Research on SPDTRS-PNN Based Intelligent Assistant Diagnosis for Breast Cancer

Mengran Zhou

Anhui University of Science and Technology

Xixi Kong (✉ kxxdyx222@163.com)

Anhui University of Science and Technology

Kai Bian

Anhui University of Science and Technology

Wenhao Lai

Anhui University of Science and Technology

Feng Hu

Anhui University of Science and Technology

Rongying Dai

Anhui University of Science and Technology

Jingjing Yan

Anhui University of Science and Technology

Oumaima Hamrani

Anhui University of Science and Technology

Research Article

Keywords: Breast cancer, Dimension reduction algorithm, machine learning, SPDTRS-PNN

Posted Date: September 23rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-850621/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Research on SPDTRS-PNN based intelligent assistant diagnosis for breast cancer

Mengran Zhou^{1,2}, Xixi Kong^{1*}, Kai Bian¹, Wenhao Lai¹, Feng Hu¹, Rongying Dai¹, Jingjing Yan¹, Oumaima Hamrani¹

Abstract

Background: Breast cancer is the second dangerous cancer in the world. How to identify breast cancer quickly and accurately is of great help to the treatment of breast cancer. Breast cancer data often contains more redundant information. Redundant information makes the breast cancer auxiliary diagnosis less accurate and time-consuming. Dimension reduction algorithm combined with machine learning can solve these problems well.

Methods: This paper proposes the single-parameter decision-theoretic rough set (SPDTRS) combined with the probability neural network (PNN) model for breast cancer diagnosis. We structure fifteen models by combining five dimensionality reduction algorithms with three classification algorithms. We compared the accuracy and test time of fifteen models under different parameters or dimensions. We find that when the parameter value of SPDTRS is 2.5, the classification effect of SPDTRS combined with PNN is better. At this point, the number of 30 attributes of the original breast cancer data dropped to 12. Then the SPDTRS-PNN model is further optimized. We compared the accuracy and test time of the model under different SPREAD values in PNN, and established a better SPDTRS-PNN model.

Result: We find the parameter value of SPDTRS is 2.5 and the SPREAD value is 0.75, the accuracy of the SPDTRS-PNN model training set is 99.25%, the accuracy of the test set is 97.04%, and the test time is 0.093s.

Conclusion: The experimental results show that the SPDTRS-PNN model can improve the accuracy of breast cancer recognition, reduce the time required for diagnosis, and achieve rapid and accurate breast cancer diagnosis.

Keywords: Breast cancer, Dimension reduction algorithm, machine learning, SPDTRS-PNN

Background

---

* Correspondence: kxxdyx222@163.com. ¹School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan 232001, China, ²State Key Laboratory of Mining Response and Disaster Prevention and Control in Deep Coal Mines, Anhui University of Science and Technology, Huainan 232001, China. Full list of author information is available at the end of the article.

Breast cancer is the most common cancer diagnosed by women aged 20-60. There are more than 2 million newly diagnosed cases of breast cancer worldwide each year. Although the incidence rate of breast cancer has gradually declined in recent years to stabilize [1], breast cancer is still one of the most common types of cancer in women, which seriously affects diseases that threaten women's life and health [2-4]. Early breast cancer is commonly identified by mammography, ultrasound, and so on. However, breast cancer still nearly 30% of cases are detected in the late stage of breast cancer [5]. We can improve the success rate and reduce the mortality rate if we find breast cancer early [6-7]. Therefore, how to diagnose malignant tumors quickly and accurately is the key in the treatment of breast cancer.

The early treatment of breast cancer is needle biopsy based on tissue biopsy. In this method, a thin hollow needle into the lump to sample cells, examining the cells sampled under a microscope. But this method may lead to misdiagnosis in the process of data collection because of some uncertain factors. In addition, when the pathologist manually inspects the abnormality, their experience may affect the diagnostic results [8-9]. To solve this problem, Wolberg, W. H. et al. (1994) [10] tried to use machine learning technology to reduce the subjectivity inherent in the visual diagnosis of needle aspiration cytology. Nowadays, many

algorithms in machine learning can distinguish benign and malignant breast cancer samples well and better assist in medical diagnosis [11-12]. For example, Al-Timemy A H et al. (2009) [13] used fine needle aspiration cytology combined with PNN to achieve rapid and accurate classification of breast tumors. H. M. Whitney et al. (2020) [14] proved the practicability of transfer learning in computer-aided diagnosis by using the breast fusion classifier based on convolutional neural network (CNN) transfer learning combined with magnetic resonance imaging (MRI). R. Nagpur et al. (2020) [15] used adaptive mean, gaussian mixture model (GMM) segmentation, and probabilistic neural network (PNN) classifier to predict whether there are benign or malignant cells in a given mammogram can help patients find diseases faster and take appropriate measures. But most of these machine learning analyzed all the features contained in the breast cancer dataset. They did not consider whether the data set contains redundant information and whether the redundant information will affect the experimental results. Some dimensionality reduction algorithms and classification models were proposed to identify malignant breast tumors in Wisconsin by using Wisconsin Breast Cancer Database (WBCD) [16]. For example, Zhou J et al. (2015) [17] used principal component analysis (PCA) to preprocess the original breast cancer data and use the improved PNN model to realize the recognition of breast

tumors, to achieve the auxiliary diagnosis of breast tumors. Kejriwal, n. et al. (2018) [18] used a univariate feature selection algorithm combined with logic and neural network algorithm to obtain a good classification effect. Bian K et al. (2020) [19] Proposed that the dimensionality reduction algorithm based on random forest (RF) and principal component analysis (PCA) combined with extreme learning machine (ELM) significantly reduced the time required for breast cancer diagnosis, showing excellent classification performance.

