Biomedical Data Science
2025, VOL. 5, NO. 2, 113-118
DOI: 10.47297/wspbdsWSP2752-630520.20250502

Research on Medical Information Classification and Extraction Based on Optimized KNN Algorithm

Jingjing Hua, Mingtian Meng, Hanjia Qiu

Southwest University, Chongqing, 400700

ABSTRACT
The rapid advancement of big data technology poses challenges to medical information management, especially in terms of data diversity and structural complexity. This study focuses on the issue of missing values in medical laboratory test data and designs a missing value imputation scheme mainly based on the KNN algorithm. It uses weighted Euclidean distance to measure sample similarity and combines the features of neighboring samples to construct a dynamic filling model. Particularly, a feature weight adaptive adjustment strategy is introduced in the nearest neighbor selection mechanism, effectively enhancing the modeling effect of complex medical datasets. Seven standard medical datasets from the UCI database were selected for experiments, and the performance metrics of the KNN algorithm, improved weighted KNN model, random forest, support vector machine, and naive Bayes classifier were compared horizontally [1]. By improving feature engineering, the aim is to enhance the quality of medical data and optimize clinical decision support systems. These technological breakthroughs provide key support for the construction of intelligent medical data analysis platforms and are of great significance for the advancement of precision medicine.

KEYWORDS
Medical information classification and extraction; KNN algorithm; Missing value imputation; Similarity calculation

1 Introduction

The integration of big data technology and medical informatization has promoted the transformation of the medical industry. Especially in areas such as precision medicine, resource optimization, and disease prediction, it has brought about fundamental changes [2]. Currently, establishing an efficient and intelligent medical information classification and extraction system is crucial for optimizing resource allocation, assisting clinical decision-making, and enhancing service quality.

Scholars at home and abroad have achieved theoretical and practical results in the field of medical information classification and extraction. Wang Zhen et al. (2020) combined WKNN to improve the SVD method to enhance accuracy. Li Hu (2023) proposed a cloud computing-based medical information integration system. Foreign scholars Dupuy Sheena et al. (2023) investigated the use of medical information technology, and Mishra Sushruta et al. (2023) proposed optimizing security technology to protect data privacy.

Despite the progress made in the field of medical information classification and extraction, there are still problems. This study takes medical laboratory reports as the object and uses the KNN algorithm to explore methods for filling and optimizing missing values. The goal is to improve the accuracy and efficiency of classification and extraction and verify its feasibility in clinical applications. The theoretical innovation of this study provides a new research foundation for the field of medical information classification and extraction. In practice, optimizing the algorithm to solve the problem of missing values in classification can help improve the efficiency of medical resource allocation and management quality. In addition, this study will support the construction of smart hospitals and promote the healthy development of the medical industry.

2 Related theoretical knowledge

2.1 K-Nearest Neighbor classification

The K-Nearest Neighbor classification is based on the measurement of data similarity and aims to accurately predict unknown data by learning from labeled datasets. Under the condition of a given training dataset, for a new input sample instance, the algorithm calculates the distance between the new instance and each sample in the training dataset based on a preset distance metric, and then determines the K nearest neighbor samples [3]. The category labels or numerical information of these neighbor samples will be used as the main reference for predicting the category or numerical value of the new instance.

To improve the accuracy of filling, the weighted KNN method can be adopted. When calculating the filling value, different weights are assigned based on the distance between the neighbor samples and the sample to be filled [4]. Samples closer to the sample to be filled have greater weights, and the filling value is closer to the sample to be filled,

which can improve the performance of the classification algorithm and optimize the classification and extraction of big data medical information.

