Natural Neighbor Algorithm for Breast Cancer Diagnosis

Shuxiang Li, Lijun Yang*, Peiyu Chen and Cheng Yang
College of computer science and technology, Southwest Minzu University, Chengdu, China
985201824@qq.com

Abstract. Machine learning to diagnose breast cancer is an important, real-world medical problem. As one of the most popular machine learning algorithm, K-Nearest Neighbors (KNN) algorithm is widely used in breast cancer classification and can provide high classification accuracy and effective diagnostic capabilities. However, selection of parameter K is still an unsolved problem for K-Nearest Neighbors (KNN). To address the problem, we introduce a novel neighbor form, Natural Neighbor (NaN), which is obtained adaptively by its search algorithm. We firstly use a noise filter, which called edited natural neighbor algorithm (ENaN), to eliminate the noises and global outliers of Wisconsin’s breast cancer prognosis dataset. Then, cleaned data set and natural neighbor algorithm are used to construct a new non parameter diagnostic system. The main advantages of the proposed algorithm are that it does not need any parameters and can maintain high classification accuracy. Experiments show that the classification accuracy of proposed method is similar to the highest classification accuracy obtained by traditional K-Nearest Neighbors algorithm with different K values.

1. Introduction
Breast cancer is a malignant tumor that occurs in the epithelial tissues of the breast gland. The global incidence of breast cancer has been on the rise since the late 1970s. In recent years, the incidence of breast cancer in China has increased by 1 to 2 percentage points higher than that high-incidence countries. The incidence of breast cancer in the national cancer registration area ranks first in female malignant tumors [1]. Therefore, accurate diagnosis of breast cancer and timely treatment are very important for the patients.

KNN algorithm is a classification method that commonly used in machine learning. Most of the studies using KNN algorithm to diagnose breast cancer focus on solving the selection of the parameter K and improving the accuracy of the different datasets. Katsuyoshi Odajima studied the selection of K values for datasets of different sizes [2]. Alberto Palacios Pawlovsky studied the difference between the maximum classification accuracy and the minimum classification accuracy for different datasets and K values [3]. Li Rong classified breast cancer data by Support Vector Machine-K-Nearest Neighbor (SVM-KNN) and evaluated the classification accuracy, and compared the accuracy with obtained by Support Vector Machine (SVM) [4]. Congqin-Yi designed geometric mean KNN (k-nearest neighbor) membership function and L-KMOD kernel function FSVM (fuzzy support vector machine) to identify breast cancer genes [5]. In general, for KNN algorithm, different parameters K have many differences in the accuracy of breast cancer classification.

This paper uses natural neighbor (NaN) algorithm to diagnose breast cancer and solve the parameter problem of KNN algorithm according to the adaptive K value. We also use the cross-validation method to calculate the classification accuracy of NaN algorithm for breast cancer dataset, and
compare it with the KNN algorithm. The classification accuracy obtained by Natural Neighbor (NaN) algorithm basically reaches the highest classification accuracy obtained by KNN algorithm and Natural Neighbor (NaN) does not need any parameters. To summarize, our main contributions are as follows:

- We introduce a new classification method, which does not need any parameters, to diagnose breast cancer.
- The classification accuracy obtained by the new method (Natural Neighbor algorithm) is more stable and higher than K-Nearest Neighbors (KNN).
- Natural Neighbor algorithm is used to clean the data leads to a much higher classification accuracy.

This work is organized as follows. Section 2 introduces the concept of natural neighbor and natural stable structure. Section 3 shows a noisy filter by using Natural Neighbor (NaN) algorithm. Section 4 shows the experiments and analysis about the performance of Natural Neighbor (NaN) algorithm. Finally, Section 6 concludes our work.

2. Natural neighbor algorithm

2.1 Natural neighbor definition

Friendship is a relationship in which each other thinks that the other as a friend. NaN algorithm comes from the process of establishing friendship in real life. Two points establish the natural neighbor relationship only when the K nearest neighbors of the two points contains each other. The natural neighbor relationship between point \( i \) and point \( j \) can be expressed as:

\[
\forall \ x_i \in NN_i(x_j) \iff \ x_i \in KNN(x_j) \land x_j \in KNN(x_i)
\]

where \( NN_i(x_j) \) is the natural neighbor of \( x_j \) after \( r \) cycles, \( KNN(x_j) \) is the set of K natural neighbors of \( x_j \).

