Application of K-Nearest Neighbors Algorithm on Breast Cancer Diagnosis Problem

Manish Sarkar and Tze-Yun Leong
Medical Computing Laboratory
Department of Computer Science
School of Computing, The National University of Singapore
Lower Kent Ridge Road, Singapore: 119260
{manish, leongty}@comp.nus.edu.sg

Abstract: This paper addresses the Breast Cancer diagnosis problem as a pattern classification problem. Specifically, this problem is studied using the Wisconsin-Madison Breast Cancer data set. The K-nearest neighbors algorithm is employed as the classifier. Conceptually and implementation-wise, the K-nearest neighbors algorithm is simpler than the other techniques that have been applied to this problem. In addition, the Knearest neighbors algorithm produces the overall classification result 1.17% better than the best result known for this problem.

1 Introduction

Motivation: In medical diagnosis, the doctor uses his experience to draw diagnostic inference from the information supplied by (a) the tests performed on the patient, (b) the patient's physical condition, and (c) the patient's history. Diagnosis is a difficult task even for an experienced doctor because (a) the information contains uncertainty, (b) the amount of the information may be insufficient, and (c) part of the information may be misleading. To achieve better diagnostic results, we cast the diagnosis problem as a pattern classification problem, and we apply machine-learning techniques for the classification. The objective of this work is to apply simple machinelearning techniques to the Wisconsin-Madison breast cancer diagnosis problem [1] so that the classification results are enhanced.

Related Work: Many researchers [10] have measured the performance of their classification algorithms on the Wisconsin-Madison breast cancer problem. Most of these methods are, however, not specifically designed for the breast cancer problem. We shall discuss only those methods that are tailored to address the breast cancer problem, and in particular relevant to the Wisconsin-Madison breast cancer problem.

Setanio proposed [12] a rule extraction algorithm called *NeuroRule*. In this work, initially an artificial neural network is designed, and the rules are then extracted from the network. Two major components of the algorithm are pruning the network and clustering the hidden nodes of the network. The pruning algorithm is used to remove the redundant connections, and the clustering is used to discretize the activation values of the input pattern into small number of clusters. The pruning and clustering are needed because this technique is semiparametric in nature, and some information regarding the structure of the training set is required.

Taha *et al.* proposed in [15] [14] [16] three rule extraction algorithms. The rules are extracted from the artificial neural networks that are trained specifically for this problem. In Reyes' work [11], a fuzzy classifier system is evolved using genetic algorithm. The classification result of this classifier is substantially better than the classification result reported in Setanio's work [12].

In the latest work on this problem [13], Setanio has preprocessed the input data to select the most relevant attributes, and then like his earlier work [12] he fed the modified data set to NeuroRule. The selection of attribute is carried out using the neural networks with one hidden unit. The selection is used to decrease the training time and to enhance the classification result.

Conceptually, the attribute selection is carried out to make the structure of the training set more compact. The attribute selection is needed because this method is semiparametric, and hence the more knowledge the algorithm has about the structure of the training set, the better it performs.

Wisconsin-Madison Breast Cancer Problem: The presence of a breast mass may indicate (but not always) malignant cancer. *Fine needle aspiration* of breast masses is a popular diagnosis technique. The University of Wisconsin Hospital has collected 699 samples using the fine needle aspiration test. Each sample consists of the following ten attributes: (1) Patient's id, (2) clump thickness, (3) uniformity of cell size, (4) uniformity of cell shape, (5) marginal adhesion, (6) single epithelial cell size, (7) bare nuclei, (8) bland chromatin, (9) normal nucleoli and (10) mitosis. Except the patient's id, all other measurements are assigned to an integer value between 1 and 10, with 1 being closest to the benign and 10 the most anaplastic. Each sample is either benign or malignant. Various classifiers have been designed that can classify this data set into the benign and malignant classes.

