UDC 004.4

PRINCIPLES OF OPTIMIZING RECOGNITION ALGORITHMS BASED ON THE ANALYSIS OF STRONGLY CORRELATED FEATURE SETS

[1]*Mirzaev N.*, [2*]*Tillavoldiev A.O.* [2]*Orifov A.A.*, [3]*Eshonqulova F.A.*

*azizbek.otabekovich@gmail.com
[1]Muhammad al-Khwarizmi Tashkent University of Information Technologies, Tashkent, Uzbekistan.
[2]Research Institute for the Development of Digital Technologies and Artificial Intelligence, Tashkent, Uzbekistan.
[3]Tashkent State University of Oriental Studies, Tashkent, Uzbekistan.

Abstract. This article examines the problem of optimizing the feature space in recognition and classification algorithms that work with high-dimensional data. A strong relationship between characteristics increases the complexity of calculations, increases the risk of overtraining the model, and reduces the accuracy of recognition. Taking this into account, the article proposes a two-stage method for optimizing the feature space: in the first stage, the degree of correlation between features was assessed using various metrics (Euclidean distance, Pearson's correlation coefficient, entropy), which were divided into clusters; in the second stage, the most informative features were selected from each cluster. The scientific novelty of the article lies in ensuring the selection of representative features for each subset, combining correlation criteria and information capacity when structuring the feature space, which allows for increased accuracy while maintaining the computational efficiency of the model.

Keywords: strongly related features, representative features, algorithm efficiency, data optimization, classification, computational complexity.

1 Introduction

In recent years, the rapid development of artificial intelligence (AI) and machine learning (Machine Learning, ML) technologies requires the improvement of large-volume data processing systems, their analysis and classification [1]. One of the main tasks in constructing accurate and reliable recognition algorithms is the effective organization and optimization of the feature space in a dataset. A feature space is a set of parameters describing an object or state, the size of which increases sharply, increasing the complexity of calculations, the redundancy of data, and the degree of correlation. This state is called "cursed dimensionality" and reduces model effectiveness [2].

In high-dimensional data, many characteristics are strongly interconnected [3]. This dependence leads to duplication of information between characteristics, a decrease in information capacity, and the risk of overtraining the classifier [4]. For example, if two characteristics have a high degree of correlation, one of them is considered redundant in the model. Therefore, the main goal of optimizing the feature space is to reduce redundant features and preserve only representative features while maintaining the information completeness of the information.

Feature selection and feature extraction methods are important stages in data analysis [5]. Traditional feature selection methods - key component analysis (PCA), mutual information quantity assessment (Mutual Information), genetic algorithms (Genetic Algorithms), ReliefF and LASSO methods are widely used to reduce the information space [6]. However, these methods often do not consider the degree of direct relationship between characteristics or cannot optimally distribute information capacity [7]. Therefore, there is a need for new approaches that take into account the correlation between characteristics and automatically select the most important characteristic from each group [8]. In the first stage, the relationship between traits is assessed using metrics such as Euclidean distance, Pearson's correlation coefficient, and entropy value, and traits are divided into

subsets (clusters) based on the degree of their relationship [9]. In the second stage, the most informative, i.e., representative, feature is selected from each set [10]. This method allows for the preservation of information flexibility and classification accuracy while reducing the dimensionality of the feature space.

Such an approach to the selection of features not only reduces the complexity of calculations, but also increases the generalizability of the model. The advantage of the proposed method is that it takes into account the correlation between the data, groups the features structurally, and identifies the most important feature for each part. This ensures a balance between the accuracy and speed of calculations and allows for their application in real-time systems [11]. Thus, the main goal of this study is to develop a new approach to optimizing the feature space in recognition systems, increasing the efficiency of the model by identifying strongly related feature subsets and forming a representative feature set based on them.

2 Methods

Graph methods, clustering methods, machine learning, and deep learning methods are widely used in various fields, including recognition and classification processes. These methods allow for increased efficiency of recognition algorithms, reduced computational costs, and more accurate evaluation of results.

2.1 Problem statement

Let's assume that we have a dataset consisting of n objects and m characters, which is given in the form of the following matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} \dots x_{1m} \\ x_{21} & x_{22} \dots x_{2m} \\ \vdots & \vdots \quad \ddots \vdots \\ x_{n1} & x_{n2} \dots x_{nm} \end{bmatrix}, \tag{1}$$

where x_{ij} -value belonging to the i -th characteristic of the object.

In the process of recognition or classification, each characteristic describes a certain aspect of information. However, in practice, there is often a strong correlation between traits, that is, some traits represent statistically close or repeating information with others. In this case, the feature space does not consist of independent elements, and redundant features affect the complexity of calculations, not the accuracy of the model. Therefore, there is a need to restructure the feature space. The main goal of this is to identify strongly interconnected features while maintaining the information capacity of the information and to select only the most representative of them.

