A FUZZY K-NN APPROACH FOR CANCER DIAGNOSIS WITH MICROARRAY GENE EXPRESSION DATA

Çiğdem Beyan, Hasan Oğul

Department of Computer Engineering, Başkent University
Eskişehir Road 20.km, Baglica Campus, 06810, Ankara
20493331@mail.baskent.edu.tr, hogul@baskent.edu.tr

ABSTRACT
Recent advances in DNA microarray technology have made it possible to measure the expression level of several thousand of genes simultaneously. The gene expression profiles obtained from microarray techniques have provided the opportunity of early diagnosis of cancer with the use of supervised learning algorithms. As a simple, effective and nonparametric classification method, k-Nearest Neighbor (k-NN) algorithm has recently been applied for the problem of cancer diagnosis and categorization. An obvious problem of traditional k-NN algorithm is that, when the density of training data is uneven, the precision of classification may reduce due to the consideration of first k nearest neighbors but not the differences of distances. A recent solution for this problem is adopting the theory of fuzzy sets and constructing a new membership function based on the similarities. This study has been conducted to demonstrate in what degree the fuzzification of k-NN algorithm can improve the prediction accuracy of cancer classification based on gene expression data. According to the results of the experiments over a six distinct benchmarking dataset spanning 27 diagnostic categories, it reveals that the fuzzy k-NN algorithm promotes the accuracy of cancer classification to a certain degree. Results also encourage the use of this fuzzification technique on similar problems in computational biology.

1. INTRODUCTION
Cancer, defined as the plaque of our time and is the major cause of death. Therefore, it is inevitable for the information technologies to give the necessary support to medicine in such an important and popular subject.

Gene expression data obtained from DNA microarray experiments are useful in the diagnosis and prognosis of diseases. Additionally, the DNA microarray technique allows measuring rapidly and simultaneously a great number of genes in different tissue sample. This technique generally produces large datasets with thousands of gene expression values but only few samples are available in order to make classification. It is being used especially in two areas in medicine; cancer and infectious diseases [1]. In the last decade, numerous works have been conducted to develop proper classification methods to recognize cancerous and normal tissue by analyzing microarray data [2, 3, 4].

From the classification point of view, one of the most important things is being aware of the strengths and weakness of available classification methods. Although lots of prior research had established accurate models for cancer diagnosis, the corresponding studies are limited in terms of the number of datasets and types of cancer involved [5, 6, 7]. Furthermore, it is noticed that successful methods suffer from the computational complexity and implementation difficulties of the proposed methodology [8].

In the last years, many approach applied k-nearest neighbors (k-NN) as a classification algorithm for microarray gene expression cancer diagnosis such as [4, 6, 8, 9]. Although the traditional k-NN algorithms are a good choice for this situation, one of the difficulties in utilizing these techniques is that all the labeled samples are given equal importance while deciding the cancer class of the gene expression sample under consideration and once a class has been assigned to a sample, there is no indication of its confidence in a particular class. This problem arises more seriously when the density of training data is uneven as in our case of experimental data obtained gene expression profiles. To solve this problem, recent studies proposed
to adopt the theory of fuzzy sets, constructing a new membership function based on the similarities between given samples and utilized this fuzzification technique on many problems such as document classification [10]. In this study, we attempted to use the fuzzy k-NN method for cancer diagnosis with microarray gene expression and presented a comparative study between the classification performance of traditional k-NN and fuzzy k-NN algorithms in terms of prediction accuracy over the experiments on several distinct benchmarking dataset. We believe that the current approach may not only play an important complementary role to previous powerful methods [8] but can also be alternative to the current solutions because of its simplicity and fast processing time.

2. MATERIALS AND METHODS

2.1 k-Nearest Neighbors

The k-nearest neighbors (k-NN) is one of the oldest and simplest non-parametric classification algorithms [11, 12]. Despite its simplicity, it has many advantages and it may give competitive performance compared to many other classification methods. And this algorithm has been successfully applied to a broad range of problems and has numerous variations.

