A kNN Method for Breast Cancer Prognosis that Uses a Genetic Algorithm for Component Selection

Alberto PALACIOS PAWLOVSKY [1] and MATSUHASHI Hiroki [2]

[1] Faculty of Biomedical Engineering and [2] Dept. of Clinical Engineering,
Toin University of Yokohama

（2017 年 3 月 18 日　受理）

***Abstract* — This paper shows the application of a genetic algorithm (GA) to improve the accuracy of a kNN method when using it for breast cancer prognosis. The GA is used to select the components of the data used by the kNN method. We have found a combination of only eleven features that rises the average accuracy of kNN to nearly 78%.**

I. Introduction

Breast cancer is one of the most common cancers in the world. Almost one third of all women cancer patients in Japan suffer from breast cancer. Breast cancer has a high probability of surviving it if detected at an early stage. The advances in technology in recent years have fostered the research of tools to help a physician to diagnose many types of diseases. Many machine learning methods for data classification have been used for breast cancer prognosis and diagnosis [1–4].

This paper deals with the improvement of the accuracy obtained with the kNN (k-Nearest Neighbor) method. kNN is a nonparametric classification method that has high accuracy. It has even been used to detect different stages of breast cancer [5, 6]. We have been researching on several ways to improve kNN's accuracy when using it for breast cancer diagnosis and prognosis.

In the kNN method we can change the percentage of data used for classification, the number of neighbors k and choose to normalize the data before using it. Normalization of the data helps to improve the accuracy of kNN [7–9] for prognosis and diagnosis. Usually, in the kNN method the similarity metric is based on the Euclid distance extended to the number of components in the data. Changing this metric usually leads to improvements too [10, 11].

When developing methods for prognosis and diagnosis researchers usually use the data sets of breast cancer of the UCI site [12]. The prognosis set contains data that is composed of 35 items of which 32 are used for classification. One can use all the items or just a few ones of them. Principal component analysis (PCA) is one way of choosing them, but we can combine it also with other methods [13, 14].

[1] Alberto PALACIOS PAWLOVSKY and [2] MATSUHASHI Hiroki

[1] Faculty of Biomedical Engineering and [2] Department of Clinical Engineering, Toin University of Yokohama. 1614 Kurogane-cho, Aoba-ku, Yokohama 225-8503, Japan

We could try to test all possible combinations of the components in the data but since it is impractical, in most cases, we usually use heuristic algorithms to try to find one near-optimal combination. In this paper, we use a genetic algorithm (GA) for this purpose.

Genetic algorithms have been applied in the reduction of the dimensionality of the data in many fields [15, 16].

We have implemented and tested a kNN that uses a genetic algorithm (GA) to select the components in the data. We used this implementation with the data for cancer prognosis of UCI. Our implementation normalizes the data and uses the Manhattan distance as a similarity metric. Details are given in the following section. We show the corresponding experimental results in section III. And at the end we discuss some topics for further research.

II. Our kNN and GA Implementation

In the following subsections, we show some details on the use of the kNN method and characteristics of the GA we have implemented.

1. kNN Method

The k-Nearest Neighbor (kNN) method is an unsupervised nonparametric machine learning method used for classification tasks. To evaluate the accuracy of it and other classification algorithms we usually divide an already classified data in two sets. One is used for the classification task and the other one is used as a test set. Then, one datum at a time is taken from the test set and compared with the data in the classification set. The percentage of correct classifications determines the accuracy of the method for the given classification set.

One method widely used to evaluate the accuracy of kNN is ten-fold cross validation. In it a complete set of classified data is divided in ten equal parts and one is used as a test set and the remaining nine ones as the set for classification (the ones surrounded by the point-line enclosure in ***Figure 1***).

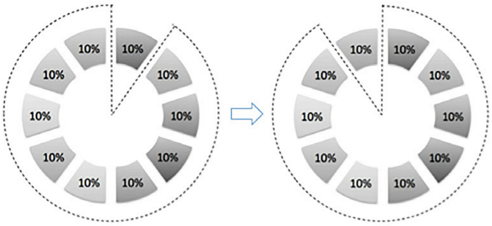

Fig. 1. Ten-fold cross validation: first and last sets.

Then the next part is used as a test set and so on, repeating this process 10 times. The accuracy is given as the average of the accuracies obtained in each step of this process.

We have found that the highest average accuracies are indeed obtained using 90% of all the data, but most of the time they also have very small minimum accuracies and sometimes also high standard deviation figures. Therefore, in the evaluation of our approach we have used nine different settings for the size of the classification set.

2. GA Implementation

Our GA implementation is a standard one. It uses all members in one generation to generate the new members of the next generation. Two contiguous members are used to generate two new members of the next generation.

