

ANALYSIS OF RELATIONSHIPS BETWEEN TRAINING SAMPLE OBJECTS ON MANIFOLDS OF LATENT FEATURE SETS

Akbarova Azizaxon Axmat qizi¹, Akbarov Baxriddin Xusniddin o'g'li²

Namangan state technical university, Namangan, Uzbekistan,

University of exact and social sciences, Tashkent, Uzbekistan

Abstract. This paper proposes an approach to pattern recognition based on latent feature sets formed through hierarchical agglomerative clustering using additive and multiplicative principles. The compactness of the resulting feature sets is evaluated via a connectivity ratio based on a hypersphere system. A computational experiment on genomic DNA data demonstrates that the multiplicative principle outperforms the additive one, achieving a recognition accuracy of 96.15% with only 5 latent features, confirming its effectiveness for solving dimensionality reduction problems in data mining tasks.

Annotatsiya. Ushbu maqolada additiv va multiplikativ tamoyillar asosida ierarxik aglomerativ guruhlash orqali shakllantiriladigan latent alomatlar to'plamlari asosida obrazni anglash yondashuvi taklif etilgan. Natijada olingan latent alomatlar to'plamlarining kompaktligi gipersharlar tizimiga asoslangan ulanish nisbati yordamida baholanadi. Inson genomi DNK ma'lumotlari bo'yicha o'tkazilgan hisoblash eksperimentida multiplikativ tamoyil additiv tamoyilga nisbatan yuqori natija berishini ko'rsatdi, ya'ni atigi 5 ta latent alomat bilan 96,15% anglash aniqligi ta'minlandi, bu esa ma'lumotlarni intellektual tahlil qilishda o'lcham la'nati muammosini hal etishda uning samaradorligini tasdiqlaydi.

Аннотация. В статье предлагается подход к распознаванию образов на основе наборов латентных признаков, формируемых методом иерархической агломеративной группировки по аддитивному и мультипликативному принципам. Компактность полученных наборов признаков оценивается с помощью отношения связанности на основе системы гиперсфер. Вычислительный эксперимент на геномных данных ДНК показал, что мультипликативный принцип превосходит аддитивный, обеспечивая точность распознавания 96,15% при всего лишь 5 латентных признаках, что подтверждает его эффективность для решения проблемы проклятия размерности в задачах интеллектуального анализа данных.

Keywords: latent features, dimensionality reduction, hierarchical agglomerative clustering, pattern recognition, compactness measure.

Kalit so'zlar: latent alomatlar, fazo o'lchamini kamaytirish, ierarxik aglomerativ guruhlash, obrazlarni anglash, kompaktlik o'lchovi.

Ключевые слова: латентные признаки, снижение размерности, иерархическая агломеративная группировка, распознавание образов, мера компактности.

The formation of latent feature sets is considered as one of the approaches to solving the curse of dimensionality problem in data mining tasks. For the implementation of linear and non-linear dimensionality reduction methods, there are a number of conditions that limit their application. These conditions include:

- orientation only toward quantitative features;
- sensitivity to the curse of dimensionality when computing the distance measure between objects;
- non-linear growth in computational complexity.

The source of the diversity of latent feature sets is the multitude of ways to group the original features, on the basis of which they are formed. When studying these sets, an analysis of changes in the proximity relationship structure of sample objects across different metrics was used [1]. To calculate the compactness measure of the sample, the connectivity ratio of objects was applied via a system of hyperspheres, at the intersections of which the boundary objects of the classes are located.

Problem Statement

The recognition problem is considered in its standard formulation. It is assumed that a set of objects $E_0 = \{S_1, \dots, S_m\}$, is given, divided into two non-overlapping classes K_1, K_2 . The objects are described using n heterogeneous features $X(n) = (x_1, \dots, x_n)$, of which ξ are measured on interval scales, and $n - \xi$ – on a nominal scale.

It is assumed that two sets of algorithms are defined on $X(n)$, differing in the principles of forming non-overlapping feature groups: Ω_1 – additive; Ω_2 – multiplicative.

Algorithms $A, B \in \Omega_1 \cup \Omega_2$ are used to form non-overlapping feature groups $\Delta(t(A)) = \{G_1, \dots, G_{t(A)}\}$, $\Delta(t(B)) = \{G_1, \dots, G_{t(B)}\}$ when selecting the description of objects E_0 by sets of latent features $Y(t(A)) = (y_1, \dots, y_{t(A)})$, $Y(t(B)) = (y_1, \dots, y_{t(B)})$. For each group $G_i \in \Delta(t(A))$ ($G_i \in \Delta(t(B))$), a latent feature $y_i \in Y(t(A))$ ($y_i \in Y(t(B))$) is synthesized.