The Classification problem and the relevant techniques: The task of pattern classification is defined as the search for the structures in a pattern set, and the subsequent labelling of the structures into categories such that the degree of association is high among the structures of the same category and low between the structures of different categories [3]. Most of the pattern classification techniques can be classified into the following three groups: (i) *parametric*, (ii) *semiparametric* and (iii) *nonparametric*. All the three techniques use a set of data that already has class labels. Henceforth, we call this data set the *training set*. The parametric and semiparametric classifiers need specific information about the structure of the data in the training set. In many cases it is difficult to collect this type of information. Hence, the nonparametric classification technique like the *K-nearest neighbors* (KNN) algorithm [4] becomes an attractive approach. It assigns the class label to the input pattern based on the class labels of the K-closest (in some distance sense) neighbors of the input. All the K-neighbors are from the training set, and the class label corresponding to most of the neighbors represents the class label of the input. The advantages of this algorithm are

1. It is simple to implement.
2. It works fast for small training sets.
3. It does not need any *a priori* knowledge about the structure of the data in the training set.
4. Its performance asymptotically approaches the performance of the Bayes classifier [2].
5. It does not need any retraining if the new training pattern is added to the existing training set.
6. The output of the KNN algorithm can be interpreted as an *a posteriori* probability of the input pattern belonging to a particular class [3]. Thus the output provides the relative class confidence levels.

2 Adopted Method

Preprocessing: The data set contains 16 samples each with one missing attribute. We have discarded these samples, as have been done by the other authors. Hence, a fair comparison of our results against their results can be made. The 683 samples (339 malignant and 444 benign) are split randomly into a training set that consists of 119 malignant and 222 benign samples. The test set consists of the remaining 120 malignant and 222 benign samples.

K-Nearest Neighbors Algorithm: In this method, for each test datum, the Euclidean distances between the test datum, and all the training data are calculated, and the test datum is assigned the class label that most of the K closest training data have [10].

The KNN algorithm assumes that all the data correspond to points in the N-dimensional space $\Re^N$. Let the test datum $\mathbf{x}_i$ be represented by the feature vector $\left[x_1^i, x_2^i, x_3^i, \ldots, x_N^i\right]'$, where x_k^i denotes the value of the kth attribute of the test datum $\mathbf{x}_i$, and $\mathbf{x}_i'$ is the transpose of $\mathbf{x}_i$. The distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{N}(x_k^i - x_k^j)^2}$. If the number of training data is n, then n such distances will be calculated, and the closest K training data are identified as *neighbors*. If $K = 1$, then the class label of the test datum is equal to the closest training datum. If $K > 1$, then the class label of the test datum is equal to the class label that most of the neighbors have. If there is a tie, then the tie is resolved arbitrarily. The output of the KNN algorithm attains a richer semantic when the output is interpreted as *a posteriori probability*. Hence, instead of labelling the output class label equal to the class label that most of the neighbors have, we assign the following class confidence values to $\mathbf{x}$:

$$p_c(\mathbf{x}) = \frac{1}{K}(\text{no of neighbors with class label } c) \quad \forall c$$
$$= \frac{1}{K}\sum_{i=1}^{K}\delta(i,c) \qquad \forall c \qquad (1)$$

where $\delta(i,c) = 1$ if x_i has the class label c and $\delta(i,c) = 0$ otherwise. Here, p_c is the *a posteriori* probability that $\mathbf{x}$ belongs to the class c. With this formulation, we can still consider the hard decision by assigning the class label j to the test datum $\mathbf{x}$ when $p_j(\mathbf{x}) = \max_{1,2,...,C}\{p_c(\mathbf{x})\}$ and C is the total number of classes.

One refinement to the KNN algorithm is to weigh the contribution of each of the K neighbors based on its distance to the test datum. Evidently, the closest neighbor should receive the highest weight. It can be accomplished by modifying Eqn. (1) into the following:

$$p_c(\mathbf{x}) = \sum_{i=1}^{K}\left(\frac{\frac{1}{d(\mathbf{x},x_i)^2}}{\sum_{j=1}^{K}\frac{1}{d(\mathbf{x},x_j)^2}}\right)\delta(i,c) \qquad \forall c \quad (2)$$

Here the denominator is used for normalisation such that $\sum_{c=1}^{C} p_c(\mathbf{x}) = 1$ holds. The KNN algorithm with this refinement is also known as the *fuzzy K-nearest neighbors algoritm* [6], and in that case $p_c(\mathbf{x})$ is interpreted as the fuzzy membership function.