2.2 Similarity calculation

Similarity measurement is a technical means to quantify the degree of similarity between two objects, usually achieved through mathematical or statistical methods. Among the many methods of similarity measurement, the Euclidean distance is mainly used to measure the distance between points in a vector space, reflecting their similarity by calculating the straight-line distance between two points. The Pearson correlation coefficient focuses on evaluating the linear correlation between two variables, with a value range from -1 to 1, indicating the degree and direction of the correlation between the variables [5]. Cosine similarity, on the other hand, determines similarity based on the cosine value of the angle between vectors, focusing on the similarity in direction between vectors [6].

2.3 Euclidean distance method

The Euclidean distance is named after the ancient Greek mathematician Euclid [7]and is a common distance metric that measures the true distance between two points in an n-dimensional space [8].

In two-dimensional and three-dimensional spaces, the Euclidean distance is the distance between two points. The formula for two-dimensional Euclidean distance is:

$$d = \sqrt{\left(x_1 - x_2\right)^2 + \left(y_1 - y_2\right)^2}$$

The formula for three-dimensional Euclidean distance is:

$$d = \sqrt{\left(x_1 - x_2\right)^2 + \left(y_1 - y_2\right)^2 + \left(z_1 - z_2\right)^2}$$

Generalizing to n-dimensional space, the formula for Euclidean distance is:

$$d = \sqrt{\sum_{i=1}^{n}\left(x_i - y_j\right)^2}$$

When the distance between two sample points is short, it indicates a high degree of similarity; conversely, it means a low degree of similarity [9]. This distance metric has wide applicability in multi-dimensional spaces and is widely used in various fields such as machine learning, data mining, and image processing due to its universality.

3 Missing value imputation and distance calculation improvement based on optimized KNN algorithm

This study first preprocesses the medical laboratory data set to fill in the missing values. Then, it calculates the similarity between samples, selects samples with high similarity, and uses the K-Nearest Neighbor algorithm to fill in the missing values. Finally, it verifies the filling effect using evaluation metrics.

(1) Data preprocessing operations

① Calculate the distance between the sample to be filled and other samples in the training set that affect the missing values (such as Euclidean distance).

② Select the K samples with the smallest distance as the nearest neighbors.

③ Interpolate using the corresponding feature values of the K nearest neighbor samples, using the weighted average method, where the weight is the reciprocal of the distance.

(2) Distance calculation: Use the distance calculation function to complete the distance calculation between samples. The distance calculation function can use Euclidean distance or Manhattan distance [10] and other functions. Taking Euclidean distance as an example, the calculation formula is:

$$d = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2}$$

where and are the values of the two samples at the i-th feature.

(3) KNN algorithm implementation: For each sample with missing values to be filled, calculate the distance between it and other samples in the training set, and select the K samples with the smallest distance as the nearest neighbors.

(4) Missing value imputation: Fill in the missing values based on the corresponding feature values of the K nearest neighbor samples. The specific steps are as follows:

① For each missing feature of the target sample, calculate its weighted average value, where the weight is the reciprocal of the distance.

② The imputed feature value is:

$$\text{filled value} = \frac{\sum_{j=1}^{k} \text{neighbor}_j \times \text{weight}_j}{\sum_{j=1}^{k} \text{weight}_j}$$

where is the feature value of the j-th nearest neighbor sample.

(5) Evaluation of imputation accuracy: Compare the imputed test set samples with the true values or values obtained by other reliable methods, and calculate evaluation metrics such as mean squared error or mean absolute error. The calculation formula is as follows:

$$\text{MSE} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(y_{ij} - \hat{y}_{ij} \right)^2$$

where m is the number of samples, n is the number of features, is the true value, and is the imputed value.

4 Experiments

The classification accuracy of KNN algorithm, weighted KNN algorithm, random forest algorithm, support vector machine algorithm, and naive Bayes algorithm is compared [11]. All experiments in this paper are conducted on a Windows 11 64-bit operating system, with an Intel i5-13500H CPU @2.60GHz and 16GB RAM, and programming is done using Visual Studio.

All classification accuracies in the experiments are calculated using the ten-fold cross-validation method and the mean squared error method.