The conditions for selecting subsets of boundary objects of the classes $\Theta(A, \rho)$ and $\Theta(B, \rho)$ — are defined with respect to the metric $\rho(x, y)$ on the sets $Y(t(A))$ and $Y(t(B))$, respectively.

Algorithm for Hierarchical Agglomerative Feature Clustering Based on the Additive Principle

The hierarchical agglomerative clustering algorithm [2] implements a greedy strategy for forming sets of latent features. The number of groups and their composition are determined algorithmically from $X(n) \setminus F$ based on the principle of dynamic programming.

Let $T (F \subset T)$ – be the set of features whose assignment to non-overlapping groups has been established. When $|T| < n$ the formation of the next group G according to the rules of hierarchical agglomerative clustering proceeds as follows:

- selection of $x_c \in X(n) \setminus T$ with the maximum weight value as the first representative of group G , $T = TU\{x_c\}$;
- inclusion of feature $x_p \in X(n) \setminus T$ into group $G = GU\{x_p\}$ under the condition

$$\frac{\gamma_i}{\theta_i} > \frac{\gamma_{i+1}}{\theta_{i+1}} \text{ и } \frac{\gamma_i}{\theta_i} - \frac{\gamma_{i+1}}{\theta_{i+1}} = \max_{x_p \in X(n) \setminus T},$$

where $\gamma_i(\theta_i)$ – is the intra-class similarity (inter-class difference) based on the values of the generalized scores on G , $i = |G|$, $|GU\{x_p\}| = i + 1$.

Multiplicative Principle for Computing Generalized Object Scores

When computing generalized object scores using the multiplicative principle, the weight values of features and their pairwise combinations are used. For data (latent and original) on interval measurement scales, it is recommended to use the interval method [3], as is done when implementing the rules of the hierarchical agglomerative feature clustering algorithm.

As the boundaries of two non-overlapping intervals $[\pi_1; \pi_2]$, $(\pi_2; \pi_3]$, the values $\pi_1 = r_1$, $\pi_2 = r_j$, $1 < j < m$, $\pi_3 = r_m$ are used. The intervals $[\pi_1; \pi_2]$ and $(\pi_2; \pi_3]$ are identified as the first and second, respectively. The feature weight of class objects is computed as the maximum of the product of intra-class similarity and inter-class difference according to the criterion:

$$\left(\frac{\sum_{d=1}^2 \sum_{i=1}^2 (u_i^d - 1) u_i^d}{\sum_{i=1}^2 |K_i| (|K_i| - 1)} \right) \left(\frac{\sum_{d=1}^2 \sum_{i=1}^2 u_i^d (|K_{3-i}| - u_{3-i}^d)}{2 |K_1| |K_2|} \right) \rightarrow \max_{\pi_1 < \pi_2 < \pi_3}, \quad (1)$$

where $u_i^d (u_{3-i}^d)$ – is the number of values of feature $x_c \in X(n)$ belonging to objects of class K_i (K_{3-i}) in the d -th interval. The set of admissible values of criterion (1) belongs to $(0; 1]$ and is used to evaluate the compactness of class objects on the number line. The boundary (threshold) between classes K_1 and K_2 for a quantitative feature $x_a \in X(n)$ is defined as:

$$\Gamma_a = \frac{\pi_2 + b}{2}, \quad (2)$$

where b is the value from the interval $(\pi_2; \pi_3]$ closest to π_2 , computed according to (1).

The replacement of gradations of nominal features with values of the class membership function is proposed for use in synthesizing latent features according to the rules of hierarchical agglomerative clustering. The algorithm and rules for clustering quantitative features are described in [3]. The process of pairwise merging of features to form a set of non-overlapping groups $\mathfrak{G} = \{G_1, \dots, G_t\}, t \geq 1$ is

based on the use of the extreme values of criterion (1) and the corresponding boundaries of the two non-overlapping intervals.