This algorithm classifies an unknown sample by comparing it to its k nearest neighbors among a set of known samples, where k is a positive and typically small integer. Firstly, the distances between the unknown sample and all the training samples which mean a set of known samples are calculated. To calculate the distance, various techniques can be used such as Euclidean distance, which is used in the current study, or Mahalanobis distance. After all the distances are calculated, they are sorted and nearest k samples are determined. With a voting scheme among them, i.e. using the majority of the class of nearest neighbors, the class of the unknown sample is assigned.

2.2 Fuzzy k-Nearest Neighbors

The fuzzy k-NN classification method was designed by [10] and this method can often improve performance in biological and medical data classification problems.

This method assigns memberships of samples to various classes rather than a particular class as in k-NN method. The following relationship provides class memberships to the sample as a function of the sample’s distance from its k-NN training samples:

\[
    u_i(x) = \frac{\sum_{j=1}^{k} u_i(x^{(j)})(\|x-x^{(j)}\|^{2/m})}{\sum_{j=1}^{k}(\|x-x^{(j)}\|^{2/m})} \quad i = 1, \ldots, c
\]

where \( m \) is a fuzzy strength parameter between 1 and 2, that determines how the distance is weighted. The variable \( k \) is the number of nearest neighbors. And \( u_i(x) \) is the membership of the test sample \( x \) to class \( i \). \( \|x-x^{(j)}\| \) is the distance between the test sample \( x \) and its nearest training samples \( x^{(j)} \). To calculate the distance, various techniques can be used like k-NN. In this study, we used Euclidean distance. \( u_i(x(j)) \) is the membership value of the j-th neighbor to the i-th class which can be defined in several ways. The “crispest” way is to give them complete membership in their own class and non-membership in all other classes. In other words; it assigns 1 if the sample belongs to class otherwise 0. Alternatively, a more fuzzy method can be used to assign the training samples memberships based on the distance from the class mean. After calculating the membership for the test sample, it is assigned to the class with highest membership. In the present study, we used the “crispest” way.

2.3 Datasets

The datasets that we used in this study described in Table 1. All datasets were produced by oligonucleotide-based microarray technology. We used 6 datasets which has 2-9 distinct categories, 50-102 samples and 2308-10509 genes. All datasets are downloaded from http://www.gems-system.org.

Brain_Tumor1 dataset contains 5 human brain tumor types which are medulloblastoma, malignant glioma, AT/RT, normal cerebellum, PNET. Brain_Tumor2 includes classic glioblastomas, classic anaplastic oligodendrogliomas, non-classic glioblastomas, non-classic anaplastic oligodendrogliomas which means 4 malignant glioma type.

In Leukemia dataset, the categories are acute myelogenous leukemia (AML), acute lymphoblastic leukemia (ALL) B-cell and acute lymphoblastic leukemia (ALL) T-cell. Furthermore, SRBCT dataset involves EWS, RMS, BL, NB tumors.

In Prostate_Tumor dataset, there are only two categories which are tumor and normal. Finally, 9Type_Tumor dataset contains 9 kinds of human tumor types which are NSCLC, Colon, Breast, Ovary, Leukemia, Renal, Melanoma, Prostate, CNS.
Table 1. The datasets that used in this study