The rough set theory is a novel mathematical tool for dealing with uncertain, fuzzy, and inconsistent data proposed by Professor Pawlak in 1982 [20-22]. The rough set provides an effective method for multi-source heterogeneous information classification without prior knowledge [23]. The rough set theory can find the dependency between data, and reduce the number of attributes of the data set. So the rough set is widely used in dimension reduction [24]. But the classical rough set is only suitable for discrete data. So the classical rough set has been popularized from many aspects [25]. Suo M [26] proposed a Single-parameter decision-theoretic rough set (SPDTRS), which can determine the loss function matrix by setting a single parameter and improves engineering practicability. The SPDTRS used a large number of data to prove the reliability of the model. The PNN is a feedforward neural network [27-28], which is

essentially a supervised network classifier based on the Bayesian minimum risk criterion. It has a simple structure and PNN commonly used in classification and identification [29-30]. Wang X X used PNN to intelligently classify photovoltaic array faults, achieving high classification accuracy [31].

The main work of this paper is to use SPDTRS to reduce the dimension of breast cancer data, divide the reduced data as the input of PNN, divide the sample training set and test set. The training set is used for modeling PNN, and the test set is used to test the model. Then we can build a breast cancer classification model based on SPDTRS-PNN and optimize it. Finally, the conclusion and prospects summarize at the end of the article.

Methods

Main contents

The main framework of this paper is in Figure 1 The main idea is to use the five-dimensional reduction algorithm and three machine learning algorithms to establish fifteen different models for distinguishing breast cancer data. We selected the better model from 15 models, the dimensionality reduction algorithm is adopted to reduce data redundancy, and the classification algorithm is adopted to classify breast cancer data. Then the selected model is optimized to achieve

fast and accurate identification of breast cancer data.

Figure 1 Main frame diagram

Dimension reduction algorithm and Classification algorithm

This paper adopts the neighborhood rough set (NRS) [32], the single-parameter decision-theoretic rough set (SPDTRS) [26], the stochastic neighborhood embedding (SNE) [33], the stochastic proximity embedding (SPE) [34] and the neighbor component analysis (NCA) [35]. A suitable dimension reduction algorithm is selected as the input of the classifier.

In this paper, three classification algorithms in machine learning, the probability neural network (PNN) [28], the learning vector quantization (LVQ) [36], and the backpropagation (BP) [37], are used to select the more appropriate classification algorithm.

Data description

This article adopts the breast cancer dataset created by Dr. William h Wahlberg in the UCI dataset. There are 569 instances in this dataset, of which 357 benign cases, 212 malignant cases.

There are 32 attributes in total, including ID number and diagnostic sample label. Therefore, there are 30 attributes representing data features in each instance, including the average value, standard deviation, and maximum value of 10 quantitative features in each nucleus in the sample organization, 1-10 attributes representing the average value of 10 quantitative features, and attributes 11-20 represent standard deviation of 10 quantitative features, the 21-30 attributes represent the maximum value of 10 quantitative features. Among them, the 10 quantitative

features are radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

Selection of training set and test set

569 cases of breast cancer dataset were randomly divided into a training set and test set. 400 cases were selected as a training set, and the remaining 169 cases were taken as test sets. The computer processor used in the experiment was Intel core i3-4005U, 4GB memory, Win7 system, and was simulated under Matlab R2014b version.

SPDTRS algorithm

Based on the theory of classical rough set, the probabilistic rough set proposes two threshold parameters α and β , in which $0 < \beta < \alpha < 1$. Decision theory rough set combines probability rough set with Bayesian minimum risk to give state set $\Omega = \{X, X^c\}$ and action set $A = \{a_p, a_B, a_N\}$, in which, X represents the entity set satisfying condition C , X^c represents the entity set that does not meet condition C , and C is the conditional attribute set, a_p, a_B, a_N represent three behaviors, namely, accepting events, delaying decision-making, and rejecting something, then, the loss function matrix can be described. When X meets condition C , the losses when making a_p, a_B, a_N actions are recorded as $\lambda_{pp}, \lambda_{BP}, \lambda_{NP}$, similarly, when X does not meet condition C , the losses when making a_p, a_B, a_N action are recorded as $\lambda_{pN}, \lambda_{BN}, \lambda_{NN}$, but six loss

functions need to be set artificially. On this basis, SPDTRS [26] sets a compensation coefficient ξ to replace the six loss functions.