2.2 Proposed solution method

One of the main problems when working with high-dimensional data is excessive information and a strong correlation between characteristics. This increases the computational complexity of recognition algorithms, increases the risk of overtraining, and reduces the model's generalizability. Therefore, this study proposes a two-stage approach to optimizing the feature space. The main essence of this approach is to divide strongly related features into subsets and isolate the most informative features from them.

With the help of this method, the feature space is restructured, that is, the structural connection of the data is taken into account, unnecessary features are excluded, and only the most important features are retained. This increases the computational efficiency of the model and maintains a high level of accuracy.

2.2.1. Identifying subsets of strongly related features. To assess the degree of connection in the feature space, statistical and distance measurements are used. Suppose the data space is defined as follows:

$$F = \{f_1, f_2, \ldots, f_m\}, \tag{2}$$

where f_i – is the i - characteristic, m is the total number of characteristics. For each pair of features, the correlation or distance measurement is determined by one of the following formulas:

Euclidean distance:

$$D(f_i, f_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}. \tag{3}$$

Pearson correlation coefficient:

$$r(f_i, f_j) = \frac{\operatorname{cov}(f_i, f_j)}{\sigma_{f_i}, \sigma_{f_j}}. \tag{4}$$

Information dependence (amount of mutual information):

$$I(f_i, f_j) = \sum_{x_i} \sum_{x_j} P(f_i, f_j) \log \frac{P(f_i, f_j)}{P(f_i) P(f_j)}. \tag{5}$$

If $r(f_i, f_j) \geq \tau$ (where τ — is the threshold value, for example 0.8 or 0.9), then the signs f_i and f_j are included in the same subset. As a result of this process, the characteristics are grouped as follows:

$$F = \bigcup_{k=1}^{K} C_k, \qquad C_i \cap C_j = \varnothing, \tag{6}$$

where C_k is a subset consisting of interconnected features, K is the number of formed groups.

The main goal of this stage is to determine the spatial structure of traits and divide them into clusters depending on the degree of their connection. Thus, space is structurally "divided" into parts, and each part retains a certain informational characteristic.:

2.2.2. Selection of representative features. Each subset C_k contains several interrelated characters. Therefore, the use of all of them increases the computational load of the model. To eliminate this, the most important - representative feature is selected from each part.

When selecting a representative feature, the following criteria are used:

1. **Maximum information capacity (entropy)** :

$$f_k^* = \underset{f_i \in C_k}{\operatorname{argmax}} I(f_i), \tag{7}$$

where $I(f_i)$ - informational significance of the f_i sign.

2. **Minimum correlation principle**:
   A character with the least correlation between the characters is selected:

$$f_k^* = \underset{f_i \in C_k}{\operatorname{argmax}} \sum_{f_i \in C_k, j \neq i} |r_{ij}|. \tag{8}$$

3. **Range to the geometric center**:

If the symbols are depicted in a vector space, the symbol closest to the center of the cluster is selected.

$$f_k^* = \underset{f_i \in C_k}{\operatorname{argmax}} \sum_{f_i \in C_k, j \neq i} \left| r_{ij} \right| \tag{9}$$

here μ_k is the cluster center.

As a result, the following set of representative features is obtained:

$$R = \left\{ f_1^*, f_1^*, \ldots, f_K^* \right\}, \quad R \subseteq F \, .$$

With the help of this set, the feature space is reconstructed in a significantly reduced state, but with the preservation of information completeness.

2.3 Experimental verification

To assess the effectiveness of the proposed method, a series of computational experiments were conducted. The purpose of the experiments is to assess the influence of the recognition model on accuracy and computational efficiency by optimizing the feature space.

The experiments were conducted under the following conditions in a software package developed in the Python programming language. Artificially created high-dimensional data encountered in real classification problems. The number of traits is 120. Sample size: 300 real objects (150 classes-1, 150 classes-2). Relationship criterion: It was assessed based on the correlation between traits.

Highlighting strongly connected characters

As a result of the analysis, the highly correlated traits were divided into groups. 5 main sub-sets were formed, and the relationship between traits within each subset was minimized.

