SYNTHESIS OF A BIT-BASED FEATURE USING THE K-NEAREST NEIGHBORS METHOD FOR DETECTING CONFLICT OBJECTS IN MEDICAL DATA

Shokhsanam Ergasheva Elmurod kizi
Teacher at Computational mathematics and information systems,
National University of Uzbekistan named after Mirzo Ulugbek,
Email: ergasheva.shohsanam98@gmail.com
https://doi.org/10.5281/zenodo.20661991

Abstract

The identification of hidden patterns and boundary objects remains one of the important problems in pattern recognition, data mining, and machine learning. Traditional classification methods focus mainly on assigning objects to predefined classes and often provide limited information about the local structural properties of the data. This paper proposes a method for synthesizing a new informative feature based on the local neighborhood structure of objects using the K-Nearest Neighbors (KNN) algorithm. The proposed approach transforms the class composition of the nearest neighbors into a binary sequence and subsequently into a synthetic numerical feature. This feature reflects the local environment of each object and can be used for the detection of conflict objects, analysis of class boundaries, and discovery of latent relationships between classes.

The method was evaluated using a heart disease dataset containing 270 observations and 13 attributes. Both Euclidean and Juravlyev distance metrics were employed for neighborhood formation. Experimental results demonstrate that the synthesized feature effectively reveals hidden structural regularities within the dataset and identifies objects with similar local environments despite belonging to different classes. Such objects may indicate uncertainty regions, class overlap, or potential anomalies. The obtained results show that the proposed approach can serve as an additional analytical tool in classification and decision-support systems.

Keywords: synthetic feature, K-nearest neighbors, binary sequence, conflict objects, local neighborhood, Juravlyev metric, Euclidean metric, heart disease dataset, pattern recognition.

1. Introduction

The rapid growth of data collection technologies has led to the emergence of large datasets in medicine, finance, engineering, and social sciences. One of the key challenges in analyzing such datasets is the extraction of meaningful and interpretable information that can support decision-making processes. Classification algorithms are widely used for assigning objects to predefined classes; however, they often provide only the final classification result without revealing the underlying local relationships among objects.

The K-Nearest Neighbors (KNN) algorithm is one of the most popular non-parametric classification methods. The algorithm determines the class of an object based on the classes of its nearest neighbors. Although KNN is simple and effective, its outputs are generally limited to predicted class labels. The neighborhood structure itself contains valuable information that is rarely utilized as a separate source of knowledge.

In many real-world datasets, objects from different classes may occupy similar regions of the feature space. Such objects often appear near class boundaries and can significantly affect classification performance. Detecting these objects is important because they may correspond to uncertain cases, atypical observations, or hidden subclasses.

To address this issue, this study proposes a method for generating a synthetic feature based on the class composition of an object's local neighborhood. The method converts the neighborhood structure into a binary sequence, providing an interpretable representation of local class distributions. This representation allows the identification of conflict objects and facilitates the discovery of latent patterns in the data.

2. Proposed Method

2.1 Local Neighborhood Representation

Let X = {x_1, x_2, ..., x_n} be a set of objects belonging to two classes: K_1and K_2. For each object (x_i), the K nearest neighbors are determined using a selected distance metric. Two distance measures are considered:

1. **Euclidean distance**
2. **Juravlyev distance**

which combines nominal and quantitative attributes by considering normalized differences between feature values. After identifying the K nearest neighbors, a binary sequence is constructed according to the class membership of each neighbor:

- K_1 - 0
- K_2 - 1

For example, if the nearest neighbors belong to classes (K_1,K_1,K_2,K_2,K_2) then the corresponding binary sequence becomes [00111]

2.2 Synthetic Feature Construction

The generated binary sequence is transformed into a decimal value: 00111_2 = 7_{10}

This decimal value serves as a synthetic feature describing the local environment of the object.

The synthesized feature possesses several useful properties:

- reflects neighborhood composition;
- characterizes class purity;
- indicates boundary regions;
- supports conflict object detection;
- provides interpretable numerical representation.

Objects located deeply inside a class tend to produce pure binary sequences such as: [00000] or [11111] whereas boundary objects generate mixed sequences such as: [00111,; 01011,; 10110] and others.