Given an information system $IS = (U, A)$, A is the attribute set, $A = C \cup D$, D is the decision attribute set, when $C \cup D \neq \emptyset, D \neq \emptyset$, the decision system $DS = (U, C \cup D)$ is defined. The inherent category represented by D is defined as the nominal decision class, represented by N , SPDTRS gives a loss function matrix according to the property that significance represents the importance of local equivalence classes in their relevant global statistical distribution, as shown in Table 1, among them, to simplify the research, let $\lambda_{pp} = 0, \lambda_{NN} = 0$, $S(X|[x])$ represents the significance of the N -labeled sample of X in $[x]$ to X , $S^c(X|[x])$ represents the significance of the N -labeled sample of X^c in $[x]$ to X , and $S(X|[x])$ and $S^c(X|[x])$ can be obtained from the distribution information of original data, so the loss function matrix is only related to the compensation coefficient ξ .

The overall risk is defined according to Bayesian risk decision $\mathfrak{R}_B$ [38]:

$$\mathfrak{R}_B = \sum_{x \in POS^*} \left(1 - P(X|[x]_B^\delta) \right) \cdot \lambda_{pN} + \sum_{x \in BND^*} \left(P(X|[x]_B^\delta) \cdot \lambda_{BP} + \left(1 - P(X|[x]_B^\delta) \right) \cdot \lambda_{BN} \right) + \sum_{x \in NEG^*} P(X|[x]_B^\delta) \cdot \lambda_{NP} \quad (1)$$

When $B(B \subseteq C)$ satisfies the following two conditions, subset B is considered to be an

attribute reduction with lower risk than C . the conditions are as follows:

$$1) \mathfrak{R}_B < \mathfrak{R}_C ;$$

$$2) \forall A \subset B, \exists \mathfrak{R}_A > \mathfrak{R}_B .$$

Table 1 Loss function matrix

	X	X^C
a_P	$\lambda_{PP} = 0$	$\lambda_{PN} = S^C(X [x])$
a_B	$\lambda_{BP} = S(X [x])(P(X [x]) - \xi)$	$\lambda_{BN} = S^C(X [x])(1 - P(X [x]) - \xi)$
a_N	$\lambda_{NP} = S(X [x])$	$\lambda_{NN} = 0$

PNN algorithm

A Probabilistic neural network (PNN) is a feedforward neural network extended from the nonparametric probability density estimation method based on Bayes classification rules and kernel density estimation [38]. PNN can use linear learning algorithms instead of nonlinear learning algorithms , and can meet the requirements of real-time processing in training.

PNN is a four-layer feedforward neural network, as shown in Figure 2. PNN includes the input layer, model layer, summation layer, and output layer. The input layer inputs the value of training samples and is responsible for transmitting feature vectors to the network; the model layer and input layer connection through the connection weight. Generally, the neuron

activation function of the model layer is Gaussian function, which is used to calculate the similarity between the input eigenvector and each mode in the training set, and send its distance to the Gaussian function to obtain the output of the model layer; the summation layer is responsible for connecting the pattern units of each class. Each class has only one summation unit. The summation unit only adds the pattern units belonging to its class and has no connection with the pattern units of other classes; the output layer is responsible for outputting the highest score of the summation layer, and the number of output neurons is equal to the number of sample categories.

The basic structure diagram of PNN is given below:

Figure 2 PNN structure diagram

Input layer: input the n -dimensional samples $x = [x_1, x_2, \dots, x_n]^T$ to be classified into the network;

Model layer: receive input data x , output of $\psi_{ij}(x)$ of j^{th} corresponding to class i i^{th} :

$$\psi_{ij}(x) = \frac{1}{(2\pi)^{b/2} \delta^b} \exp \left[-\frac{(x - x_{ij})^T (x - x_{ij})}{2\delta^2} \right] \quad (2)$$

where, $j = 1, 2, 3L$, $c_i, i = 1, 2, 3L$, n , where c_i is the number of class i training samples; n is the total number of training samples; b is the dimension of each sample; x_{ij} is the j center vector of class i of the model layer; δ is the smoothing factor, which plays an important role in classification;

Summation layer: add the mode units of the same kind and calculate their average value $g_{i,c_i}(x)$:

$$g_{i,c_i}(x) = \frac{1}{c_i} \sum_{j=1}^{c_i} \psi_{ij}(x) \quad (3)$$

Output layer: the n outputs obtained by the summation layer are multiplied by the a priori probability $p_i = \frac{c_i}{n}$ of each category, and the maximum output value is the predicted test sample label value $\mu(x)$,

$$\mu(x) = \arg \max [p_i g_{i,c_i}(x)] \quad (4)$$

Discussion

Model selection

To obtain a better classification and recognition effect on the premise of using no more than half of the attributes (≤ 15), this paper tries five dimensionality reduction methods.

For SPDTRS, the compensation coefficient ξ represents the tolerance of the decision-maker to uncertainty. The smaller ξ , the greater the amount of acceptable uncertainty. When the conservatism of the object is unknown, the value range of the compensation coefficient ξ should be $(0, 0.4]$. Therefore, we select the value of ξ every 0.05, combined with PNN, LVQ, and BP classifiers to construct three different models of SPDTRS-PNN, SPDTRS-LVQ, and SPDTRS-BP. We compare the accuracy and test time of the training set and test set of the three models under different ξ values. Taking the accuracy as the main evaluation index and combined with the test time, we select a better value of ξ .

