Breast Cancer Prediction Based on K-Means and SOM Hybrid Algorithm

Haoquan Lin¹ and Zhenzhou Ji²,*

¹School of Computer Science and Technology, Harbin Institute of Technology weihai, Weihai, China
²School of Computer Science and Technology, Harbin Institute of Technology weihai, Weihai, China

*Corresponding author email: jizhenzhou@hit.edu.cn

Abstract. Breast cancer is one of the most serious diseases that threaten women's health, affecting about 12.5% of women worldwide. Early detection of breast cancer is critical to saving lives. Therefore, If the physical examination indicators related to the human body can be extracted and the breast cancer can be analyzed through machine learning, which will play a key role in predicting and preventing breast cancer. As the high complexity and low precision of SOM neural network algorithm and shortcomings of K-means clustering algorithm needs to determine the number of clustering advanced and randomly select initial clustering centers, a hybrid algorithm combining K-means and SOM neural network is proposed in this study. The results show that the hybrid algorithm can accurately cluster the data sets, and compared with K means model and SOM neural network model, the performance of the hybrid algorithm model is better in clustering accuracy and computing speed.

Keywords: Breast cancer prediction; K-Means algorithm; SOM neural network model; Hybrid algorithm.

1. Introduction

Breast cancer is a disease with high incidence rate in many women worldwide. The report, Global Cancer Statistics 2018[1], published in the official journal of the American Cancer Society, assessed that the incidence and death toll of 36 cancers in 185 countries. It is estimated that in 2018, there were 18.1 million new cancer cases in the world, including 2.1 million breast cancer cases, accounting for 11.6%; and 9.6 million new cancer deaths, including 630000 breast cancer cases, accounting for 6.6%. In order to solve the problem of breast cancer and early diagnosis of breast cancer, this study uses clustering data mining technology to study the health status of breast cancer patients. Domestic and foreign experts and scholars have used data mining technology to predict the risk of breast cancer, and have achieved certain achievements. Such as Delen, et al[2] compared logistic regression techniques, ANN, and decision tree for breast cancer prediction analysis. The experiment results show that ANN with 91.2% accuracy, decision tree with 93.6% accuracy, and logistic regression with 89.2% accuracy; Rajesh, et al[3] who used C4.5 classification algorithm to predict breast cancer in SEER dataset, and the accuracy of the algorithm is 92.2%. Peter, et al[4] used j48 decision tree and naive Bayes to predict breast cancer for RASUS breast cancer dataset. The accuracy of naive Bayes classifier is 82.6%, while the accuracy of j48 decision tree is 94.2%.

A hybrid algorithm based on K-means algorithm and SOM neural network is presented in this paper. The advantages of K-means algorithm are simple, easy to understand, convenient to calculate and fast[5]. SOM neural network can represent all points in high-dimensional space with points in low-
dimensional target space to enhance useful information\cite{6}. Therefore, combining the advantages of these two algorithms, this paper proposes a hybrid algorithm based on the combination of K-means and SOM neural network for breast cancer prediction, and compares the hybrid algorithm model with the traditional SOM neural network model and K-means algorithm model respectively. The experimental results show that the hybrid algorithm model can predict breast cancer more accurately and provide effective help for medical recommendation and other applications.

2. Methodology

2.1. K-Means Algorithm

K-means clustering is a simple and fast method used commonly because of its straightforward implementation and small number of iterations\cite{7}. The basic calculation is as follows:

(1) The number of clusters is set as \( k \), and \( k \) objects are randomly selected from \( n \) data objects as the initial cluster center \( \{w_1, w_2, \ldots, w_k\} \). Where \( w_j = i_j, j \in \{1,2,\ldots,k\}, l \in \{1,2,\ldots,n\} \).

(2) Each cluster \( C_j \) is made corresponding to cluster center \( w_j \).