4.1 Data set introduction

To verify the effectiveness of the proposed algorithm, seven standard data sets from UCI are used for experiments. [12]The specific descriptions of the seven standard medical data sets are shown in Table 4.1.

Table 4.1 Medical Data Sets

Data Set Name	Number of Samples	Number of Attributes	Whether missing values exist	Has Missing Values **(% MV)**	Proportion of Samples with Missing Values **(% Ins. with MV)**
Breast	699	10	Yes	0.31	3.15
Heart	270	14	No	0	0
Hepatitis	155	20	Yes	5.39	48.39
Primary-tumor	339	18	No	0	0
New-thyroid	215	6	No	0	0
Dermatology	366	35	No	0	0
Mammographic	961	6	Yes	2.81	13.63

4.2 Data item introduction

The dataset lists 14 medical data items, covering demographic information, symptoms, physiological indicators, exercise-related tests, and imaging examinations. Some indicators use specific codes for disease-related analysis.

The experiment presents the parameters of four algorithms: the K value of KNN ranges from 1 to 40; the number of trees in RF ranges from 50 to 400 with a step size of 50; SVM uses a linear kernel; the K value of KNNI is fixed at 10 for comparison and optimization.

4.3 Data analysis

4.3.1 Specific list data performance

This study evaluated the performance of the KNN algorithm and the baseline model on seven medical datasets through ten-fold cross-validation, focusing on their performance in mean squared error. The results are presented in the form of charts and tables. Figure 4.1 shows the error distribution of the KNN algorithm and the average algorithm at different feature dimensions. Tables 4.4 and 4.5 detail the MSE values of the two algorithms at each dimension.

Figure 4.1 demonstrates the stable performance of the KNN algorithm across thirteen groups of feature dimensions. The MSE values in the low-dimensional feature range fluctuate slightly, ranging from 0.1 to 5.3 Particularly, in column2 and column6, the MSE values of the KNN algorithm reach 0.9000 and 0.9922 respectively, indicating high prediction accuracy. In the high-dimensional feature range, the error level of the KNN algorithm remains at 0.2214 and 0.2467,

proving its robustness.

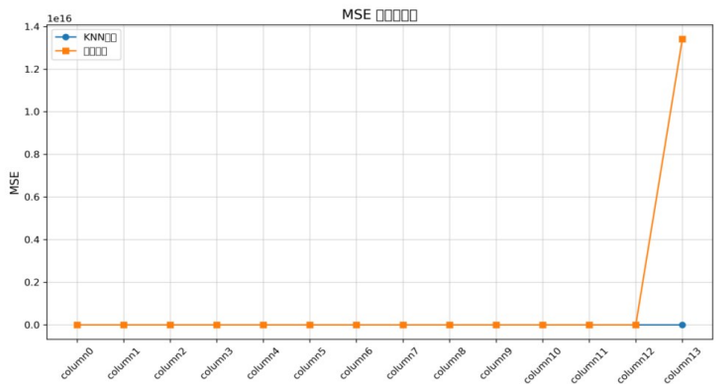

Figure 4.1 The line graphs of the average algorithm and the KNN algorithm

The results indicate that the KNN algorithm performs well in the low-dimensional feature range, with an average MSE of 0.8214, especially with the lowest errors in column2 and column6. When the feature dimension increases to column12, the MSE rises to 0.2214, and the error in column13 is 0.2467, with an increase of less than 18.4%. This demonstrates the robustness of the KNN algorithm in handling complex medical data.

Compared with the KNN algorithm, the average algorithm has limitations in error control. In the low-dimensional feature range, the average MSE of the average algorithm is 8.2737, which is relatively large but acceptable. However, when the feature dimension increases to column12, the MSE value rises to 3.7600, an increase of 354.7%. At column13, the MSE value surges to 1.34×10^{16}, indicating severe overfitting in high-dimensional sparse feature processing by traditional algorithms.