The training results are in Table 2. We can see that the accuracy of the training set and test

set of the three models under different ξ values are more than 85.00%, which can be seen from the data in the table when the ξ value of the SPDTRS-PNN model is 0.25, the accuracy of the training set is 97.00%, the accuracy of the test set is 97.04%, the test time is 0.12s, and the training effect is good. When the ξ value is 0.40, although the accuracy of the training set is 99.75%, the accuracy of the test set is only 92.90%, the accuracy is low, and the test time is 0.33s, and the time is long, Therefore, when constructing the SPDTRS-PNN model, the value of ξ is 0.25.

When the ξ value of the SPDTRS-BP model is 0.30, the accuracy of the training set is 97.04%, the accuracy of the test set is 96.16%, and the test time is 0.87s. The training effect is good. When the ξ value is 0.05, although the accuracy of the training set is 98.25%, the

accuracy of the test set is 95.55%, and the test time is 0.91s.

When the ξ value of the SPDTRS-LVQ model is 0.40, the accuracy of the training set is 92.00%, the accuracy of the test set is 91.72%, and the test time is 0.10s. The training effect is good. When the ξ value is 0.25, although the accuracy of the test set is 94.08%, the accuracy of the training set is only 90.00%, the accuracy is low, and the test time is 0.10s.

We can see that the SPDTRS-PNN model has a better performance by comprehensively comparing the accuracy and test time of the three models, and the number of attributes sent to the classifier after model reduction is 12 (<15), which meets the expectation of this paper. Among them, when the value of ξ is 0.25, the 12 attributes selected by SPDTRS-PNN model are {21,27,22,29,20,25,2,18,5,8,11,6}.

Table2 Comparison of accuracy and time of different ξ values

SPDTRS	Dimension	PNN			LVQ			BP		
		Training set (%)	Test set (%)	Time (s)	Training set (%)	Test set (%)	Time (s)	Training set (%)	Test set (%)	Time (s)
0.05	12	97.50	94.67	0.15	90.00	88.76	0.17	98.25	95.55	0.91
0.10	12	98.50	94.08	0.14	90.00	86.98	0.12	90.15	90.07	0.86
0.15	12	97.45	93.49	0.14	90.00	88.17	0.10	95.21	95.23	0.83
0.20	12	97.25	94.67	0.13	91.00	91.02	0.10	94.71	96.62	0.85
0.25	12	97.00	97.04	0.12	90.00	94.08	0.10	91.27	96.98	0.94
0.30	12	98.75	94.08	0.14	87.25	86.39	0.10	97.04	96.16	0.87
0.35	12	99.00	92.90	0.15	87.50	86.98	0.10	91.41	90.21	0.85
0.40	14	99.75	92.90	0.33	92.00	91.72	0.10	95.21	96.11	0.89

For NRS, fifteen attributes are selected according to their importance. As shown in Figure 3, the top fifteen attributes selected according to their importance ranking are {21,27,22,29,20,25,2,18,5,8,11,6}.

For NRS, this paper uses the NRS algorithm to reduce breast cancer data to 1-15 dimensions and constructs three models of NRS-PNN, NRS-LVQ, and NRS-BP. We can obtain the comparison diagrams of the accuracy of the training set and the test set and test time of the three models in different dimensions.

Figure 4 (a) and Figure 4 (b) show the comparison of the accuracy and test time of the three models under different dimensions. It can be seen from the figure that the accuracy of the BP training set and test set is higher than LVQ and PNN, but the test time of BP is longer, and there is no difference in the test time of the other two models. Through comprehensive comparison, we can see that the NRS-BP model is better. The original data is reduced to 14 dimensions with NRS. After recognition with BP, the accuracy of the NRS-BP training set is 95.16% and the test set is 93.79%, the test time is 0.96s.

Figure 3 Comparison chart of absolute value of importance

(a)Accuracy comparison chart

(b) Comparison diagram of test time

Figure 4 Comparison diagram of NRS-PNN, NRS-LVQ and NRS-BP models

For SNE, this paper uses the SNE algorithm to reduce breast cancer data to 1-15 dimensions and constructs three models of SNE-PNN, SNE-LVQ, and SNE-BP. We can obtain the comparison diagrams of the accuracy of the training set and the test set and test time of the three models in different dimensions.

Figure 5 (a) and Figure 5 (b) show the comparison of the accuracy and test time of the three models under different dimensions. It can be seen from the figure that although the accuracy

of PNN's training set is higher, the accuracy of PNN's test set is lower, and there may be an overfitting phenomenon, on the whole, the training effect of LVQ is good. For PNN, the overall training effect is poor. Through comprehensive comparison, we can see that the SNE-LVQ model is better. After reducing the dimension of the original data to 15 dimensions by SNE, combined with LVQ for identification. At this time, the accuracy of the training set is 90.00%, the accuracy of the test set is 92.31%, and the test time is 0.15s.

(a) Accuracy comparison chart

(b) Comparison diagram of test time

Figure 5 Comparison diagram of SNE-PNN, SNE-LVQ and SNE-BP models

For SPE, this paper uses the SPE algorithm to reduce breast cancer data to 1-15 dimensions and constructs three models of SPE-PNN, SPE-LVQ, and SPE-BP. We can obtain the comparison diagrams of the accuracy of the

training set and the test set and test time of the three models in different dimensions.