In the KNN algorithm, the class labels of the training data may be discrete values (i.e., lying in $\{0,1\}$) or real values (lying in [0, 1]). Since the class labels for our problem are discrete, we have restricted our discussion for discrete class labels. The time complexity of the algorithm for testing is $O(nr)$, where n and r are the sizes of the training and test sets, respectively.

The KNN algorithm uses different decision boundary every time it encounters a test input, whereas other methods fix the decision boundary before any test input is observed. In other words, this technique is nonparametric in nature, and therefore, it does not need any information about the structure of the training set.

```
INPUT: (a) Already labelled training data
           {x_i | i = 1,2,...n }.
       (b) The test datum x.

ALGORITHM:
FOR i = 1,2, ...upto n
   Determine the distance between x and x_i.
   IF (i ≤ K)
      Include x_i in the set of K-nearest
      neighbors.
   ELSE IF (x_i is closer to x than any
   previous nearest neighbor)
      Delete the farthest of the K-nearest
      neighbors.
      Include x_i in the set of K-nearest
      neighbors.
   END IF
END FOR
FOR c = 1 to C

  p_c(x) = (1/K)(no. of neighbors in class c)  (3)
END FOR
Crisp class label of x is j
when p_j = max{p_1, p_2, ..., p_C}

OUTPUT: (a) Class label of x.
        (b) Class confidence values p_c ∀c.
```

Fig. 1: The K-nearest neighbors algorithm. The input consists of a set of labelled patterns and a test pattern. The output is the class confidence values of the test pattern. Here C is the total number of classes. If Eqn. (3) is replaced by Eqn. (2), then the resultant algorithm is the fuzzy K-nearest neighbor algorithm.

Table 1: Comparison of the classification results of [13] and that of the KNN algorithm on the Wisconsin-Madison breast cancer problem. The number of training samples for the malignant and benign cases are 119 and 222. The number of test samples for the malignant and benign cases are 120 and 222. In the best cases the KNN and fuzzy KNN (FKNN) algorithms enhance the overall classification result by 1.17% and 0.88% respectively.

	Training Set		Test Set			Training and Test Sets		
	Result in [13]	Our Result	Result in [13]	Our Result		Result in [13]	Our result	
				KNN	FKNN		KNN	FKNN
Malignant Sample	118/119 (96.00%)	119/119 (100.00%)	119/120 (99.17%)	115/120 (95.83%)	117/120 (97.50%)	237/239 (99.17%)	234/239 (95.83%)	236/239 (98.74%)
Benign Sample	218/222 (98.20%)	222/222 (100.00%)	216/222 (97.30%)	221/222 (99.55%)	221/222 (99.55)	434/444 (97.75%)	443/444 (99.77%)	443/444 (99.77%)
Overall	336/341 (98.53%)	341/341 (100.00%)	335/342 (97.95%)	336/342 (98.25%)	338/342 (98.83%)	671/683 (98.24%)	677/683 (99.12%)	679/683 (99.41%)

3 Results and Discussion

We have randomly chosen the data to construct the training set. Unlike the parametric and semiparametric classifiers, the KNN algorithm does not have any training session. We have experimented with different values of K from $K = 1$ to 15. With the KNN algorithm, the classification result of the test set fluctuates between 99.12% and 98.02%. The best performance is obtained when K is 1. Table 1 shows the best classification result, which is 0.88% better than that of [13]. When the fuzzy KNN is used, the best case performance is 1.17% better than that of [13]. With the fuzzy KNN algorithm, the classification performance varies in between 99.41% and 99.12%. Note that the worst case performance with the fuzzy KNN algorithm is better than the best performance of the classifier reported in [13]. Since the output class confidence values can be interpreted as an *a posteriori* probability or fuzzy membership values, the output values have richer semantics than just crisp class labels.

Compared to the methods reported in [12], [14] [11], [13], the advantages of the KNN algorithm are that the algorithm is very simple, and its implementation is very easy. Since there is no need of any training session, there is no convergence problem. In contrast, the other approaches employing neural networks may face the convergence problem, and may need long training time. New training data can also be added to the KNN algorithm without any retraining. But for the other techniques, adding new training data needs retraining because the new training data disturb the structure of the existing training set, and all the parametric or semiparametric classifiers critically depend on this structure.

