A Study on the Use of the kNN Algorithm for Prognosis of Breast Cancer

Alberto Palacios Pawlovsky [1], Mai Nagahashi [2]

桐蔭横浜大学医用工学部

(2014年3月20日　受理)

Abstract

This paper details the implementation of the kNN (k Nearest Neighbors) algorithm and the results of its use for prognosis of breast cancer. We used its implementation with the breast cancer data of the UCI repository and found that it has nearly 73% of average accuracy when it prognosticates the recurrence of cancer.

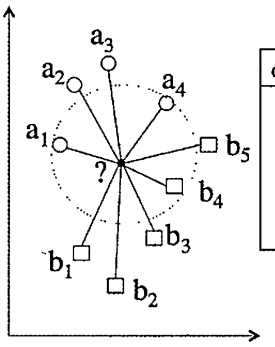
Keywords: kNN algorithm, machine-learning, prognosis, breast cancer, classification tool.

1. Introduction

Nearly 12,000 women in Japan die of breast cancer every year. It is the most frequent type of cancer among women in Japan [1]. Therefore, it is very important to forecast its recurrence after a first treatment. There are some algorithms in machine learning that have been used to predict the survival of a patient with cancer [2], for diagnosis of breast cancer [3] and for prognosis of it [4]. The k-NN (k-Nearest Neighbor) algorithm is a simple but powerful algorithm proposed by Fix and Hodges [5]. It has been used in many fields and is easy to implement. In the following section we briefly describe it. In section 3 we show the results obtained with it using the breast cancer data for prognosis of the UCI repository [6]. We end this study with some conclusions and topics for future work.

2. kNN Algorithm

The kNN algorithm is a non-parametric instance-based algorithm that can be used for regression and classification.



datum	distance
b_4	3.8
a_1	4.1
a_4	4.4
b_5	4.7
:	:

Fig. 1. Example of classification using a 3-NN.

[1]Alberto Palacios Pawlovsky : Faculty of Biomedical Engineering, Dept. of Clinical Engineering, Toin University of Yokohama, 1614 Kurogane-cho, Aoba-ku, Yokohama, Japan 225-8503
[2]Mai Nagahashi:Faculty of Engineering , Dept. of Clinical Engineering

It classifies a target object based on the number of members of the class nearest to it. Usually we measure the distance of all the objects, used for classification, to the target object and assign to it the class more common within its k nearest neighbors. An example is shown in Figure 1 where the target object is shown as the black point with an interrogation mark. The example shows nine objects with two classes. The table in the figure shows the Euclidean distances of some neighbors to the target object sorted in ascending order. If we use 3 of the neighbors for classification the class assigned to the tar-get object will be in this case "a".

To evaluate the kNN algorithm we usually take a group of classified data and divide it in two sets. The first set is used for classification and the second one is used for testing (see Figure 2).

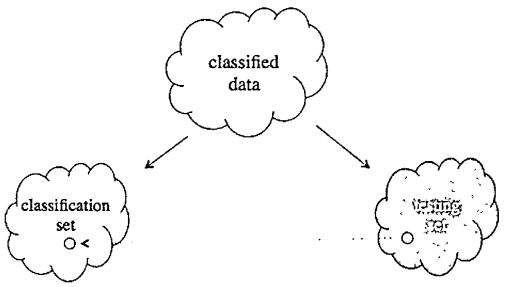


Fig. 2. Evaluation scheme used for the kNN algorithm.

Then we take one datum at a time from the testing set and classify it using the classification set. After classifying all the data in the testing set, we calculate the number of times the classification result matched the class of the target object and determine the accuracy of the algorithm.

To measure the similarity of a target datum to the data in the classification set the kNN algorithm uses the Euclidean distance in our implementation. However, it is possible to use other distances. The work of [7] shows results that indicate that the Euclidean and Manhattan distances give the best results.

In this study we divided the 194 records of the UCI data for prognosis using 10%, 20%, ..., 80% and 90% of it as the classification set, and varied the number of neighbors k from 1 to 10. The results of this implementation are shown in the following section.

3. Experimental Results

We implemented the kNN algorithm and run simulations using the UCI breast cancer prognosis data. The original data has 198 records of breast cancer patients, but four of these records lack the number of lymph nodes and we excluded them in this study. Each record has 35 features. The first one is the ID of the patient, the second one indicates with one letter the recurrence (R) or not recurrence of cancer (N) . The third feature is the time to recurrence or the time without it. The following 30 features are related to the sample taken from the patients and detail characteristics of the cells of the sample. The first 10 values of these 30 features are related to the dimensions and characteristics of the cells. The following 10 ones are the standard error of the first 10 cell features and the last 10 values in this group are the worst values of them. The 34th value is the size of the tumor and the last 35th is the number of lymph nodes.

We run simulations with groups for classification that take 10% to 90% of all the data in increments of 10% for a total of nine settings for the size of the classification set. We also varied the number of neighbors, k, from 1 to 10 for a total of ten k settings. Since the set used for classification is formed with data randomly chosen from all the available data we repeated each simulation

from 100 to 1,000 times in increments of 100, for a total of ten settings for the number of runs. In total we used 900 different settings for a total of 495,000 simulations.