(3) \( i_j \), where \( l \in \{1,2,\ldots,n\} \), is assign to the nearest cluster center \( w_j^* \) belongs to the cluster \( C_j^* \). That is:

\[
|i_j - w_j^*| \leq |i_j - w_j| \quad j \in \{1,2,\ldots,k\}
\]

(1)

(4) Cluster center \( C_j \), where \( j \in \{1,2,\ldots,k\} \), is updated to the center point of all current samples. That is:

\[
 w_j = \frac{\sum_{i \in C_j} i_j}{|C_j|}
\]

(2)

(5) According to equation (3), it can be judged whether the clustering is convergent or not.

\[
E = \sum_{j=1}^{k} \sum_{i \in C_j} |i_j - w_j|^2
\]

(3)

Steps (3) to (5) are repeated until the sum-of-squared-error criterion function \( E \) reaches the qualification, indicating that the algorithm tends to stabilize and the algorithm is terminated.

2.2. SOM Neural Network

Self-organizing mapping neural network (Self-organizing Map, SOM) is a clustering algorithm. It is a biologically reasonable model of artificial neural network, which can convert the input signal of any dimension into a one-dimensional or two-dimensional discrete mapping by calculating the mapping and implementing the process in an adaptive manner. The typical model of the SOM neural network is shown in figure 1.

\[\text{Figure 1. Classic SOM neural network model}\]

SOM neural network clustering algorithm includes competition, cooperation, and updating\cite{8}.

(1) In the process of competition, the neuron with the largest output is chosen as the winning neuron.
Because the excitation function of neuron is a linear function. Therefore, the maximum output of a neuron depends on the inner product of its input vector and weight vector.

(2) In the process of cooperation, it is to determine the strengthening center of the winning neuron. The center of the topology field is the neuron that wins in the competition, while the neurons located in the neighborhood of the winning neuron are called excited neurons.

(3) In the process of updating, the weight vector of the winning neurons and the neurons in the strengthening center are updated through using the Hebb learning rule.

### 2.3. K-Means and SOM Hybrid Algorithm

The K-Means and SOM hybrid algorithm used in this paper is a two-stage calculation method\[9\]. In the first stage, the experimental data is first clustered with SOM clustering algorithm. All data with similar characteristics are classified into the same category, so that the sample data can be grouped into different categories, and the number of categories and the center points of each class can be obtained. In the second stage, the results of the first stage are entered into the K-Means algorithm as input values and further clustered, thus forming the final clustering results. The specific steps of the hybrid algorithm based on K-means and SOM are as follows:

In the first stage, SOM initial clustering is used to get the center of each cluster.

1) Set \( X(n) = [x_1(n), x_2(n), \ldots, x_N(n)]^T \) as the input vector, i.e., the training sample; set \( W_i(n) = [w_{i1}(n), w_{i2}(n), \ldots, w_{iN}(n)]^T \) as the weight vector, where \( i = 1, 2, \ldots, M \); set the total number of iterations as \( N \).

2) Using small random value to initialize the weight vector \( W_i \); set the initial learning rate to \( \eta(0) \); normalize the initial value of the weight vector \( W_i(0) \) and all input vectors.

\[
X' = \frac{X}{\|X\|} \quad (4)
\]

\[
W_i'(0) = \frac{W_i(0)}{\|W_i(0)\|} \quad (5)
\]

where \( \|W_i(0)\| = \sum_{j=1}^{N} |w_{ij}(0)|^2; \|X\| = \sum_{j=1}^{N} (x_j)^2 \), they are the Euclidean norm of weight vector and input vector respectively.

3) For each input vector \( X(n) = [x_1(n), x_2(n), \ldots, x_N(n)]^T \):

1) Select training samples \( X' \) from the input space.

2) According to the equation (6), the winning neuron \( \varepsilon \) is selected to realize the competitive process of the neuron.

\[
\|X - W_i^\varepsilon\| = \min_j \|X - W_i^j\|, j = 1, 2, \ldots, M \quad (6)
\]

3) According to the equation (7), the weight vectors of the excited neurons in the neighbourhood of the winning neuron topology \( N_c(n) \) are updated, thus realizing the process of neuron cooperation and renewal.

\[
W_i^j(n+1) = W_i^j(n) + \eta(n)(X' - W_i^j(n)) \quad (7)
\]

4) According to the equation (8) and the equation (9), the learning rate \( \eta(n) \) and the topology neighborhood are updated respectively, and the weights after learning are re-normalized.

\[
\eta(n) = \eta(0) \left(1 - \frac{n}{N}\right) \quad (8)
\]
(9)

\[ N_c(n) = \text{int} \left[ N_c(0) \left(1 - \frac{n}{N}\right) \right] \]

5) Determine whether the number of iterations \( n \) exceeds \( N \). If \( n \leq N \), jump to step 1). Otherwise, end the iteration process.