Figure 2 shows the components of one datum of the breast cancer prognosis data. As shown before only the 32 last ones are used in classification.

The members of a generation in the GA are patterns of zeros and ones that mask out some components in one datum (***Figure 3***). The first generation of patterns (members) is randomly generated.

Subsequent generations are formed using one-point crossover and mutation applied to each zero and one of a member. Each member in one gener-

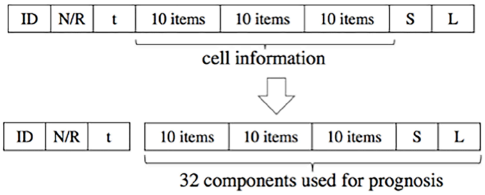

Fig. 2. UCI breast cancer prognosis datum details.

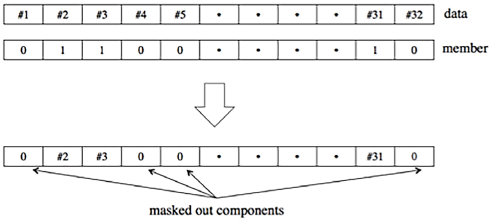

Fig. 3. Masking of components using a GA member.

ation is evaluated using kNN. Additional details follow.

3. Pre-processing: Component Selection

The GA described above is used to select the best patterns of all tried. In the preprocessing stage, we start with a first generation generated randomly (***Figure 4***).

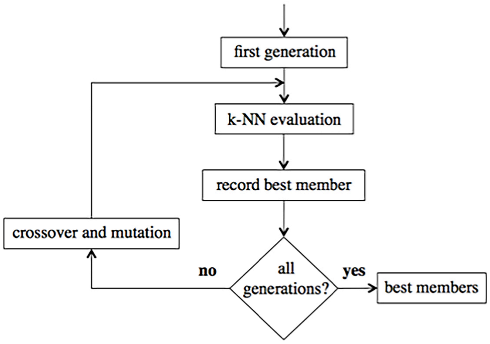

Fig. 4. Evaluation of GA members using kNN.

We then evaluate each of its members (patterns) using our implementation of the kNN method. It uses the Manhattan distance as similarity metric. And evaluates each member with nine different sizes of the classification set (they go from 10% to 90% in increments of 10%). For each classification set size all the possible number of neighbors are used. This maximum number is 19 for a classification set size of 10% and 175 for a classification set size of 90% of all available data (194 patients).

After each evaluation, the best member is recorded. Each evaluation with the same setting is repeated 10 times and the selection of the best member is done based on the best average accuracy got with all the member in the generation. We repeat the preprocessing stage for a given number of generations. In our pre-processing stage, we use 10 generations each of 100 members. If we have not used all generations, we form a new one and repeat the process.

This preprocessing stage is repeated for 91 combinations of the probability of crossover pc and probability of mutation pm. Both probabilities have nine settings from 10% to 90% in increments of 10%. After the evaluation of all generations and corresponding members (a total of 91000 patterns in the preprocessing stage) the best patterns recorded are sorted and used as input in the following stage (***Figure 5***).

The last stage, briefly detailed in ***Figure 5*** uses only some of the top patterns of the preprocessing stage to perform a more detailed evaluation. In it only those patterns with an average accuracy equal or greater than 76% are used. It is a kind of verification stage where each pattern is evaluated, but only for the size of the classification set where it was found to excel.

Since the set for evaluation in our kNN is formed choosing data randomly from the whole

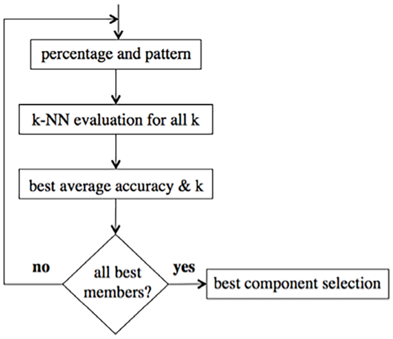

Fig. 5. Verification process of the top patterns.

available data, the evaluation of the target pattern is repeated a hundred times. Even though the value of k that makes the pattern give a high accuracy is known, at this stage all possible values of k are again used in the evaluation. After the evaluation of all patterns chosen from the preprocessing stage, the results are sorted again. In our implementation, only the top three ones are given as best results. Details of those patterns are given in the next section

III. Experimental Results

The best pattern results at the preprocessing stage with a classification set size of 30% are shown in ***Figure 6***.

This pattern gave a best average accuracy of 77.57% with any number of neighbors k greater than 31.