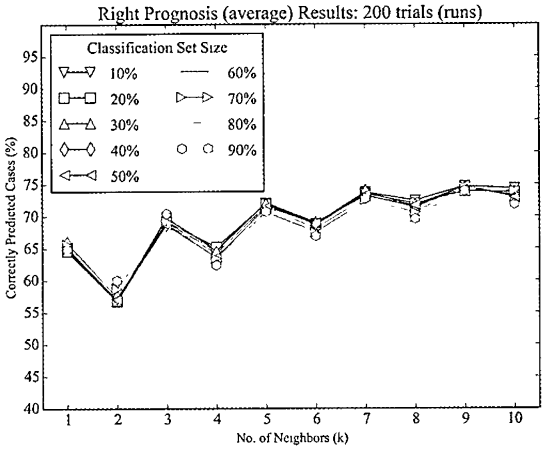


Fig. 3 Average accuracy with 200 runs.

Fig. 3 shows the average accuracy results obtained with 90 settings of k and the size of the classification set for 200 runs with each setting. In our study the data belongs to either of two classes, N (non-recurrent) or R (recurrent) . When we use an even number for k there are some cases were the majority vote comes to a draw. In our implementation the first class after sorting is chosen as the class of the target datum. This causes that the accuracy decreases for these values of k. We can also see that the accuracy tends to increase with k. It seems also that the average accuracy reaches a maximum near 74%.

Figure 4 shows the results when running each setting 800 times. We can see that there are almost no differences when comparing them to those of figure 3.

We also obtained the maximum and minimum accuracy results for all the settings we tried with 19 values of k.

Figures 5 and 6 show the average, maximum, and minimum accuracy results with classification set sizes of 20% and 40%, respectively.

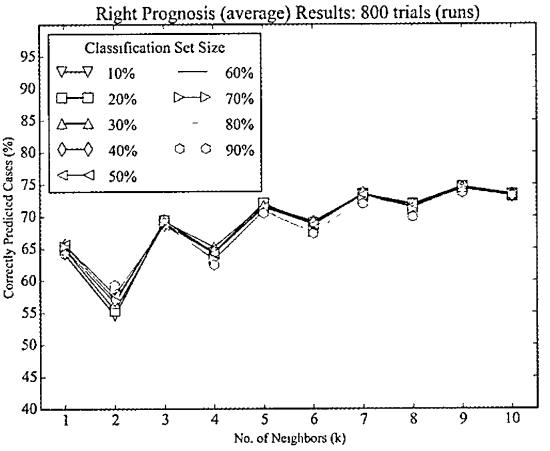


Fig.4 Average accuracy with 800 runs.

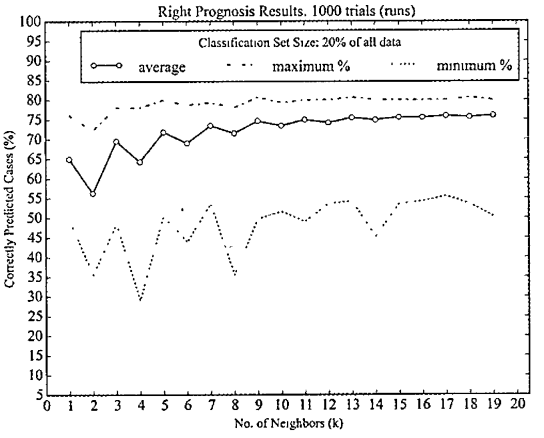


Fig.5 Accuracy values for a 20% classification set.

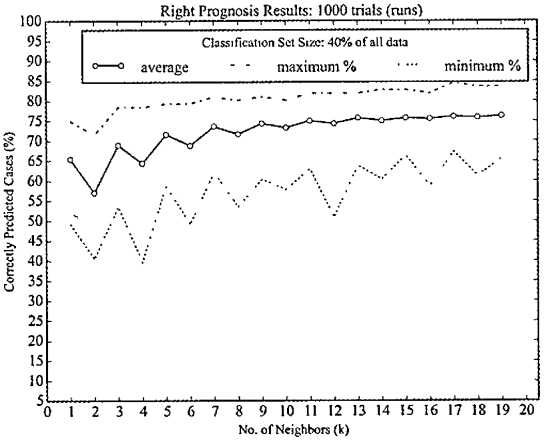


Fig.6 Accuracy values for a 40% classification set.

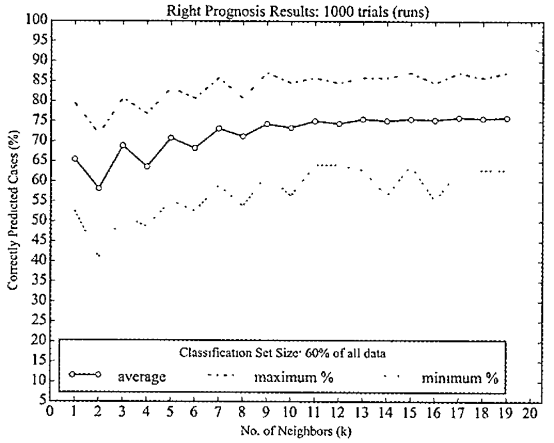


Fig.7 Accuracy values for a 60% classification set.