2.2 Natural stable structure

Natural stable structure allows points to determine the neighbor status of each point by its own distribution and density characteristics. As the number of cycles \( r \) increases, the number of points with natural neighbors increases. Since there are global outliers in dataset, when the number of cycles increases to \( r+1 \), the number of points with natural neighbor does not change, indicating that the dataset has reached natural stable structure. Natural stable structure can be expressed as:

\[
\forall \ x_i \exists x_j (r \in N) \land (x_i \neq x_j) \rightarrow (x_i \in NN_i(x_j)) \land (x_j \in NN_j(x_i))
\]

\( r-1 \) called Natural Neighbor Eigenvalue (NaNE).

From formula (1) and (2), we derive Algorithm 1. Algorithm 1 describes the principle of the NaN algorithm.

**Algorithm 1.** Natural Neighbor algorithm (NaN)

**Input:** \( X \) \{Training Set\}

**Output:** NaN, Nb

1: Initializing:
\[ r = 1, Nb(i) = 0, NN_i(i) = \emptyset, NaN_i(i) = \emptyset \]

2: Create a kd-tree from data set

3: For each point \( i \) in \( X \), find its \( r \)-th neighbor \( j \) by using \( T \)

   a. \( Nb(j) = Nb(j) + 1 \)  
   b. \( NN_i(i) = NN_i(i) \cup \{ j \} \)  
   c. \( NaN_i(j) = NaN_j(j) \cup \{ i \} \)  

   (from formula (1))
4: Compute Num \% Num is the number of data point \( i \) that \( Nb(i) = 0 \)
5: If the Num doesn’t change
    \( r = r + 1 \) then execute step 3 and step 4
    if the Num still doesn’t change
        go to step 7
    else go to step 3
6: Else
    \( r = r + 1 \) then go to step 3 (from formula (2))
7: Output NaN, Nb

\( NaN_r(i) \) represents the natural neighbor of \( i \) after \( r \) cycles, \( Nb(j) \) is the number that point \( j \) is contained by the neighbor of other points, which also is the number of \( NaN_r(i) \). \( NN_r(i) \) is the natural neighbor of \( i \) after \( r \) cycles.

3. **Data prepossessing**

We classify Wisconsin’s breast cancer prognosis dataset culled from UCI machine learning repository. There are 11 attributes, the first one is ID, the last attribute is label, and the nine attributes affect the classification accuracy. The nine attributes are: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses.

The dataset consists of 699 clinical cases and contains missing data for 16 observations. We exclude the missing data and performed data preprocessing on the remaining 683 samples.

3.1 **Noisy filter**

For \( \forall x_i \in NN_r(x_j) \), if the labels of \( x_i \) and \( x_j \) are inconsistent, and the number of \( x_i \) is greater than half of the number of cycles \( r \) of NaN algorithm, then \( x_j \) is recorded as a noise point. Algorithm 2 describes the noisy filter (edited natural neighbor algorithm) based on Natural Neighbor (NaN) algorithm.

**Algorithm 2.** The Edited Natural Neighbor algorithm

**Input:** \( X \) (Training Set)

**Output:** NP (Noise point)

1: Initializing: \( r, Nb(i), NN_r(i), NaN_r(i) \) % the results inheritance from Algorithm 1
2: Determine the type of natural neighbor of \( i \) and the sample size of each types
3: If the sample size of one type which the same as \( i > r/2 \)
   \[ NP = NP \]
   Else
   \[ NP = NP \cup \{ i \} \]
4: Output NP

3.2. **Eliminating global outliers**

When the number of cycles increase to \( r \), the number of points with natural neighbors unchanged, the training set reaches natural stable structure [6]. The points without natural neighbors are recorded as global outliers. Figure 1 describes the process of NaN algorithm eliminating global outliers.
Figure 1. The process of NaN eliminating global outliers.

At the end of the first cycle, point 3, 6, 7, and 9 are not treated as neighbors by other points and point 1, 2, 4, 5, 8, 10 are treated as neighbors by other points. In (a), point 3, 6, 7 and 9 enclosed by red box have no neighbor in the first cycle. At the end of the second cycle, only point 6, 9 are not treated as neighbors by other points. Point 3, 7 are treated as neighbors by other points. After second cycle, only points 6 and 9 remain no neighbors in (b). At the end of the third cycle, point 6 and 9 are still not treated as neighbors by other points. NaN algorithm ends and point 6 and 9 are recorded as noise points. NaN algorithm reaches the natural stable structure and constructs a natural neighbor graph as (c) shows. The arrow 3 pointing to 7 shows that 3 is 7’s NaN. The double arrow between 2 and 3 indicates that 2 and 3 are mutual natural neighbor.