4 Summary and Conclusions

Summary: This paper treats the Wisconsin-Madison Breast Cancer diagnosis problem as a pattern classification problem. The KNN algorithm is used as the nonparametric classifier. The KNN algorithm assigns the class label of the new datum based on the class label that most of the K-closest training data possess. The KNN algorithm yields the best classification performance that is obtained so far on this problem.

Conclusion: The good performance of the KNN algorithm does not imply that the KNN algorithm will be always good for all diagnosis problems. In fact there is no known single algorithm that performs well on all the diagnosis problems (if there were, we would have observed only one classification algorithm available for diagnosis). This work, however, highlights the potential usefulness of the KNN algorithm on different diagnosis problems.

Limitations: Some of the drawbacks of the KNN approach are (a) we need to store all the training data; hence for a large training set it may take a lot of space, and (b) for every test datum, the distance should be computed between the test datum and all the training data. Thus a lot of time may be needed for the testing. Fortunately, some fast versions of the KNN algorithm [9], [8] are available, and they have been successfully applied to other computation intensive tasks like script recognition and speech recognition. For instance, the data can be stored in the form of *kd-tree* [5] so that the nearby data are stored at the same or nearby nodes. The internal nodes of the tree sort the new query to the relevant leaf by testing the selected attributes of **x**. This paper does not attempt to improve the space and time complexity of the KNN algorithm, but shows the better classification results using the simple technique. Moreover, our work does not attempt to extract rules from the data.

Future work: There are some advanced versions of the KNN algorithm like the editing KNN algorithm [7], which in many cases provide better results than the KNN algorithm considered here. In future, we would like to investigate these algorithms in the context of the Breast Cancer problem and other relevant problems.

Acknowledgements: This work has been supported by a Strategic Research Grant No. RP960351 from the National Science and Technology Board and the Ministry of Education, Singapore.

References

[1] C. L. Blake and C. J. Merz. UCI repository of machine learning databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[2] T. M. Cover and P. E. Hart. Nearest neighbor pattern classifiers. *IEEE Transactions on Information Theory*, pages 21–27, 1967.

[3] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1979.

[4] E. Fix and J. L. Hodge. Discriminatory analysis: Nonparametric discrimination: Consistency properties. Project 21-49-004 4, USAF School of Aviation Medicine, Randolf Field, Texas., 1953.

[5] J. Friedman, J. Bentley, and R. Finkel. An algorithm for finding best matches in logarithmic time. *ACM Transaction on Mathematical Software*, 3(3):209–226, 1977.

[6] J. M. Keller and D. J. Hunt. Incorporating fuzzy membership functions into the perceptron algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 693–699, July/August 1985.

[7] L. I. Kuncheva. Editing for the K-nearest neighbors rule by a genetic algorithm. *Pattern Recognition Letters*, 16:809–814, 1995.

[8] L. Mico and J. Oncina. Comparison of fast neighbor classifiers for handwritten character recognition. *Pattern Recognition Letters*, 19:351–356, 1998.

[9] L. Mico, J. Oncina, and R. C. Carrasco. A fast branch and bound nearest neighbour classifier in metric space. *Pattern Recognition Letters*, 17:731–739, 1996.

[10] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[11] C. A. P. Reyes and M. Sipper. A fuzzy-genetic approach to breast cancer diagnosis. *Artificial Intelligence in Medicine*, 17:131–155, 1999.

[12] R. Setanio. Extracting rules from pruned networks for breast cancer diagnosis. *Artificial Intelligence in Medicine*, 8(1):37–51, 1996.

[13] R. Setanio. Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in Medicine*, 18:205–219, 2000.

[14] I. Taha and J. Gosh. Characterization of the Wisconsin breast cancer database using a hybrid symbolic-connectionist system. Technical Report UT-CVISS-TR-97-007, Computer and Vision Research Center, University of Texas, Austin, 1996.

[15] I. Taha and J. Gosh. Symbolic interpretation of artificial neural networks. Technical Report TR-97-01-106, Computer and Vision Research Center, University of Texas, Austin, 1996.

[16] I. Taha and J. Gosh. Evaluation and ordering of rules extracted from feed forward networks. In *Proceedings of IEEE International Conference on Neural Networks*, pages 221–226, 1997.