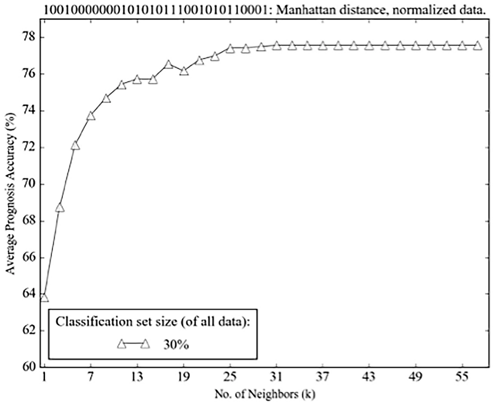

Fig. 6. Best pattern's pre-processing accuracy results.

This pattern was found with a probability of crossover of 30% and a probability of mutation of 80%. The corresponding results in the verification stage are shown in ***Figure 7***. The average accuracy obtained was lower at this stage with a best average accuracy of 76.46%. This pattern shows a very small standard deviation (±2.04%) and almost constant maximum (≈ 81%) and minimum (≈ 72%) average accuracy values.

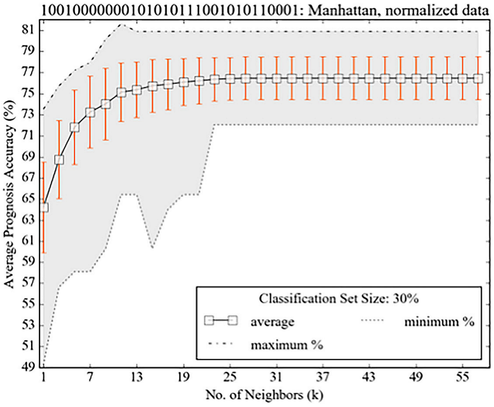

Fig. *7*. Best pattern's verification accuracy results.

In ***Figure 8*** we show the average accuracy results obtained with the second-best pattern at the preprocessing stage. It showed a best average value of 77.32% when using 23 neighbors and a classification set size of 50%.

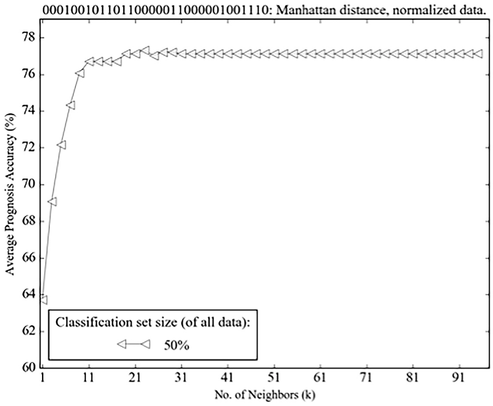

Fig. 8. Second best pattern's accuracy results.

The corresponding average accuracy values at the verification stage are shown in ***Figure 9***. It has a little higher standard deviation of ± 2.9%. The range of values for its average accuracy goes from a minimum of 69% to a maximum of 81.4%. Its range of variation (12%) is almost constant for higher values beyond k = 23, but the best average accuracy value was obtained only with this number of neighbors. It also showed a small unstable variation for neighbor's values between 24 and 30. The average accuracy then remained constant at

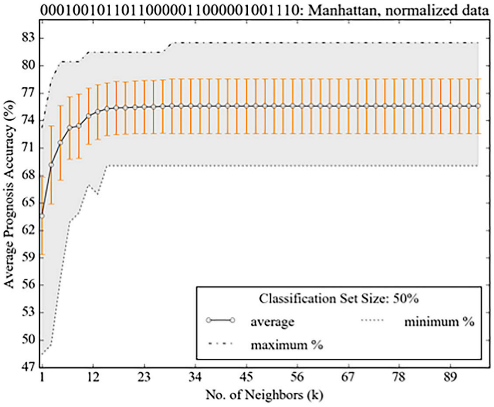

Fig. 9. 2nd. best pattern's verification accuracy results.

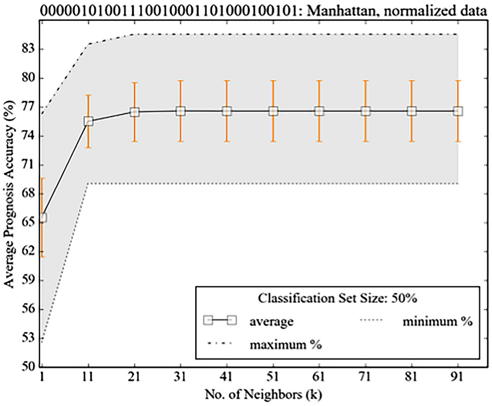

Fig. 11. 3rd. best pattern's verification accuracy results.

77.1% from k = 31 and any other higher number of neighbors.

The average accuracy results obtained with the third best pattern at the preprocessing stage are shown in ***Figure 10***. This pattern was found with a probability of crossover of 30% and a probability of mutation of 70%.

Its best average accuracy was of 77.01% for a number of neighbors of 17 (the same accuracy was obtained also with k = 19).