6) After the algorithm converges, the initial clustering of experimental data is completed according to the response of output nodes.

(4) Get the output -- the number of clusters \( K \) and the center point of the cluster \( Z = (Z_1, Z_2, \ldots Z_K) \).

In the second stage, the output of the first stage, i.e. the number of clusters \( K \) and the cluster center point \( Z \), are taken as the initial input value of the K-means algorithm.

(5) The output of step (4) is used as the initial input value of the K-means algorithm for iterative calculation until convergence.

(6) Output the clustering results of the K-means and SOM hybrid algorithm.

3. Breast Cancer Prediction Based on K-Means and SOM Hybrid Algorithm

3.1. Experimental Data

This paper proposes a breast cancer prediction model based on K-means and SOM hybrid algorithm. The experimental data is read from the open-source Wisconsin breast cancer data set in the machine learning library of UCI. The data set includes 569 cases (benign 62.74%, malignant 37.26%) with 32 patient attributes, including 1 patient ID number record, 30 tumor diagnosis information and 1 tumor diagnosis result record (benign and malignant). As shown in the table1:

| Attribute | Description | Mean     | Standard error | Largest value |
|-----------|-------------|----------|----------------|---------------|
| 1         | Radius      | 6.98-28.11 | 0.11-2.87      | 7.93-36.04    |
| 2         | Texture     | 0.71-39.28 | 0.36-4.89      | 12.02-49.54   |
| 3         | Perimeter   | 43.79-188.50 | 0.76-21.98    | 50.41-251.20  |
| 4         | Area        | 143.50-2501.00 | 6.80-542.20 | 185.20-4254.00 |
| 5         | Smoothness  | 0.05-0.16  | 0.00-0.03      | 0.07-0.22     |
| 6         | Compactness | 0.02-0.35  | 0.00-0.14      | 0.03-1.06     |
| 7         | Concavity   | 0.00-0.43  | 0.00-0.40      | 0.00-1.25     |
| 8         | Concave points | 0.00-0.20 | 0.00-0.05     | 0.00-0.29     |
| 9         | Symmetry    | 0.11-0.30  | 0.00-0.08      | 0.16-0.66     |
| 10        | Fractal dimension | 0.05-0.10 | 0.00-0.03    | 0.06-0.21     |

3.2. Experimental Process

Step 1: First, the open source Wisconsin Breast Cancer Dataset from the machine learning library of UCI was read. Second, the classification labels in row y are encoded, where Benign tumor was represented by 0, and malignant tumor was represented by 1. Third, checking whether the data is complete and a relatively complete dataset which can be used for training model was obtained.

Step 2: The dataset was divided into independent training set and independent test set randomly, where 80% of breast cancer data samples were allocated to X-train and y-train and the remaining 20% of the sample data were allocated to X-test and y-test.

Step 3: The dataset was standardized, so as to eliminate the adverse effects caused by different scale between indicators.

Step 4: The training set was clustered by K-Means algorithm. Firstly, the range of cluster number \( a \) was determined to b. The optimal number of clusters was solved by the elbow chart. As shown in the figure 2, the optimal number of clusters was two, which was consistent with the actual results. The K-Means and SOM hybrid algorithm was also used c for clustering. The initial parameters of the SOM neural network model were set to: the input layer had 30 input nodes and the output layer had 2 nodes. Theoretically, all samples could be divided into two categories, the initial weight was taken from d
randomly, the training times were 100. The training times of the K-means and SOM hybrid model were changed to 50.

![Figure 2. The elbow chart of K-means algorithm](image)

Step 5: The trained models were saved, and the three models were tested with the test set. Finally, the precision, recall and F1-score of each model were calculated and compared.

