of the objects belonging to one of two disjoint intervals from *F*, a gradation is written in the new space 1 or 2. The contribution of the attribute $y_d \in Y(k)$ to the generalized estimate by the gradation $j \in \{1, 2\}$ and the weight $v_d \in V$ is defined as

$$\mu_d(j) = v_d \left(\frac{\alpha_{dj}^1}{|K_1|} - \frac{\alpha_{dj}^2}{|K_2|} \right), \quad (6)$$

where $\alpha_{dj}^1, \alpha_{dj}^2$ is the number of gradation values j of the attribute y_d, respectively, in classes K_1 and K_2. The generalized estimate of the object $S_r \in E_0$ according to its description in the nominal measurement scale $S_r=(a_{r1},\ldots,a_{rk})$ and contributions (6) is calculated as

$$Z(S_r) = \sum_{i=1}^{k} \mu_i (a_{ri}). \quad (7)$$

III. Computation Experiment

For the study we used data of 123 patients with CLL A-C by Binet [9, 10] aged 47 to 82 years with known values of overall survival rate, obtained at the Hematology Department of the Ulyanovsk regional clinical hospital. At the time of diagnosis of CLL, the age of patients was registered, the Charlson comorbidity index was calculated, standard biochemical parameters were measured: alanine aminotransferase (ALT), aspartate aminotransferase (AST), total bilirubin, indirect bilirubin, glucose, urea, and the glomerular filtration rate (GFR) was calculated using the MDRD formula: GFR=*186*{[serum creatinine (plasma) + 88.4] -1.154}*age-0.0203*. Additionally, the number of chemotherapy courses performed and the actual survival rate in months were recorded. Patients with HIV infection and other oncological conditions were excluded from the study.

By gender, two data samples were generated. The sample of these male patients consisted of 60 objects (64.6 ± 9.0 years old), female of 63 objects (67.0 ± 8.4 years old). The objects of each sample were divided into two disjoint classes K_1 (actual survival rate is less than the prognosed overall survival) and K_2 (actual survival rate is greater than or equal to the prognosed overall survival).

The strongest patterns were obtained on the male patients data. Classes K_1 and K_2 were represented by 36 and 24 objects respectively. Based on the results of computation experiments on two sets (identified as the first and second) of the initial features, nonlinear combinations were obtained to completely separate the objects of two classes. In the second set, there was a feature of comorbidity index missing. The complete separability of classes is confirmed by the value of criterion (2), equal to 1.

Here below the sequence of formation of the first two latent features for each of the two sets of initial features is given.

The sequence of formation of the first latent feature from the first set:

x0=0.1428*(comorbidity index – 4.0);

x1=0.0175*(GFR – 76.0);

x2=1.9463*(*x0***x1* + 0.01);

y1=1.2572*((0.1203**x0* – 0.2442**x1* – 0.5339**x2*) – 0.0142).

The sequence of formation of the first latent feature from the second set:

x0=0.0243*(age – 63.0);

x1=0.0175*(GFR – 76.0);

x2=2.3798*(*x0***x1* + 0.010270);

y1=2.6311*((–0.0985**x0* + 0.2442**x1* – 0.1790**x2*) + 0.0332);

x0=1.0699**y1*;

x1=0.0083*(creatinine – 84.0);

y2=1.1537*((–0.9346**x0* – 0.2717**x1*) + 0.0460).

The logical regularity in the form of half-planes obtained from the second set is written as *φ[y2< – 0.003]=β, β∈{true, false}*. When *φ [y2 <- 0.003] = true*, the patient will live less than the estimated period of overall survival. The threshold value for the half-plane was determined by the result (2) of decomposing the descriptions of the sample objects according to the feature y2 into two intervals $[c_1; c_2]$ (c_2; c_3] as $(c_2 + b)/2$, where b is the value closest to c_2 from (1) and $b > c_2$.

There is a functional dependence of GFR on age and creatinine value depending on gender. Determination of the points needed to calculate the comorbidity index demands additional effort from a user to collect data from patients. For practical reasons, the use of the second option (without the comorbidity index) for calculating the latent feature looks preferable for the prognosis. To prognoses the patient's survival potential, it will be enough for the user to set the values of measurable indicators - age and creatinine.

The values of the first two latent features in [4] are recommended to be used for visual representation of objects on the plane. Recommendations are based on the analysis of the results of breaking into disjoint groups of descriptions of class objects in R^2. Another argument when choosing the first two latent features for visualization is their values according to criterion (2). In the table 1 demonstrates the existence of a relationship between the number of the latent feature and its value according to (2) for female patients.