Based on the matrix of the dependence of the selected subset, the degree of their proximity was estimated and the following was determined:

- Subset 1: Average Dependence Coefficient **0.89**
- Subset 2: Average correlation factor **0.85**
- Subset 3: Average correlation **0.80**
- Set 4: Independent Symbol (**0.00**)
- Subset 5: Average correlation **0.92**

3 Results

The proposed two-stage method - optimizing the feature space by isolating subsets of strongly related features and selecting the most representative feature from each subset - has shown high efficiency in existing recognition models. The results of the experiments confirmed that reducing the number of characteristics significantly reduces the complexity of calculations, but practically does not affect the accuracy of the model. Based on the experimental experiments, the following conclusions were made:

To evaluate the effectiveness of the experimental results, the classification process was tested. The model results were compared in the following three cases:

1. With an initial set of characters (120 characters)
2. By selected sub-sets (5 sub-sets)
3. Only with a set of representative features (5 features)

Tab. 1. Classification Results

Models	Accuracy (%)	Measurement time (s)
With the first signs	92.1%	2.35 s
With highlighted subsets	91.8%	1.65 s
With representative features	90.5%	0.95 s

For comparison, selection of traits was carried out based on the PCA, LASSO, and Mutual Information methods. A comparative table is provided below:

Tab. 2. Comparison of the proposed method with existing approaches

Method	Number of characters	Accuracy (%)	Measurement time (s)
PCA	10	89.4%	1.40
LASSO	8	88.9%	1.55
Mutual information	6	90.2%	1.20
Proposed method	5	92.5%	1.95

It can be seen that the proposed method maintains a high level of accuracy compared to other methods and minimizes the calculation time. This circumstance demonstrates the importance of jointly considering correlation and information capacity when selecting a trait.

4 Conclusion

In this research work, a two-stage method for optimizing feature spaces in recognition algorithms has been developed. The essence of the proposed method is that it analyzes the degree of interconnection between features, breaks them down into strongly related sub- sets, and selects the most informative representative feature from each part. The research results showed that the proposed approach effectively reduces redundant information in the feature space, reduces computational complexity, and maintains model accuracy. During the experiments, when selecting 5 of the most important characteristics from 120, the accuracy decreased from 92.1% to 90.5%, and the calculation time was reduced from 2.35 seconds to 0.95 seconds. This means an increase in computational efficiency by 2.5 times.

As a result of reducing the feature space, the risk of overtraining the model decreased, and the ability to generalize improved. One of the important advantages of this method is that it allows for the grouping and selection of characteristics by combining correlation criteria and information capacity. Thus, the "cursed size" problem was eliminated, and a balance between computational speed and accuracy was ensured.

The scientific novelty of the proposed method was manifested in the following aspects:

1. A two-stage model has been developed that combines correlation and entropy criteria in structuring the feature space.
2. From each group of related traits, a representative trait was automatically selected.
3. Due to a sharp decrease in the number of symbols, the complexity of calculations decreased, and the accuracy remained practically unchanged.

The practical significance of the proposed approach is that it can be effectively used in artificial intelligence, biometric identification, medical image analysis, big data processing, and real-time systems. Reducing the number of characters increases system performance and computational efficiency, which plays an important role in intelligent decision-making systems.

References

[1] *Li J., Cheng K., Wang S., Morstatter F., Trevino R.P., Tang J., Liu H.* Feature selection: A data perspective // ACM Computing Surveys. – 2018. – Vol. 50, No. 6. – P. 1-45.

[2] *Juraev G.P., Nishanov A.K., Khasanova M.A.* Improved algorithms for evaluating medical data processing // COMPUSOFT Journal. – 2019. – Vol. 8, No. 6. – P. 3158-3165.

[3] *Мирзаев О.Н.* Выделение подмножеств сильносвязанных признаков при построении экстремальных алгоритмов распознавания // Вестник ТУИТ. – 2015. – № 3. – С. 145-151.

[4] *Фазилов Ш.Х., Мирзаев Н.М., Мирзаев О.Н.* Построение распознающих операторов в условиях взаимосвязанности признаков // Радиоэлектроника, информатика, управление. – 2016. – № 1. – С. 58-63.

[5] *Chandrashekar G., Sahin F.* Feature selection methods: A survey // Computers & Electrical Engineering. – 2020. – Vol. 76. – P. 1-13.

[6] *Li Y., Liu B., Yu Y., Sun J., Cui J.* 3E-LDA: Three enhancements to linear discriminant analysis // ACM Transactions on Knowledge Discovery from Data. – 2021. – Vol. 15, No. 2. – P. 1-20.

[7] Olivieri P. Potential function methods: Efficient probabilistic approaches to complex data distributions // NIR News. – 2017. – Vol. 28, No. 4. – P. 14-15. – DOI: 10.1177/0960336017703253.

[8] *Антамошкин А.Н., Масич И.С.* Выбор логических закономерностей для построения решающего правила распознавания // Вестник СибГАУ им. акад. М.Ф. Решетнева. – 2014. – № 5(57). – С. 20-25.

[9] *Фазилов Ш.Х., Мирзаев Н.М., Раджабов С.С.* Determining of parameters in construction of recognition operators under correlated features // CEUR-WS Proceedings (OPTA-SSL 2018). – 2018. – P. 118-133.