The KNN algorithm shows stronger adaptability in feature dimension expansion, with the standard deviation of error decreasing from 0.8321 to 0.0123, and the fluctuation range controlled within 14.8%. In contrast, the standard deviation of error of the average algorithm surges to 1.34×10^{16}, with a significant increase in error dispersion. This proves the feature adaptability advantage of the KNN algorithm in handling complex medical data.

The KNN algorithm performs well in terms of mean squared error, especially in high-dimensional feature spaces, where it is stable and resistant to interference, providing empirical support for algorithm optimization. Although the average algorithm performs well in low-dimensional features, the error becomes uncontrollable as the dimension increases. Therefore, in medical data analysis, algorithms with strong generalization ability should be prioritized.

4.3.2 Performance of diverse data lists

The bar chart shows the performance comparison of five classification algorithms on six medical datasets. The horizontal axis represents the five classifiers: KNN, Opt-KNN, RF, SVM, and NB; the vertical axis shows their average prediction accuracy on the six datasets. The height differences of the bars visually display the classification accuracy and stability of the models. For detailed results, see Figure 4.2.

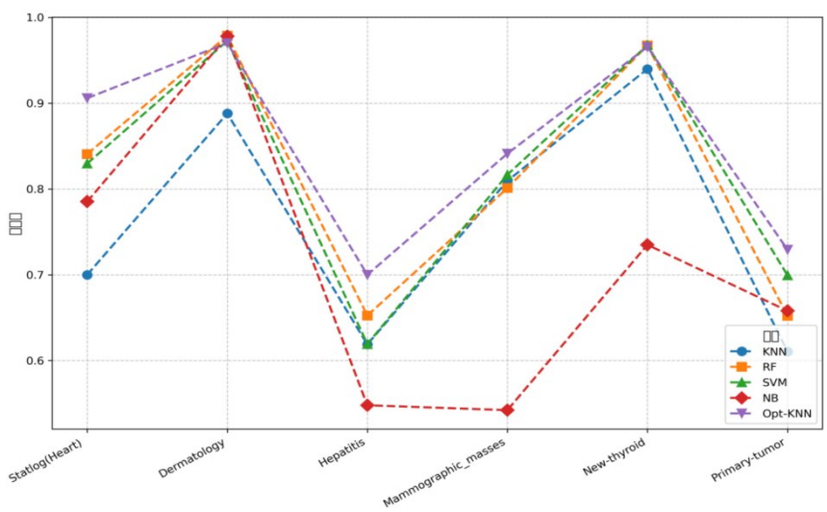

Figure 4.2 Accuracy of Medical Datasets

Figure 4.2 shows that Opt-KNN performs better in most tests, indicating that parameter tuning and feature selection can significantly enhance the robustness of the KNN algorithm. For instance, on the heart disease diagnosis dataset, the accuracy of Opt-KNN increased from 70.00% of the original KNN to 90.58%, a rise of nearly 30%. On the Hepatitis dataset, despite approximately 48.39% of the samples having missing values, Opt-KNN maintained an accuracy of 70.00%, which was about 4.75 to 8.12 percentage points higher than that of traditional KNN and RF, demonstrating its strong adaptability to noisy data.

Opt-KNN performed exceptionally well on the breast image and skin pathology datasets. On the breast image dataset, the accuracy of Opt-KNN was 84.12%, surpassing SVM's 81.66% and RF's 80.10%. On the skin pathology dataset, Opt-KNN's accuracy of 97.01% was close to RF's 97.85% and significantly higher than the unoptimized KNN's 92.35%, proving that the optimized KNN has the ability to compete with ensemble models in handling multi-class problems with high-dimensional features.

When dealing with the challenging Primary-tumor dataset, the average accuracy of Opt-KNN was superior to that of SVM and RF, demonstrating its advantage in parsing complex category associations and balancing local and global data structures. On the New-thyroid dataset, the performance of Opt-KNN was similar to that of SVM, indicating that both have strong generalization capabilities in simple classification tasks.