TABLE I. The Relationship Between the Serial Number of the Latent Feature and its Value According to (2)

Number of the latent feature	The procedure for combining the initial features in latent	Criterion value (2)
1	(((GFR, number of chemotherapy courses), glucose), AST)	0.7656
2	(((comorbidity index, total bilirubin), creatinine), ALT)	0.4890
3	(age, urea)	0.4101
4	indirect bilirubin	0.3016

Ordering by (2) can be used to select informative sets of latent features from $Y(4)=(y_1, y_2, y_3, y_4)$. The number of options during selection is minimized due to the sequential removal of features with the smallest value (2). According to the results of calculating the generalized estimates of objects according to (7) on the sets $Y(4)$, $Y(4)\backslash\{y_4\}$, $Y(4)\backslash\{y_3, y_4\}$, the number of errors was 4, which corresponds to the accuracy on the training 93.65%. When determining the threshold between classes, we used the partition of the

values of generalized estimates into two intervals according to (2).A visual representation of objects (female patients) by the first two latent features is shown in “Fig. 1”.

Anomalous deviations in class objects can be visually detected by their latent attribute values [11]. The probability of such deviations is indicated by values (2) less than 1 for all latent features from the table 1. In the interpretation of latent features obtained by the algorithmic method, the concept of “the set of permissible values” is not used.

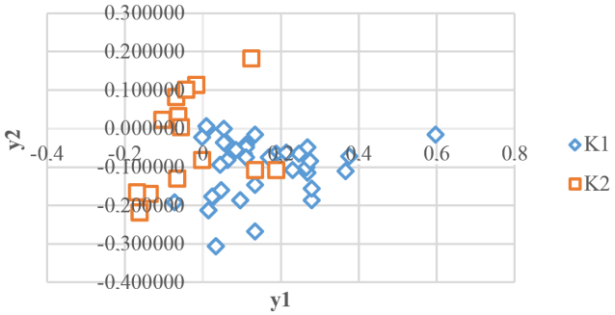

Fig. 1. Visual representation of female patients.

IV. Conclusion

The search for hidden patterns in the data of CLL patients is described. The algorithm to generate the analytical representation of arithmetic expressions for calculating the values of latent features is given. The prognosis of the deviation of the actual survival rate of male patients towards either decreasing or increasing from overall survival is determined by a logical regularity in the form of half-planes. The forecast of deviation of the actual terms of survival of patients in the direction of decreasing or increasing from the terms of overall survival is determined by logical patterns in the form of half-planes. The found patterns can be recommended for use in specialized medical institutions.

References

[1] N.G. Zagoruyko, O.A. Kutnenko, A.O. Zyryanov and D.A. Levanov, “Pattern recognition training without retraining,” Machine Learning and Data Analysis, vol. 1, no. 7, pp. 891-901, 2014.

[2] N.A. Ignatyev, “Structure Choice for Relations between Objects in Metric Classification Algorithms,” Pattern Recognition and Image Analysis, vol. 28, pp. 590-597, 2018.

[3] E.N. Zguralskaya, “Analysis of the structure of the relationship between the descriptions of objects of classes and evaluation of their compactness,” CEUR Workshop Proceedings, vol. 2416, pp. 283-289, 2019.

[4] D.Y. Saidov, “Data visualization and its proof by compactness criterion of objects of classes,” International Journal of Intelligent Systems and Applications, vol. 9, no. 8. pp. 51-58, 2017.

[5] M. Andreas and G. Sarah, “Introduction to machine learning using Python,” Data Professionals Guide, SPb: Alpha Book, 2017.

[6] N.A. Ignat’ev, “Computing generalized parameters and data mining,” Automation and Remote Control, vol. 72, no. 5, pp. 1068-1074, 2011.

[7] K. Cresswell, M. Callaghan, S. Khan, H. Mozaffar and A. Sheikh, “Investigating the use of data-driven artificial intelligence in computerised decision support systems for health and social care: A systematic review,” Health Informatics Journal, vol. 26, no. 3. pp. 2138-2147, 2020.

[8] A.K. Nikitina and N.O. Sarajeva, “Treatment efficacy and survival of patients with chronic lymphatic leukemia depending on renal function,” Trans-Baikal Med. Bulletin, no. 4, pp. 122-127, 2014.

[9] J.L. Binet, A. Auquier and G. Dighierol, “A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis,” Cancer, vol. 48, pp. 198-206, 1981.

[10] K.R. Rai, A. Sawitsky and E.P. Cronkite, “Clinical staging of chronic lymphocytic leukemia,” Blood, no. 46, pp. 219-234, 1975.

[11] D.V. Semenova and E.E. Goldenok, “The method of generalized assessments on the example of diagnosing the severity of the course of acute pancreatitis,” Materials of the XVIII International Conf. ITMM, Tomsk, vol. 1, pp. 152-157, 2019.