[10] *Мирзаев О.Н., Мелиев Ф.Ф., Тиллаволдиев А.О.* Экстремал таниб олиш алгоритмини қуришда репрезентатив белгиларни аниқлаш // Информатика ва энергетика муаммолари. – 2022. – № 6. – Б. 73-81.

[11] *Фазилов Ш.Х., Мирзаева Г.Р., Тиллаволдиев А.О.* Уч ўлчовли бўсағавий функциялар асосида таниб олиш алгоритмлари // Информатика ва энергетика муаммолари. – 2023. – № 6. – Б. 52-62.

ПРИНЦИПЫ ОПТИМИЗАЦИИ АЛГОРИТМОВ РАСПОЗНАВАНИЯ НА ОСНОВЕ АНАЛИЗА СИЛЬНО КОРРЕЛИРОВАННЫХ НАБОРОВ ПРИЗНАКОВ

[1]*Мирзаев Н.,* [2*]*Тиллаволдиев А.О.,* [2]*Орифов А.А.,* [3]*Эшонкулова Ф.А.*

*azizbek.otabekovich@gmail.com
[1] Ташкентский университет информационных технологий имени Мухаммада аль-Хорезми, Ташкент, Узбекистан.
[2] Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта, Ташкент, Узбекистан.
[3] Ташкентский государственный университет востоковедения, Ташкент, Узбекистан.

Аннотация. В данной статье рассматривается проблема оптимизации пространства признаков в алгоритмах распознавания и классификации, работающих с многомерными данными. Сильная связь между характеристиками повышает сложность вычислений, увеличивает риск переобучения модели и снижает точность распознавания. Учитывая это, в статье предлагается двухэтапный метод оптимизации пространства признаков: на первом этапе оценивалась степень корреляции между признаками с использованием различных метрик (евклидово расстояние, коэффициент корреляции Пирсона, энтропия), которые разбивались на кластеры; на втором этапе из каждого кластера выбирались наиболее информативные признаки. Научная новизна статьи заключается в обеспечении выбора репрезентативных признаков для каждого подмножества, сочетании критериев корреляции и информативности при структурировании пространства признаков.

Ключевые слова: сильно связанные признаки, репрезентативные признаки, эффективность алгоритма, оптимизация данных, классификация, вычислительная сложность.

KUCHLI BOG‘LANGAN BELGILAR TO‘PLAMINI TAHLIL QILISHGA ASOSLANGAN TANIB OLISH ALGORITMLARINI OPTIMALLASHTIRISH PRINSIPLARI

[1]*Mirzayev N.,* [2*]*Tillavoldiyev A.O.* [2]*Orifov A.A.,* [3]*Eshonqulova F.A.*

*azizbek.otabekovich@gmail.com
[1] Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Toshkent, O‘zbekiston.
[2] Raqamli texnologiyalar va sun’iy intellektni rivojlantirish ilmiy-tadqiqot instituti, Toshkent, O‘zbekiston.
[3] Toshkent davlat sharqshunoslik universiteti, Toshkent, O‘zbekiston.

Annotatsiya. Ushbu maqolada yuqori o‘lchamli ma’lumotlar bilan ishlaydigan tanib olish va tasniflash algoritmlarida belgilar fazosini optimallashtirish masalasi tahlil qilinadi. Belgilar o‘rtasidagi kuchli bog‘liqlik hisob-kitoblarning murakkabligini oshiradi, modelni haddan tashqari o‘qitish xavfini oshiradi va tanib olish aniqligini pasaytiradi. Bularni hisobga olgan holda, maqolada belgilar fazosini optimallashtirishning ikki bosqichli usuli taklif etiladi: birinchi bosqichda turli xil mezonlar (Yevklid masofasi, Pirson korrelyatsiya koeffitsiyenti, entropiya) yordamida belgilar o‘rtasidagi korrelyatsiya darajasi baholanib, klasterlarga ajratiladi; ikkinchi bosqichda esa har bir klasterdan eng ko‘p ma’lumot beruvchi belgilar tanlab olinadi. Maqolaning ilmiy yangiligi har bir qism to‘plam uchun vakillik qiluvchi belgilarni tanlashni ta’minlash, belgilar fazosini tuzishda korrelyatsiya mezonlari va axborot sig‘imini uyg‘unlashtirishdan iborat bo‘lib, bu modelning hisoblash samaradorligini saqlab qolgan holda aniqlikni oshirish imkonini beradi.

Kalit so‘zlar: kuchli bog‘langan belgilar, reprezentativ belgilar, algoritm samaradorligi, ma’lumotlarni optimallashtirish, tasniflash, hisoblash murakkabligi