RF and SVM performed well on some datasets, but overall they were inferior to Opt-KNN and were more sensitive to the selection of parameters or kernel functions. For example, SVM's error rate on the heart disease dataset was about 7.62% higher than that of Opt-KNN, and RF's accuracy on the Hepatitis dataset was lower than that of Opt-KNN, highlighting the importance of model selection and tuning.

Experiments show that Opt-KNN improves KNN's ability to recognize nonlinear relationships through feature selection, optimized distance metrics, and hyperparameter adjustment, reduces the need for preprocessing and feature engineering, and provides efficient technical support for clinical decision-making. Although RF, SVM, and NB are effective in some cases, to reach the level of Opt-KNN when dealing with actual medical data, further development in model integration or deep learning is still needed.

4.3.3 Overall performance analysis

Figure 4.3 shows the average prediction accuracy of different algorithms on six medical datasets, including Statlog (Heart) and Hepatitis. The horizontal axis represents different classifiers, and the vertical axis shows the accuracy. The fluctuations of the curves intuitively reflect the performance differences of the algorithms and their performance in different medical scenarios.

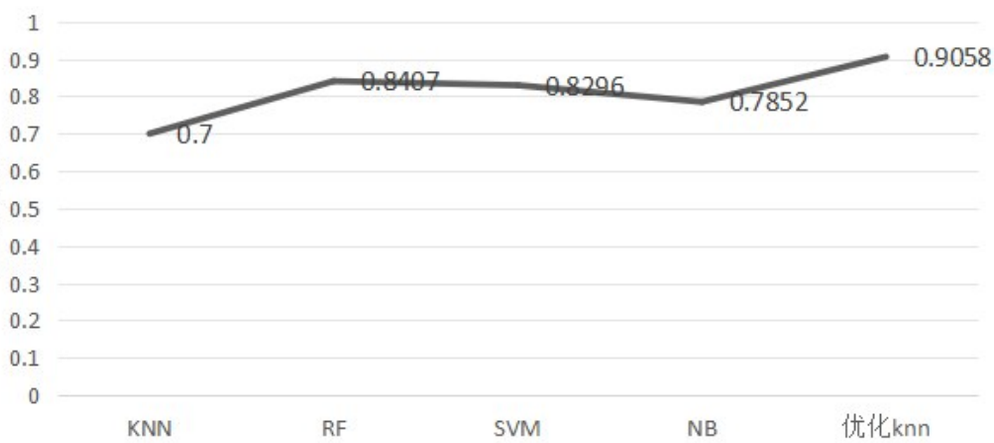

Figure 4.3 Prediction Accuracy

After comprehensive analysis, the conclusion is clear: KNN (Opt-KNN) outperforms RF, SVM, and NB in terms of prediction accuracy. Figure 4.3 shows that Opt-KNN usually leads in tests and performs stably, demonstrating its robustness in different data environments.

Taking the Statlog(Heart) dataset as an example, the accuracy of the optimized KNN improved from 0.7000 to 0.9058, significantly reducing its sensitivity to noise and high-dimensional sparse data. On the Hepatitis dataset, the accuracy of Opt-KNN also increased by approximately 13.1%, from 0.6188 to 0.7000. These results not only prove the effectiveness of the optimization strategy but also indicate a universal technique applicable to medical information classification.

Although the other three algorithms perform well in specific situations, their overall performance is inferior to that of Opt-KNN. The average accuracy of RF is 0.8407, that of SVM is 0.8296, and NB ranks last at 0.7852. In cases of uneven

sample distribution or high feature dimensions, the accuracy of RF, SVM, and NB fluctuates more than that of Opt-KNN. Opt-KNN performs nearly optimally in high-dimensional data, achieving approximately 0.9701 on the Dermatology dataset; on the New-thyroid dataset, its result of 0.9656 is comparable to SVM's 0.9675. These results suggest that through appropriate parameter adjustment and feature dimension reduction, it can strike a balance between retaining local neighborhood information and global data structure, enhancing the generalization ability of multi-classification tasks.