Figure 6 (a) and Figure 6 (b) show the comparison of the accuracy and test time of the three models under different dimensions. It can

be seen from the figure that the accuracy of the PNN training set is lower as a whole; the test time of BP is long. Through comprehensive comparison, it can be seen that the SPE-LVQ model is better. After reducing the dimension of

the original data to 14 dimensions by SPE and combining LVQ for identification, the accuracy of the training set is 91.00%, the accuracy of the test set is 91.13%, and the test time is 0.099s.

(a) Accuracy comparison chart

(b) Comparison diagram of test time

Figure 6 Comparison diagram of SPE-PNN, SPE-LVQ and SPE-BP models

For NCA, this paper uses the NCA algorithm to reduce breast cancer data to 1-15 dimensions and constructs three models of NCA-PNN, NCA-LVQ, and NCA-BP. We can obtain the comparison diagrams of the accuracy of the training set and the test set and test time of the three models in different dimensions.

used for identification. At this time, the accuracy of the training set is 100.00%, the accuracy of the test set is 95.86%, and the test time is 0.12s.

Figure 7 (a) and Figure 7 (b) show the comparison of the accuracy and test time of the three models under different dimensions. It can be seen from the figure that the accuracy of PNN is higher. On the whole, the NCA-PNN model is better. After reducing the dimension of the original data to 11 dimensions by NCA, PNN is

This paper compares the above-selected model with the accuracy and test time of putting the original data directly into the classifier, as shown in Table 3. We can see that the training effect of putting the original data into BP is the best from the table. The accuracy of the training set is 97.38% and the accuracy of the test set is 95.46%, but compared with other models, the test time is longer, the model constructed by the dimension reduction algorithm and the classification algorithm can achieve higher

recognition accuracy with fewer attributes and shorten the testing time. The SPDTRS-PNN model is more accurate and the test time is 0.12s.

Therefore, the SPDTRS-PNN model is used to identify and analyze breast cancer data.

(a) Accuracy comparison chart

(b) Comparison diagram of test time

Figure 7 Comparison diagram of NCA-PNN, NCA -LVQ and NCA -BP models

Table 3 Comparison of different models

Model	Dimension	Training set (%)	Test set (%)	Time(s)
PNN	30	100	72.78	0.17
LVQ	30	89.75	92.9	0.16
BP	30	97.38	95.46	1.6
SPDTRS-PNN	12	97	97.04	0.12
NRS-BP	14	95.16	93.79	0.96
SNE-LVQ	15	90	92.31	0.15
SPE-LVQ	14	91	91.13	0.099
NCA-PNN	11	100	95.86	0.12

Optimization of SPDTRS-PNN model

In this paper, we used the SPDTRS algorithm to reduced breast cancer data, and we set the SPDTRS compensation coefficient ξ as 0.25. After obtaining dimension reduction, the twelve attributes are {21,27,22,29,20,25,2,18,5,8,

11,6}, and the dimensionality reduction data are taken as the input of PNN. To achieve better diagnosis and discrimination of breast cancer data, we further optimize the SPDTRS-PNN model.

For PNN, the distribution density SPREAD is the expansion coefficient of the radial basis

function. Reasonably selecting the value of SPREAD is also a significant step in classification. When the value of SPREAD is close to 0, it can form the nearest neighbor classifier. When the SPREAD value is more, the output result will become smooth and can form a proximity classifier for several training samples, however, too large a SPREAD value will make numerical calculation difficult. Therefore, in this paper, the value range of SPREAD is set as $(0.5, 1.5]$, the step size is 0.05, the accuracy is the main evaluation index, and in combination with test time to select the appropriate SPREAD parameter value.

The accuracy of the training and the test set and test time obtained under different SPREAD values are in Figure 8. The left axis is the accuracy, the right axis is the test time, the blue

solid line represents the training set accuracy, the green solid line represents the test set accuracy, and the red dotted line represents the test time. It can be seen from the figure that with the increase of SPREAD value. The accuracy of the training set of the SPDTRS-PNN model shows a downward trend as a whole. The accuracy of the test set is stable at 95.00% - 97.50%, and the test time fluctuates between 0.09s-0.12s. It can be seen from the test time curve that when the SPREAD value is 1.55, although the shortest test time is 0.09s, the accuracy of the training set is only 96.50%, and the accuracy of the training set is low. When the SPREAD value is 0.75, the accuracy of the training set and the test set are 99.25% and 97.04%, and the test time is only 0.003s later than the test time when the SPREAD value is 1.25. Therefore, this paper sets the SPREAD value to 0.75.

Figure 8 Comparison of accuracy and time of different SPREAD values

In this paper, we will use the 12 attributes of SPDTRS to reduce the dimension of breast cancer data as the input feature vector of the PNN model.

The output eigenvector of PNN is the sample label, which is benign and malignant breast cancer. We set the SPREAD parameter of PNN

as 0.75. A model combining SPDTRS-PNN is used to identify the breast cancer data.