<table>
<thead>
<tr>
<th>Dataset’s Name</th>
<th>Dataset’s Task</th>
<th>Number of Category (Class)</th>
<th>Number of Samples</th>
<th>Number of Genes</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain_Tumor1</td>
<td>5 human brain tumor types</td>
<td>5</td>
<td>90</td>
<td>5920</td>
<td>[13]</td>
</tr>
<tr>
<td>Brain_Tumor2</td>
<td>4 malignant glioma types</td>
<td>4</td>
<td>50</td>
<td>10367</td>
<td>[14]</td>
</tr>
<tr>
<td>Leukemia</td>
<td>Acute lymphoblastic leukemia (ALL), B-cell and T-cell; acute myelogenous leukemia (AML)</td>
<td>3</td>
<td>72</td>
<td>5327</td>
<td>[4]</td>
</tr>
<tr>
<td>SRBCT</td>
<td>Small, round blue cell tumors</td>
<td>4</td>
<td>83</td>
<td>2308</td>
<td>[15]</td>
</tr>
<tr>
<td>Prostate_Tumor</td>
<td>Prostate tumor and normal tissue</td>
<td>2</td>
<td>102</td>
<td>10509</td>
<td>[16]</td>
</tr>
<tr>
<td>9Type_Tumor</td>
<td>9 kinds of human tumor types</td>
<td>9</td>
<td>60</td>
<td>5726</td>
<td>[17]</td>
</tr>
</tbody>
</table>

2.4 Measurement Accuracy

For all genes in datasets we applied k-nearest neighbor and fuzzy k-nearest neighbor algorithm and we determined each gene’s category by using the other genes in the dataset as a training sample. Then we calculated the prediction accuracy according to the relationship (2) for these two classification methods.

\[
\text{accuracy} = \frac{\sum_{s=1}^{N} p(s)}{N} \times 100
\]  

(2)

In this relationship; \(p(s)\) value is assigned 1 if s-th gene’s category (class) determined correctly, otherwise it is assigned 0. And \(N\) is the total number of genes in a dataset.

3. RESULTS AND ANALYSES

For k-NN algorithm; tests has been done with various values of nearest neighbor k. We took k at least 5 and for each datasets the best results were achieved when \(k=5\). When we increased k, the accuracies went down.

On the other hand; for fuzzy k-NN algorithm; the tests has been done with several number of fuzzy strength parameter \(m\) and various values of nearest neighbor k. The total prediction accuracy on the number of nearest neighbors, k, is shown in Figure 1 and 2. It can be seen that the prediction accuracy does not change significantly when k is 5 or 10. Additionally; in most of the datasets the best accuracies were achieved while \(m=1.2\), and when we increased \(m\), it can be seen that accuracies went down.

To make a comparison; in Table 2, the accuracies that were obtained while \(k=5\) and \(m=1.2\) are shown. And finally; it can be inferred from Table 2 that fuzzy k-NN algorithm’s classification performance for microarray gene expression cancer diagnosis is better than the k-NN algorithm’s classification performance.

Table 2. Accuracies

<table>
<thead>
<tr>
<th></th>
<th>k-NN</th>
<th>Fuzzy k-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain_Tumor1</td>
<td>84.4%</td>
<td>85.6%</td>
</tr>
<tr>
<td>Brain_Tumor2</td>
<td>74.0%</td>
<td>80.0%</td>
</tr>
<tr>
<td>Leukemia</td>
<td>88.9%</td>
<td>90.3%</td>
</tr>
<tr>
<td>SRBCT</td>
<td>90.4%</td>
<td>94.0%</td>
</tr>
<tr>
<td>Prostate_Tumor</td>
<td>80.4%</td>
<td>82.4%</td>
</tr>
<tr>
<td>9Type_Tumor</td>
<td>51.7%</td>
<td>58.3%</td>
</tr>
</tbody>
</table>
4. CONCLUSION

In this paper, we have presented an approach for classification of high dimensional DNA microarray gene expression data. Additionally, we have presented a comparative study of two classification methods that applied to the DNA microarray gene expression cancer diagnosis. In this work we investigated the following two classification algorithms: k-Nearest Neighbors and Fuzzy k-Nearest Neighbors. We compared fuzzy k-NN algorithm's classification performance with k-NN algorithm’s performance by calculating prediction accuracies for six distinct datasets spanning 27 diagnostic categories.

The results demonstrate that fuzzy k-NN approach can attain higher classification accuracy than the k-NN approach for all datasets that we used. In conclusion, we believe that fuzzy k-NN classification algorithm can play an important complementary role and can be helpful because of its simplicity for microarray gene expression cancer diagnosis.

REFERENCES