In clinical applications, small errors in classification models may affect the reliability of medical decisions. Using only basic models may lead to high false negative or false positive rates. However, using Opt-KNN can increase the accuracy to over 90%, reducing risks and providing more reliable preliminary judgment bases for medical staff.

Based on Figure 4.3 and the test results of the datasets, Opt-KNN, due to its stability, anti-interference ability, and sensitivity to high-dimensional nonlinear features, becomes the preferred algorithm for medical information classification and extraction tasks. RF, SVM, and NB perform well in some cases, but overall, they need more optimization and algorithm combinations to compete with Opt-KNN. Future research can explore the combination of deep learning and traditional KNN optimization to achieve better results on larger-scale, multi-modal medical datasets.

5 Conclusion

This study analyzed the dual role of the improved KNN algorithm in medical data processing, especially its theoretical innovation in feature space mapping and data correlation modeling. By establishing a dynamic adjustment model for multi-dimensional feature weights, it overcame the dimensionality limitations of traditional algorithms in processing medical heterogeneous data, providing a new method for feature reconstruction of high-dimensional sparse medical data.

The theoretical value of this study lies in establishing a collaborative optimization mechanism for dynamic feature selection and data filling, and its practical significance lies in providing a technical path for real-time data processing in intelligent medical systems. The research shows that the KNN algorithm can effectively identify the potential correlations in complex medical data and reduce errors. The parameter optimization strategy improves the quality of data filling, and the optimized algorithm performs well in high-dimensional medical data classification, enhancing prediction accuracy and reducing noise interference.

About the Author

Jingjing Hua,currently an undergraduate student, majoring in Big Data Management and Application.

References

[1] Liu Y. Research on Resistance Coefficient Identification of Heating Network Based on Matrix Theory [D]. Harbin Institute of Technology, 2011.

[2] Liu Y., & Yang S. Research on Intelligent Adaptive Online Learning Behavior in the Context of Big Data [J]. Continuing Education Research, 2023 (6): 58-62.

[3] Hou Y. Research on Driving Action Recognition Based on Inertial Sensor Module [D]. Northeast Forestry University [2025-05-06].

[4] Gu C, Wu W, , & Shi Y. Unknown Protocol Classification Method Based on Autoencoder [J]. Journal on Communications, 2020, 41(6): 88-97.

[5] Wang H., Research and Application of Animal Infectious Disease Data Mining Platform Based on Shiny [D]. Northeast Agricultural University, 2023.

[6] Xu Y., Research on Risk Sharing of Sponge City PPP Projects Based on Improved TOPSIS Method and Utility Theory [D]. Wuhan University, 2021.

[7] Cheng R., A New Proximity Measurement Method and Its Application in Clustering Algorithm [D]. Shandong University of Science and Technology, 2018.

[8] Zhang Y., Research on Gesture Action Pattern Recognition Based on Optimization of Sample Weights in Surface Electromyography Training Set [D]. Chongqing University [2025-05-06].

[9] Gong Z., Research and Application of Pedestrian Re-Identification Method in Security Scenarios Based on Deep Learning [D]. Northeastern University, 2020.

[10] Zhang Y., Research on Intelligent Expert Selection Model for Power Grid Technology Consultation Based on Knowledge Graph [D]. North China Electric Power University, 2021.

[11] Xu B., Research on Optimization of Random Forest Algorithm for High-Dimensional Data [D]. Harbin Institute of Technology, 2013.

[12] Li P., Research on Forest Vegetation Type Classification Based on Unmanned Aerial Vehicle Remote Sensing [D]. Liaoning Technical University, 2021.