Table 1. Natural neighbor and the number of natural neighbors.

| No. | Nb(x) | NaN(x) |
|-----|-------|--------|
| 1   | 2     | 5,10   |
| 2   | 3     | 3,4,6  |
| 3   | 3     | 2,6,7  |
| 4   | 7     | 1,2,3,5,6,7,8 |
| 5   | 5     | 1,2,8,9,10 |
| 6   | 0     |        |
| 7   | 3     | 3,4,8  |
| 8   | 4     | 4,7,9,10 |
| 9   | 0     |        |
| 10  | 3     | 1,5,9  |

Table 1 shows NaNs and the number of NaN for each point. For example, point 1 has two NaNs that are point 5 and point 10, and point 4 has seven NaNs that are point 1, 2, 3, 5, 6, 7, 8.

3.3 Experimental analysis

3.1 and 3.2 show the process of the noisy filter and eliminating global outliers by natural neighbor (NaN) algorithm respectively. Put the 683 samples into natural neighbor (NaN) algorithm for training. When the training stop, the number of cycles \( r \) is 26, which means each point has found the nearest 26 neighbors. Combined \( Nb(j) \), the number that point \( j \) contained by the neighbor of other points, we can get the natural neighbors of each point.

Put the 683 samples into natural neighbor (NaN) algorithm for training. When the training stop, there are 23 noise points in total. Points without natural neighbors are global outliers, when the training stop, there are 8 global outliers in total. After cleaned the training set, the number of samples of Wisconsin’s breast cancer prognosis dataset for classification is 652.

4. Breast cancer diagnosis by using Natural Neighbor (NaN) algorithm
Cross validation, a practical method for cutting data samples into smaller subsets, was proposed by Seymour Geisser [7]. Cross validation is often used to test and evaluate the accuracy of an algorithm or model. The basic idea of it is to group the dataset, one part as a training set and the other as a testing set. First, training set is used to train the algorithm, and then testing set is used to test the model that trained by training set. The results obtained are used as an indicator to evaluate the performance of the algorithm [8].

We divide the pre-processed 652 Wisconsin’s breast cancer samples into 10 groups, and classify each point in the test set and compare it with its actual type, change the test set and calculate the classification accuracy of different test sets. We use NaN algorithm to classify the testing set and the average result of each 10 calculations is regarded as the result of one cross-validation. The cross-validation accuracy of testing set after pre-processing is shown in Table 2.

| No. | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|-----|----|----|----|----|----|----|----|----|----|----|
| 1   | 98.48 | 93.94 | 96.97 | 96.97 | 94.03 | 98.48 | 89.39 | 98.48 | 98.48 | 100.00 |
| 2   | 93.94 | 96.97 | 95.45 | 98.48 | 98.51 | 93.85 | 100.00 | 90.91 | 98.48 | 98.48 |
| 3   | 100.00 | 96.92 | 96.97 | 96.97 | 96.97 | 96.97 | 96.97 | 98.48 | 90.91 | 96.97 |
| 4   | 93.94 | 93.94 | 96.97 | 95.45 | 98.48 | 100.00 | 89.39 | 95.38 | 98.51 | 95.45 |
| 5   | 93.94 | 98.51 | 96.97 | 96.97 | 96.97 | 96.92 | 100.00 | 98.48 | 92.42 | 98.48 |
| 6   | 96.97 | 95.45 | 98.48 | 95.38 | 96.97 | 98.48 | 93.94 | 96.97 | 97.01 | 98.48 |
| 7   | 96.97 | 96.92 | 95.52 | 98.48 | 98.48 | 98.48 | 93.94 | 98.48 | 98.51 | 94.54 |
| 8   | 98.48 | 95.45 | 96.97 | 98.48 | 93.94 | 98.48 | 90.91 | 98.51 | 100.00 | 98.48 |
| 9   | 96.97 | 93.94 | 90.91 | 95.45 | 100.00 | 95.45 | 98.46 | 96.97 | 98.51 | 95.45 |
| 10  | 98.48 | 95.45 | 93.94 | 98.46 | 95.45 | 92.42 | 95.45 | 100.00 | 97.01 | 98.48 |
| Average accuracy | 96.82 | 95.75 | 95.92 | 97.11 | 96.98 | 96.96 | 95.00 | 96.51 | 97.29 | 97.88 |

The average result of each 10 calculations is regarded as the result of one cross-validation [9]. After 10 times cross-validation, the pre-testing set classification accuracy and post-pre-testing set classification accuracy obtained by NaN algorithm are shown in Figure 2. We use K-Nearest Neighbors (KNN) algorithm to classify the Wisconsin’s breast cancer prognosis dataset and compare the accuracy with obtained by NaN algorithm. The results are shown in Figure 3.