The 400 samples of the training set are input into PNN, and the training effect is shown in Figure 9. The red asterisk is the output value of the network prediction, and the blue circle is the actual output value of the network. From the graph, 3 malignant tumors in the breast cancer data can be mistakenly divided into benign tumors. In general, the training set has 3 errors in the network prediction, and the accuracy rate of the training set is 99.25%.

The 169 samples of the test set were input into PNN, and the training effect is shown in Figure 10. The red star is the output value of the network prediction, and the blue circle is the actual output value of the network. From the chart, we can see that 4 of the breast cancer data have been mistaken for benign tumors and become benign, and 1 actually benign tumor is wrongly predicted for malignant tumors. The network prediction has five errors, and the accuracy of the test set is 97.04%.

Figure 9 Training set classification rendering

Figure 10 Effect drawing of test set classification

Result

The results show that: (1) The SPDTRS can express more comprehensive information in the original breast cancer data with fewer features by selecting the appropriate value. It can reduce the complexity of the model and improves the efficiency of the model. (2) The appropriate SPREAD value was selected, then using the PNN model for classification, the test time of SPDTRS-PNN is short and fast. (3) The SPDTRS-PNN model is suitable for breast cancer diagnosis. It can objectively distinguish breast cancer from benign and malignant samples and meet the needs of a rapid and accurate breast cancer diagnosis. Although this research has made some achievements, there are still some shortcomings. To a certain extent, the SPDTRS-PNN model can reduce the redundant information of breast cancer and improve prediction accuracy. But parameters in the algorithm need artificial settings, that is, manual optimization. In future work, we need to add some automatic parameter optimization algorithms to improve the model performance and make the model performance closer to our ideal state.

Conclusion

In this paper, we combed a dimension reduction algorithm with a machine learning algorithm. Then we proposed a new auxiliary medical diagnosis method, that is, SPDTRS-PNN. The SPDTRS is used to reduce the quantitative

characteristic data of breast tumor images to 12 dimensions. Then the PNN model was set up to test the predictive effect of breast cancer. We have demonstrated that the rapid and accurate diagnosis of breast cancer can be achieved by using the attributes of fewer breast cancer data.

Abbreviations

SPDTRS: single-parameter decision-theoretic rough set; PNN: probability neural network; NRS: neighborhood rough set; SNE: stochastic neighborhood embedding; SPE: stochastic proximity embedding; NCA: neighbor component analysis; LVQ: learning vector quantization; BP: back propagation.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets analysed during the current study are available in the UCI repository, [<http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>]

Competing interests

The authors declare that they have no competing interests

Funding

We'd like to thank for major science and technology program of Anhui province (No. 201903a07020013), the New Generation of Information Technology Innovation Project(No.2019ITA01010), Demonstration project of science popularization innovation and scientific research education for College Students (No. KYX202117), University-level Key Projects of Anhui University of Science and Technology (No. xjzd2020-06), and the National Key Research and Development Program of China (No. 2018YFC0604503), These funding bodies did not play any role in the design of study, the interpretation of data, or the writing of this manuscript.

Author contributions

XXK conceived the study. MRZ developed the method and supervised the study. XXK, KB, and WHL implemented the algorithms. XXK, FH, RYD, JJY, and OH analyzed the data. KB wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank the Departments of Computer Science, Surgery, and Human Oncology at the University of Wisconsin, Madison, United States to make the database available.

Authors' information

ZMR, Professor, School of electrical and information engineering, Anhui University of Science and Technology, State Key Laboratory of Mining Response and Disaster Prevention and Control in Deep Coal Mines, Anhui University of Science and Technology. KXX, doctor, School of electrical and information engineering, Anhui University of Science and Technology. BK, doctor, School of electrical and information engineering, Anhui University of Science and Technology. LWH, doctor, School of electrical and information engineering, Anhui University of Science and Technology. HF, doctor, School of electrical and information engineering, Anhui University of Science and Technology. DRY, master, School of electrical and information engineering, Anhui University of Science and Technology. YJJ, master, School of electrical and information engineering, Anhui University of Science and Technology. Oumaima Hamrani, doctor, School of electrical and information engineering, Anhui University of Science and Technology.

Reference

- [1] Marchetti, P., Antonov, A., Anemona, L. et al. New immunological potential markers for triple negative breast cancer: IL18R1, CD53, TRIM, Jaw1, LTb, PTPRCAP. *Discov Onc* 12, 6 (2021). <https://doi.org/10.1007/s12672-021-00401-0>.
- [2] Dina A. Ragab, Omneya Attallah, Maha Sharkas, Jinchang Ren, Stephen Marshall, A framework for

breast cancer classification using Multi-DCNNs, Computers in Biology and Medicine, Volume 131, 2021, 104245, ISSN0010-4825, <https://doi.org/10.1016/j.combiomed.2021.104245>.

[3] M. R. Basunia, I. A. Pervin, M. Al Mahmud, S. Saha and M. Arifuzzaman, "On Predicting and Analyzing Breast Cancer using Data Mining Approach," 2020 IEEE Region 10 Symposium (TENSYP), 2020, pp. 1257-1260, doi: 10.1109/TENSYP50017.2020.9230871.