3. Conflict Objects and Local Similarity

A particularly important situation occurs when two objects belonging to different classes generate identical binary sequences.

For example:

Object	Class	Binary Sequence
A	K1	10111
B	K2	10111

Although these objects belong to different classes, their local neighborhoods are structurally identical.

Such objects are referred to as **conflict objects**.

Conflict objects indicate:

- class overlap;
- uncertainty regions;
- hidden structural dependencies;
- possible labeling ambiguities;
- latent subclasses.

The identification of conflict objects provides additional information that cannot be obtained directly from conventional classification results.

4. Experimental Study

The proposed approach was evaluated on a heart disease dataset consisting of:

- 270 objects;
- 13 attributes;
- 150 objects in Class 1;
- 120 objects in Class 2.

The dataset contains both nominal and quantitative features describing medical characteristics of patients. Neighborhood structures were generated using Euclidean and Juravlyev distance metrics. Classification performance was evaluated using Leave-One-Out Cross Validation (LOOCV). The obtained accuracy values are summarized below.

K	Juravlyev (%)	Euclidean (%)
3	81.11	79.26
5	81.11	80.00
7	80.37	81.11
9	81.11	80.37
11	82.96	81.48

The highest classification accuracy was achieved using the Juravlyev metric with (K=11).

The results indicate that larger neighborhood sizes reduce the influence of local noise and improve classification stability.

5. Visualization and Interpretation

To better understand the synthesized feature, Principal Component Analysis (PCA) was applied to project the data into two- and three-dimensional spaces.

Visualization revealed that:

- pure binary sequences tend to appear in dense class regions;
- mixed sequences are concentrated near class boundaries;
- conflict objects occupy overlapping regions between classes;
- clusters of identical binary patterns correspond to hidden local structures.

The PCA representation confirmed that the synthesized feature captures meaningful geometric information about the dataset.

6. Conclusion

This paper presented a method for synthesizing a new feature based on the local neighborhood structure of objects using the K-Nearest Neighbors algorithm. The proposed

approach converts neighborhood class compositions into binary sequences and subsequently into numerical synthetic features.

Experimental results obtained on a heart disease dataset demonstrated that the synthesized feature effectively characterizes local environments and supports the identification of conflict objects. Objects with identical local structures but belonging to different classes were successfully detected, revealing hidden patterns and uncertainty regions within the dataset.

The proposed methodology extends the traditional use of KNN from a classification tool to a mechanism for knowledge extraction and structural analysis. Future research may focus on multi-class problems, adaptive neighborhood sizes, and integration of the synthesized feature into ensemble learning and explainable artificial intelligence systems.

Adabiyotlar, References, Литературы:

1. Guyon I., Elisseeff A. An Introduction to Variable and Feature Selection // Journal of Machine Learning Research. 2003. Vol. 3. P. 1157–1182.
2. Тухтабаев К.А., Эргашева Ш.Э. Аналитические выражения для вычисления значений латентных признаков // Proceedings of the Seminar dedicated to the memory of professor M.I. Isroilov, CMT2024. Tashkent, 2024. P. 159–162.
3. Tenenbaum J.B., de Silva V., Langford J.C. A Global Geometric Framework for Nonlinear Dimensionality Reduction // Science. 2000. Vol. 290. No. 5500. P. 2319–2323.
4. Ignatev N.A., Akbarov B.K., Tuhtabayev K.A. Deriving Analytical Expressions for Calculating the Values of Latent Features in Recognition Problems // Problems of Computational and Applied Mathematics. 2023. No. 6/1(54). P. 68–76.
5. Ergasheva Sh.E. Dimensionality Reduction of Feature Space Using Nonlinear Transformations of Heterogeneous Features // Education, Science and Innovative Ideas in the World. 2024.
6. Belkin M., Niyogi P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation // Neural Computation. 2003. Vol. 15. No. 6. P. 1373–1396.
7. Ignatev N.A. On Nonlinear Transformations of Features Based on the Functions of Objects Belonging to Classes // Pattern Recognition and Image Analysis. 2021. Vol. 31. No. 2. P. 197–204.