4. Experimental Results

4.1. Evaluation Criteria

Generally, the accuracy, precision and recall of the classifier are used to evaluate the performance of the model. Whether the sample predicts the category is positive or negative can be known from the classifier's prediction report. At the same time, the categories of data under the real situation can be known. Therefore, the count values of four basic indicators can be obtained as shown in the table 2.

**Table 2. Confusion matrix**

| Health condition | Positive | Negative |
|------------------|----------|----------|
| True             | TP       | FN       |
| False            | FP       | TN       |

Therefore, it can be obtained that the Precision, Recall and F1 values predicted of the experimental data as evaluation indexes, and the equations are as follows:

\[
\text{Precision} = \frac{TP}{TP + FN} \quad (10)
\]

\[
\text{Recall} = \frac{TP}{TP + FP} \quad (11)
\]

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)
\]

4.2. Experimental Results and Comparison

The experimental results are shown in the table 3. It can be seen from the table that the SOM algorithm model has the longest running time and more iterations, with the F1-score of 92.4%; the K-means algorithm has the fastest running speed, but the F1-score is the lowest of 89.9%; the overall time of K-means and SOM hybrid algorithm model is relatively low, with fewer clustering times, and its F1-score is the highest of the three models.
Table 3. Performance comparison of three algorithm

| Algorithm     | Precision | Recall  | F1-score | Iteration | Running time |
|---------------|-----------|---------|----------|-----------|--------------|
| K-Means       | 0.857     | 0.947   | 0.899    | --        | 0.04         |
| SOM           | 0.888     | 0.965   | 0.924    | 100       | 8.85         |
| K-Means+SOM   | 0.921     | 0.983   | 0.950    | 50        | 4.62         |

Comprehensive evaluation of the above three algorithms: K-Means algorithm model running time is short, but the accuracy is not high; although the accuracy of SOM algorithm model has been improved, but the running time is relatively long. K-means and SOM hybrid algorithm model has better overall performance in accuracy and running speed, and is more suitable for breast cancer prediction.

5. Conclusion and Future Work
In this paper, a hybrid algorithm of K-Means and SOM neural network is proposed for breast cancer prediction in the medical industry. Experimental results show that the K-means and SOM hybrid algorithm inherit the advantages of K-means algorithm and SOM algorithm respectively, and overcome their respective shortcomings. Compared with k-means algorithm, the hybrid algorithm has higher accuracy, while compared with traditional SOM algorithm, it not only improves the accuracy, but also shortens the running time.

Although the accuracy and running time of the hybrid algorithm have been improved, it takes longer than k-means algorithm. How to reduce the running time while maintaining high accuracy of the algorithm is the next work. In addition, the amount of data used in the experiment is not large enough. More experimental dataset are needed to test the algorithm in the future.

Acknowledgments
Supported by the National Natural Science Foundation of China (Grant No. 2019603B001).

References
[1] Bray F et al 2018 Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries 68(6) 394-424
[2] Venkatesan and Perumal 2009 Breast Cancer Survival Prediction using Artificial Neural Network
[3] Wang H F and Yoon S W 2015 Breast Cancer Prediction Using Data Mining Method (Proceedings of the 2015 Industrial and Systems Engineering Research Conference)
[4] Kehinde W et al 2015 Breast cancer risk prediction using data mining classification techniques (Transactions on Networks and Communications) 3. 10.14738/tnc.32.662
[5] Wang Z et al 2009 A k-means algorithm based on optimized initial center points (Pattern Recognition and Artificial Intelligence No.2) 133-138
[6] Liang B M 2009 Study on improvement and application of serf-organizing map neural network (Computer Engineering and Applications(31)) 138-141
[7] Radha R and Rajendiran P 2014 Using K-Means Clustering Technique to Study of Breast Cancer (World Congress on Computing and Communication Technologies)
[8] Gao J 2003 Principles and simulation examples of artificial neural network (China Machine Press) 147-151
[9] Zhou H et al 2010 SOM+K-Means Two-phase Clustering Algorithm and Its Application (Modern Electronics Technique Vol 16) 113-115