[4] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018 IEEMA Engineer Infinite Conference (eTechNxT), 2018, pp. 1-5, doi: 10.1109/ETECHNXT.2018.8385355.

[5] K. S. Bhangu, J. K. Sandhu and L. Sapra, "Improving diagnostic accuracy for breast cancer using prediction-based approaches," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2020, pp. 438-441, doi: 10.1109/PDGC50313.2020.9315815.

[6] L. H. A. al-sammarraie and A. A. Ibrahim, "Predicting Breast Cancer in Fine Needle Aspiration Images Using Machine Learning," 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2020, pp. 1-4, doi: 10.1109/ISMSIT50672.2020.9254891.

[7] K. S. Bhangu, J. K. Sandhu and L. Sapra, "Improving diagnostic accuracy for breast cancer

using prediction-based approaches," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2020, pp. 438-441, doi: 10.1109/PDGC50313.2020.9315815.

[8] F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," 2013 13th International Conference on Intelligent Systems Design and Applications, 2013, pp. 121-125, doi: 10.1109/ISDA.2013.6920720.

[9] Dennison, G., Anand, R., Makar, S.H. and Pain, J.A. (2003), A Prospective Study of the Use of Fine-Needle Aspiration Cytology and Core Biopsy in the Diagnosis of Breast Cancer. The Breast Journal, 9: 491-493.<https://doi.org/10.1046/j.1524-4741.2003.09611.x>

[10] William H. Wolberg, W.Nick Street, O.L. Mangasarian, Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates, Cancer Letters, Volume 77, Issues 2-3,1994,Pages 163-171,ISSN 0304-3835,[https://doi.org/10.1016/0304-3835\(94\)90099-X](https://doi.org/10.1016/0304-3835(94)90099-X).

[11] T. Thomas, N. Pradhan and V. S. Dhaka, "Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey," 2020 International Conference on Inventive Computation Technologies (ICICT), 2020, pp. 192-196, doi: 10.1109/ICICT48043.2020.9112464.

- [12] Hayashi Y. Does Deep Learning Work Well for Categorical Datasets with Mainly Nominal Attributes? *Electronics*. 2020;9(11):1966. <https://doi.org/10.3390/electronics9111966>
- [13] A. H. Al-Timemy, F. M. Al-Naima and N. H. Qaeb, "Probabilistic Neural Network for Breast Biopsy Classification," 2009 Second International Conference on Developments in eSystems Engineering, 2009, pp. 101-106, doi: 10.1109/DeSE.2009.31.
- [14] H. M. Whitney, H. Li, Y. Ji, P. Liu and M. L. Giger, "Comparison of Breast MRI Tumor Classification Using Human-Engineered Radiomics, Transfer Learning From Deep Convolutional Neural Networks, and Fusion Methods," in *Proceedings of the IEEE*, vol. 108, no. 1, pp. 163-177, Jan. 2020, doi: 10.1109/JPROC.2019.2950187.
- [15] R. Nagpure, S. Chandak and N. Pathak, "Breast Cancer Detection using Neural Network Mammogram," 2020 International Conference on Convergence to Digital World - Quo Vadis (ICCDW), 2020, pp. 1-6, doi: 10.1109/ICCDW45521.2020.9318635.
- [16] Street W N, Wolberg W H, Mangasarian O L. Nuclear feature extraction for breast tumor diagnosis[J]. *Proc Spie*, 2012, 1993:861-870.
- [17] J. Zhou, T. Zhong and X. He, "Auxiliary Diagnosis of Breast Tumor Based on PNN Classifier Optimized by PCA and PSO Algorithm," 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2017, pp. 222-227, doi: 10.1109/IHMSC.2017.164.
- [18] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018 IEEMA Engineer Infinite Conference (eTechNxT), 2018, pp. 1-5, doi: 10.1109/ETECHNXT.2018.8385355.
- [19] Bian Kai, Zhou Mengran, Hu Feng, Lai Wenhao, RF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute Selection and Feature Extraction, *Frontiers in Genetics*, 11, 2020, 1082, doi:10.3389/fgene.2020.566057.
- [20] Qun-Feng Zhang, Su-Yun Zhao and Yun-Chao Bai, "On the application of rough sets to data mining in economic practice," 2009 International Conference on Machine Learning and Cybernetics, 2009, pp. 272-276, doi: 10.1109/ICMLC.2009.5212452.
- [21] Roman W. Swiniarski, Andrzej Skowron, Rough set methods in feature selection and recognition, *Pattern Recognition Letters*, Volume 24, Issue 6, 2003, Pages 833-849, ISSN 0167-8655, [https://doi.org/10.1016/S0167-8655\(02\)00196-4](https://doi.org/10.1016/S0167-8655(02)00196-4).
- [22] Chen Y, Chen Y. Feature Subset Selection Based on Variable Precision Neighborhood Rough Sets[J]. *International Journal of Computational Intelligence Systems*, 2021, 14(1). <https://doi.org/10.2991/ijcis.d.210106.003>

[23] Xiaoli Chu, Bingzhen Sun, Xue Li, Keyu Han, JiaQi Wu, Yan Zhang, Qingchun Huang, Neighborhood rough set-based three-way clustering considering attribute correlations: An approach to classification of potential gout groups, *Information Sciences*, Volume 535, 2020, Pages 28-41, ISSN 0020-0255, <https://doi.org/10.1016/j.ins.2020.05.039>.