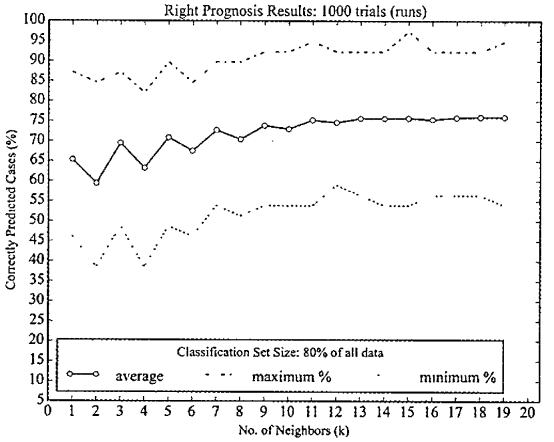


Fig.8 Accuracy values for a 80% classification set.

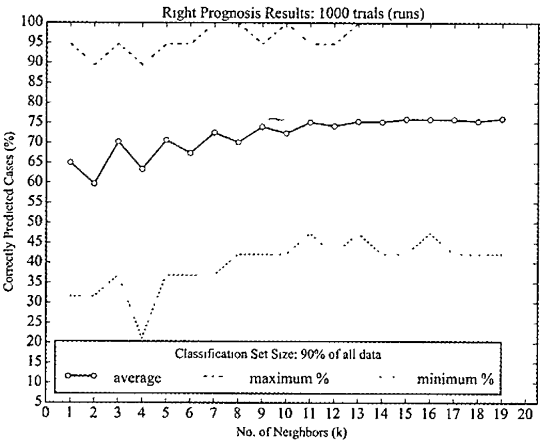


Fig.9 Accuracy values for a 90% classification set.

Figures 7, 8 and 9 show the accuracy results with classification sets of 60%, 80% and 90% of all data.

We can see that the accuracy maximum and minimum values are closer to the average values when we increase the classification set size towards 40%, but the range of their values widen as the set size increases above 40%.

We can also see that when using 90% of all the data it is possible to obtain a maximum accuracy of 100%. However, depending of the value of k, it is also possible to correctly predict only 20% of all cases (see Figure 9).

4. Conclusions

We have implemented and run several simulations with the kNN algorithm to evaluate its accuracy when using it for breast cancer recurrence prediction. The implementation of the algorithm was done in Python [8]. The running time is of almost two hours in a 2.7 GHz PC for the 940,500 simulations run when generating average, maximum, and minimum values for nineteen settings of k as shown in figures 5 to 9. Our results show that the kNN algorithm is simple and powerful, but it also could give very poor results. Our implementation is the simplest one and uses all the 32 values of the record of a patient. We plan to study it when applying principal component analysis to determine the most important features in a record. We also plan to study the effect of using other distance definitions as a measure of similarity [9]. There are a lot of variants for the kNN algorithm [10]. We plan to study and evaluate them for cancer prognosis and expect to develop and suggest new ap-proaches to improve it.

[Reference]

[1] http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/kakutei12/ dl/11_h7.pdf

[2] Arihito Endo, Takeo Shibata and Hiroshi Tanaka, "Comparison of Seven Algorithms to Predict Breast Cancer Survival," Biomedical Soft Computing and Human Sciences, Vol. 13, No.2, pp. 11-16, 2008.

[3] Jini R. Marsilin and G. Wiselin Jiji, "An Efficient CBIR Approach for Diagnosing the Stages of Breast Cancer Using KNN Classifier," Bonfring International Journal of Advances in Image Processing, Vol. 2, No. 1, pp.1-5, March 2012.

[4] Shomona G. Jacob and R. Geetha Ramani, "Efficient Classifier for Classification of Prognosis Breast Cancer Data Through Data Mining Techniques," Proceedings of the World Congress on Engineering and Computer Science 2012, Vol. I, October 2012.

[5] E. Fix and J. L. Hodges, "Discriminatory analysis, nonparametric discrimination: Consistency properties," Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[6] http://archive.ics.uci.edu/ml/datasets.html

[7] Seyyid A. Medjahed, Tamazouzt A. Saadi, and Abdelkader Benyettou, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules," International Journal of Computer Applications, Vol. 62, No.1, pp.1-5, January 2013.

[8] https://www.python.org

[9] Sung-Hyuk Cha, "Comprehensive Survey of Distance/Similarity Measures between Probability Density Func-tions," International Journal of Mathematical Models and Methods in Applied Sciences, Vol. 1, Issue 4, pp.300-307, 2007.

[10] Nitin Bhatia and Ashev Vandana, "Survey of Nearest Neighbor Techniques," International Journal of Computer Science and Information Security, Vol. 8, No. 2, pp.302-305, 2010.