Let $Y(n-q)$ denote the set of latent and original features at the q -th step of hierarchical clustering, $Y(n)=X(n)$ when $q=0$. The verification of the condition for including feature $y_p \in Y(n-q)$ into group $G \in \mathcal{G}$ is preceded by normalization of its values using the interval boundaries $[\pi_{1p}; \pi_{2p}][\pi_{2p}; \pi_{3p}]$ according to the formula:

$$a_p = (y_p - \pi_{2p}) / (\pi_{3p} - \pi_{1p}). \quad (3)$$

For a quantitative feature, the interval boundary values $[\pi_{1p}; \pi_{2p}][\pi_{2p}; \pi_{3p}]$ are determined by criterion (1). After replacing the gradations of a nominal feature with membership function values, the selection of boundaries π_{1p}, π_{3p} is analogous to (1).

The synthesis of latent feature values $y \in Y(n-q-1)$ with normalization by (3) for a pair $(y_i, y_j) \in Y(n-q)$ is performed according to the formula:

$$y = \eta_{ij} (t_i w_i a_i + t_j w_j a_j) + (1 - \eta_{ij}) t_{ij} w_{ij} (a_i a_j - \pi_{2ij}) / (\pi_{3ij} - \pi_{1ij}), \quad t_{ij} \in \{-1, 1\}, \quad \eta_{ij} \in [0; 1], \quad (4)$$

where w_i, w_j, w_{ij} — are the feature weights and boundary values $[\pi_{1ij}; \pi_{2ij}][\pi_{2ij}; \pi_{3ij}]$, determined by (1).

When implementing the rules of hierarchical agglomerative clustering, the principle of dynamic programming is applied. Group G and the latent feature y obtained from it via (4) are considered formed if there exists no pair (y, y_i) , $y_i \in Y(n-q)$ upon merging which into $y^* \in Y(n-q-1)$, the feature weight by (1) satisfies $w(y^*) \geq w(y)$.

Computational Experiment

Data on the 3rd chromosome for 2504 objects [4] are analyzed, containing information on single mutations in certain coding gene regions of the DNA molecule. For DNA classification by genotype ELP6-201, data from the mediator RNA molecule were used as the target feature. The experiment considers a variant of partitioning the objects into two classes with one and two mutations.

When forming the sample for machine learning, only unique objects were used. Out of 52 objects, 27 had one mutation and 25 had two mutations.

The synthesis of two sets of latent features L_1 and L_2 from 38 original features for ELP6-201 was carried out according to the rules of hierarchical agglomerative clustering algorithms using the additive and multiplicative principles, respectively. The size of the set formed by the additive principle is $|L_1|=9$, and by the multiplicative principle — $|L_2|=5$. The results of computing the compactness measure (1) and the recognition accuracy by boundary (2) for each latent feature are presented in Tables 1 and 2.

Table 1**Compactness measure and recognition accuracy values for set L_1**

Latent feature No.	Values	
	compactness (1)	accuracy in % by (2)
1	0,3197	69,23
2	0,3216	71,15
3	0,2848	65,38
4	0,0743	3,85
5	0,0743	3,85
6	0,0743	3,85
7	0,0743	3,85
8	0,1419	50,00
9	0,1419	50,00

Table 2**Compactness measure and recognition accuracy values for set L_2**

Latent feature No.	Values	
	compactness (1)	accuracy in % by (2)
1	0,8550	96,15
2	0,3200	69,23
3	0,2286	50,00
4	0,0744	3,85
5	0,0744	3,85

Conclusion

The study demonstrates that the multiplicative principle of hierarchical agglomerative feature clustering produces significantly more compact and accurate latent feature sets compared to the additive principle, as evidenced by the first latent feature of L_2 achieving a compactness measure of 0.855 and a recognition accuracy of 96.15%. The proposed approach of analyzing proximity relationships between training sample objects on latent feature manifolds provides an effective solution to the curse of dimensionality in data mining tasks, particularly for DNA classification problems involving nominal and quantitative features.

References

- [1]. Ignatyev N. A. Structure Choice for Relations between Objects in Metric Classification Algorithms // Pattern Recognition and Image Analysis. 2018. V. 28. № 4. P. 590–597.
- [2]. Ignatev N. A. and Rahimova M. A. Formation and Analysis of Sets of Informative Features of Objects by Pairs of Classes // Scientific and Technical Information Processing, 2022, Vol. 49, No. 6, pp. 439–445.
- [3]. Saidov D.Y. Data visualization and its proof by compactness criterion of objects of classes // (IJISA), 2017. Vol 9. № 8. P. 51–58.
- [4]. Sudmant P. H. et al. "An integrated map of structural variation in 2,504 human genomes," Nature vol. 526, pp. 75–81, Oct. 2015. doi:10.1038/nature15394.