[24] Yao Ping, Lu Yongheng, Neighborhood rough set and SVM based hybrid credit scoring classifier, *Expert Systems with Applications*, Volume 38, Issue 9, 2011, Pages 11300-11304, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2011.02.179>.

[25] Xiaodong Fan, Weida Zhao, Changzhong Wang, Yang Huang, Attribute reduction based on max-decision neighborhood rough set model, *Knowledge-Based Systems*, Volume 151, 2018, Pages 16-23, ISSN 0950-7051, <https://doi.org/10.1016/j.knosys.2018.03.015>.

[26] Mingliang Suo, Laifa Tao, Baolong Zhu, Xuewen Miao, Zhichao Liang, Yu Ding, Xingliu Zhang, Tong Zhang, Single-parameter decision-theoretic rough set, *Information Sciences*, Volume 539, 2020, Pages 49-80, ISSN 0020-0255, <https://doi.org/10.1016/j.ins.2020.05.124>.

[27] S. Naaz and S. Parveen, "A PNN based Malign Attack Detection and Classification Model," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 933-938, doi: 10.1109/ICOSEC49089.2020.9215424.

[28] M. Kusy, "Selection of Pattern Neurons for a Probabilistic Neural Network by Means of Clustering and Nearest Neighbor Techniques," 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), 2019, pp. 598-603, doi:10.1109/CoDIT.2019.8820385.

[29] S. Naaz and S. Parveen, "A PNN based Malign Attack Detection and Classification Model," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 933-938, doi: 10.1109/ICOSEC49089.2020.9215424.

[30] Jun Guo, Chen-xu Gu, Jun-jie Yang, Yin Zhang, Heng Yang, Data mining and application of ship impact spectrum acceleration based on PNN neural network, *Ocean Engineering*, Volume 203, 2020, 107193, ISSN 0029-8018, <https://doi.org/10.1016/j.oceaneng.2020.107193>.

[31] X. x. Wang, L. Dong, S. y. Liu, Y. Hao and B. Wang, "A Fault Classification Method of Photovoltaic Array Based on Probabilistic Neural Network," 2019 Chinese Control And Decision Conference (CCDC), 2019, pp. 5260-5265, doi: 10.1109/CCDC.2019.8832338.

[32] Y. Han, X. Wu, J. Wu, R. Jia, B. Zhang and X. Yao, "A New Algorithm for Knowledge Reduction Based on Neighborhood Rough Set," 2010 International Conference on Artificial Intelligence and Computational Intelligence, 2010, pp. 15-18, doi: 10.1109/AICI.2010.10.

[33] Bunte K, Haase S, Biehl M, et al. Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences[J]. *Neurocomputing*, 2012, 90(8):23-45.

Kerstin Bunte, Sven Haase, Michael Biehl, Thomas Villmann, Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences, *Neurocomputing*, Volume 90, 2012, Pages 23-45, ISSN 0925-2312, <https://doi.org/10.1016/j.neucom.2012.02.034>.

[34] Dmitrii N Rassokhin, Dimitris K Agrafiotis, A modified update rule for stochastic proximity embedding, *Journal of Molecular Graphics and Modelling*, Volume 22, Issue 2, 2003, Pages 133-140, ISSN 1093-3263, [https://doi.org/10.1016/S1093-3263\(03\)00155-4](https://doi.org/10.1016/S1093-3263(03)00155-4).

[35] Liu, C., X. Li, and Y. Yang. "Text Classification Algorithm Based on Neighborhood Component Analysis." *Computer Engineering* 38.15(2012):139-141. doi: 10.3969/j.issn.1000-3428.2012.15.038

[36] A. Sumarsono and S. Supatman, "IMAGERY IDENTIFICATION OF TOMATOES WHICH

CONTAIN PESTICIDES USING LEARNING VECTOR QUANTIZATION", *J. Tek. Inform. (JUTIF)*, vol. 2, no. 1, pp. 9-16, Jan. 2021. <https://doi.org/10.20884/1.jutif.2021.2.1.15>.

[37] Murty E M. PREDIKSI PENGADAAN DAN PENGELOLAAN INVENTORI JARINGAN SYARAF TIRUAN ALGORITMA BACKPROPAGATION PADA PERUM BULOG[J]. *KOMPUTEK*, 2021, 5(1):1. doi:10.24269/jkt.v5i1.637

[38] Xiuyi Jia, Wenhe Liao, Zhenmin Tang, Lin Shang, Minimum cost attribute reduction in decision-theoretic rough set models, *Information Sciences*, Volume 219, 2013, Pages 151-167, ISSN 0020-0255, <https://doi.org/10.1016/j.ins.2012.07.010>.

[39] L. Shi-ya, C. Zhao-hua, L. Jing and Z. Zhi-fei, "A Medical Diagnosis Model Based on Pnn-Cadaboost Algorithm," 2018 11th International Conference on Intelligent Computation Technology and Automation (ICICTA), 2018, pp. 1-4, doi: 10.1109/ICICTA.2018.00008.