Its best average accuracy value at the verification stage was 76.32%, a little bit higher than the second-best pattern (***Figure 11***).

However, it had a maximum accuracy of 84.5% and a minimum of 69% (wider range = 15.46%), and a standard deviation of 3%.

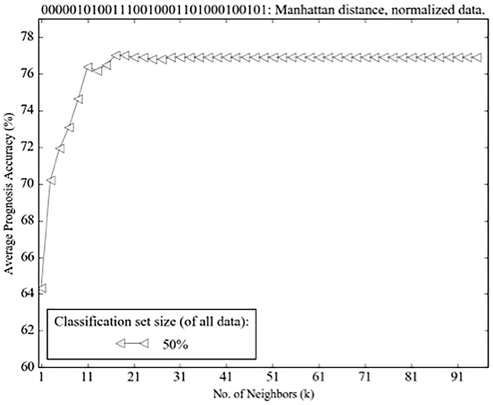

Fig. 10. Third best pattern's accuracy results.

IV. Conclusions

We detailed in this paper an implementation of a GA that could be used to select components in data targeted by a kNN method. We show in the experimental section that for the breast cancer prognosis data set of the UCI [12], it is possible to find components that increase the average accuracy by almost 2%. We are now working in a modification that will search for common items to build a member (pattern) that will be used again with a GA for a second preprocessing stage.

[References]

1) Shelly Gupta, Dharminder Kumar and Anand Sharma, "Data Mining Classification Techniques Applied for Cancer Breast Diagnosis and Prognosis," Indian Journal of Computer Science and Engineering, Vol. 2, No. 2, pp. 188–195, May 2011.
2) Shweta Kharya, "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease," International Journal of Computer Science and Information Technology, Vol. 2, No. 2, pp. 55–66, April 2012.
3) Gouda I. Salama, M. B. Abdelhalim, and Magdy Abd-elghany Zeid, "Breast Cancer Diagnosis on Three Different Data sets Using Multi-Classifiers,"

International Journal of Computer and Information Tech- nology Vol. 1, Issue 1, pp.36–43, September 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) Manish Sarkar and Tze-Yun Leong, "Application of k-Nearest Neighbors Algorithm on Breast Cancer Diagnosis Problem," Proceedings of American Medical Informatics Association (AMIA) Annual Symposium, pp. 759–763, 2000.

6) 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.

7) Alberto Palacios Pawlovsky and Mai Nagahashi, "A Method to Select a Good Setting for the kNN Algorithm when Using it for Breast Cancer Prognosis," Proceedings of the 2nd. IEEE International Conference on Biomedical and Health Informatics (BHI 2014), pp.189–192, Sevilla, Spain, June, 2014.

8) Katsuyoshi Odajima and Alberto Palacios Pawlovsky, "A Detailed Description of the Use of the kNN Method for Breast Cancer Diagnosis," Proceedings of the 7th International Conference on Biomedical Engineering and Informatics (BMEI 2014) Dalian, pp. 606–610, China, October 2014.

9) Alberto Palacios Pawlovsky, Daisuke Kurematsu, "Improving the Accuracy of the kNN Method when Using an Even Number k of Neighbors," International Conference on Biomedical and Health Informatics, ICBHI 2015, Haikou, China, October 8–10, 2015.

10) 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.

11) Alberto Palacios Pawlovsky, Takumi Akiyoshi, Junja Hasegawa and Akihiro Asanuma, "Improving the kNN Method for Breast Cancer Diagnosis," 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015, Milan, Italy, August 25–29, 2015.

12) http://archive.ics.uci.edu/ml/datasets.html

13) A. Bucinski, J. Zaluski, J. Krysinski and R. Kalisszan, "A Principal Component Analysis of Patients, Disease, and Treatment Variables: A New Prognostic Tool in Breast Cancer After Masectomy," Rep. Pract. Oncol. Radiother. Vol.5, pp.8389, 2000.

14) S. Saxena, V. P. S. Kirar and K. Burse, "A Polynomial Neural Network Model for Prognostic Breast Cancer Prediction," International Journal of Advanced Trends in Computer Science and Engineering, Vol.2, No.1, pp.103–106, 2013.

15) M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn and A. K. Jain, "Dimensionality Reduction Using Genetic Algorithms," IEEE Trans. on Evolutionary Computation, Vol.4, No.2, pp.164–171, 2000.

16) M. Dolled-Filhart, L. Ryden, M. Cregger, K. Jirstrom, M. Harigopal,R. L. Camp and D. L. Rimm, "Classification of Breast Cancer Using Genetic Algorithms and Tissue Microarrays," American Association for Cancer Research, Clin. Cancer Res. Vol.12, No.21, pp.6459–6468, 2006.