![Figure 2. The accuracy of testing set by NaN](image1.png)

![Figure 3. The accuracy of testing set by NaN and KNN](image2.png)

Before the data pre-processing, the classification accuracy of Natural Neighbor (NaN) algorithm for Wisconsin’s breast cancer prognosis dataset is 94.66%. After data pre-processing, the accuracy is 96.62%, which is higher 1.96% than before data pre-processing. Natural Neighbor Eigenvalue (NaNE) is
25, which means the number of neighbors sought by each point in NaN algorithm is 25. It can be concluded from Figure 3 that the accuracy of KNN algorithm is up to 97%, but there is a large fluctuation [10], the accuracy decreases with the increase of K. The accuracy of NaN algorithm is 96.62%. The classification accuracy obtained by NaN algorithm basically reaches the highest classification accuracy obtained by KNN algorithm and NaN does not need any parameters.

5. Discussion
Selection of parameter K is an unsolved problem for K-Nearest Neighbors (KNN). Besides, concerning the time complexity, because of using the kd-tree in the stage of searching the natural neighbor, the time complexity of Natural Neighbor (NaN) algorithm is \(O(n \log(n))\), while KNN has \(O(n^2)\). Therefore, NaN is somewhat faster than KNN. The dimension of data set can be reduced to improve the speed and accuracy of the algorithm. It can also reduce the training sample, improve the speed of the algorithm, and reduce the algorithm space consumption.

6. Conclusions
In this paper, we use an improved K-Nearest Neighbors (KNN) method called Natural Neighbor algorithm (NaN) to classify the Wisconsin’s breast cancer prognosis dataset. The method uses the novel concept of natural neighbor to remove noisy patterns and outliers. The main advantages are that NaN does not need any parameters and can maintain high classification accuracy. Experiments show that Natural Neighbor algorithm automatically filters noisy patterns while improves the classification accuracy. In addition, the classification accuracy obtained by Natural Neighbor algorithm (NaN) basically reaches the highest classification accuracy compared with KNN algorithm with different K values. Moreover, Natural Neighbor algorithm does not need any parameters.

7. References
[1] S. Bharati, M. A. Rahman and P. Podder, "Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA," 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), Dhaka, Bangladesh, 2018, pp. 581-584.
[2] K. Odajima and A. P. Pawlovsky, "A detailed description of the use of the kNN method for breast cancer diagnosis," 2014 7th International Conference on Biomedical Engineering and Informatics, Dalian, 2014, pp. 688-692.
[3] A. P. Pawlovsky and M. Nagahashi, A method to select a good setting for the kNN algorithm when using it for breast cancer prognosis, IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Valencia, 2014, pp. 189-192.
[4] L. Rong and S. Yuan, Diagnosis of Breast Tumor Using SVM-KNN Classifier, 2010 Second WRI Global Congress on Intelligent Systems, Wuhan, 2010, pp. 95-97.
[5] Congqin-Yi, Ruyan Zhou and Kening Hu, Fuzzy Support Vector Machine for breast cancer gene classification, 2017 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, 2017, pp. 676-679.
[6] Lijun Yang, Qingsheng Zhu, Jinlong Huang, Dongdong Cheng, Adaptive edited natural neighbor algorithm, Neurocomputing, Volume 230, 2017, pp. 427-433.
[7] L. Lingqiao, Y. Huihua, H. Qian, Z. Jianbin and G. Tuo, Design and Realization of the Parallel Computing Framework of Cross-Validation, 2012 International Conference on Industrial Control and Electronics Engineering, Xi’an, 2012, pp. 1957-1960.
[8] C. Alippi and M. Roveri, Virtual k-fold cross validation: An effective method for accuracy assessment, The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, 2010, pp. 1-6.
Acknowledgements
This work was supported by the Fundamental Research Funds for the Central Universities (No.2018NQNo5).