WEIGHTED NEIGHBORHOOD CLASSIFIER FOR THE CLASSIFICATION OF IMBALANCED TUMOR DATASET

SHU-LIN WANG
School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China
Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Science, Hefei, China
slwang@hotmail.com

XUELING LI
Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Science, Hefei, China
xlli@iim.ac.cn

JUN-FENG XIA
Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Science, Hefei, China
School of Life Science, University of Science and Technology of China, Hefei, Anhui 230027, China
jfxia@mail.ustc.edu.cn

XIAO-PING ZHANG*
Dept. of Electrical & Computer Engineering, Ryerson University, Toronto, Ontario, Canada, M5B 2K3
xzhang@ee.ryerson.ca

Machine learning is widely applied to gene expression profiles based molecular tumor classification, but sample imbalance problem is often overlooked. This paper proposed a subclass-weighted neighborhood classifier to address the imbalanced sample set problem and a novel neighborhood rough set model to select informative genes for classification performance improvement. Experiments on three publicly available tumor datasets demonstrated that the proposed method is obviously effective on imbalanced dataset with obscure boundary between two subtypes and informative gene selection and it can achieve higher cross-validation accuracy with much fewer tumor-related genes.

Keywords: Gene expression profiles; molecular tumor classification; imbalanced dataset; weighted neighborhood classifier; Kruskal-Wallis rank sum test; neighborhood rough set model.

*Corresponding author.
1. Introduction

DNA microarray promises a novel tumor subtype diagnosis based on cancer-specific molecular markers. Modern medicine indicates that many or even all human diseases may be accompanied by changes in expression levels of specific genes, so a set of differentially expressed genes can discriminate one subtype of tumors from others. Compared with traditional tumor diagnosis, which is prone to human error and cannot discriminate different tumor subtypes with similar histopathology, gene expression profiles (GEP) based tumor diagnosis is more accurate and reliable. Therefore, GEP can serve as promising tools for tumor classification, diagnosis and therapeutics.

In fact, research on GEP-based tumor diagnosis has been active in biomedicine. Numerous researchers have applied machine learning method to the GEP-based tumor classification. For example, unsupervised methods including tumor sample clustering by non-negative matrix factorization for gene selection, self-organizing maps, and supervised methods, such as artificial neural networks, support vector machines (SVMs), rough set theory and k-nearest neighbor (k-NN), have been successfully applied to tumor classification. However, many learning algorithms overlook the sample imbalance problem in a tumor dataset. And the imbalance problem may have a great influence on gene selection and practical tumor diagnosis.

Furthermore, feature extraction such as independent component analysis can improve the performance of classification model, but gene selection has more biological meanings. However, tumor-related gene selection poses another challenge in tumor classification due to the curse of dimensionality that the number of genes far exceeds the number of samples. If a sample contains \( n \) genes, \( 2^n \) candidate gene subsets must be checked to select optimal gene subsets with the highest accuracy and the least cardinal number, which can improve the prediction performance of classification model, otherwise too many redundant or irrelevant genes can degrade the classification accuracy, but this is a NP-hard problem in the case of no prior knowledge. Particularly, the selected gene subsets may have important biomedical meanings and be applied to the discovery of cancer biomarkers and drug targets. Many gene selection algorithms such as sequential forward search (SFS), support vector machine based recursive feature elimination, sequential forward floating search (SFFS) algorithm and Markov blanket-embedded genetic algorithm are successfully applied to tumor classification. However, these methods are very time-consuming because each candidate subset has to be evaluated before selected.

In recent years, a rough set theory that is suited to deal with the uncertainty of decision system has been developed, and successfully applied to GEP-based gene selection. Fang et al. utilized rough set approach to predict leukemia and found eight tumor-related genes and eight informative rules on the leukemia dataset. However, before gene reduction by classical rough set method, gene expression values must be discretized, which can lead to information loss before classification. To avoid
this problem, Hu et al.\textsuperscript{18} proposed a neighborhood rough set model based on δ neighborhood and neighborhood relations, which directly processes real values without discretization as demonstrated by experiments on UCI datasets.

This paper applied the neighborhood rough set model to gene selection and proposed a weighted neighborhood classifier (W-NEC) to address imbalanced tumor dataset problem, which combines both the neighborhood classifier (NEC)\textsuperscript{18,19} and the sample imbalance processing method in Ref. 20. Neighbor-weighted $k$-nearest neighbor (NW$^k$-NN) is proposed by Tan\textsuperscript{20} for categorization of imbalanced text corpus. Our W-NEC and NW$^k$-NN have much in common for subclass decision in that both W-NEC and NW$^k$-NN estimates the subclass of an unknown sample from its weighted neighbors. However, there is difference between the two classifiers. For NW$^k$-NN classifier, the neighbor number of an unknown sample $s$ is a constant $k$, while for W-NEC classifier, the number of training samples within δ neighborhood of an unknown sample $s$ vary with different sample $s$. On the contrary, for W-NEC, the range of δ$(s)$ is a constant in the process of classification, while for NW$^k$-NN, the range of $k$-nearest neighbors varies with different unknown sample $s$. The purpose of our method is both to alleviate the effect of sample imbalance and improve the effectiveness of informative gene selection on tumor classification by using the neighborhood rough set model.

2. Tumor Classification Model

2.1. Problem description
Let $G = \{g_1, \ldots, g_n\}$ be a set of genes and $S = \{s_1, \ldots, s_m\}$ be a set of samples. The gene expression matrix can be represented as $X = \{x_{i,j}\} | 1 \leq i \leq m, 1 \leq j \leq n\}$, where $x_{i,j}$ is the expression level of gene $g_j$ on sample $s_i$, and usually $n \gg m$. The matrix $X$ is composed of $m$ row vectors $s_i \in R^n, i = 1, 2, \ldots, m$. Here, $m$ is the number of samples and $n$ is the number of genes. Each vector $s_i$ in the gene expression matrix may be thought of as a point in $n$-dimensional space, and each of the $n$ columns consists of an $m$-element expression vector for a single gene. Our task is to classify all samples into subclass set $\Omega : \{\omega_1, \omega_2, \ldots, \omega_c\}$, and each subclass $\omega_i$ is labeled as $l_i$. Here $L = \{l_1, l_2, \ldots, l_c\}$ denotes a label set. Suppose $\omega_i, i \in \{1, \ldots, c\}$ be the subset of sample set $S$, satisfying $\omega_i \cap \omega_j = \phi, i \neq j$, $\bigcup_{i=1}^{c} \omega_i = S$, which means that each vector ideally belongs to only one subclass $\omega_i, i \in \{1, \ldots, c\}$.

2.2. Neighborhood-based rough set model
Let $NDT = \langle S, G \cup D, V, f \rangle$ be a neighborhood decision table, where $S = \{s_1, \ldots, s_m\}$ is a nonempty sample set called sample space, and $G = \{g_1, \ldots, g_n\}$ is a nonempty set of genes also called condition attributes, $D = \{L\}$ is an output variable called decision attribute, $V_\alpha$ is a value domain of attribute $\alpha \in G \cup D$, $f$ is an information function, $f : S \times (G \cup D) \rightarrow V$, where $V = \bigcup_{\alpha \in G \cup D} V_\alpha$. 

Weighted Neighborhood Classifier for the Classification of Imbalanced Tumor Dataset 261
Given \( \forall s_i \in S \) and \( B \subseteq G \), the neighborhood \( \delta_B(s_i) \) of \( s_i \) in the subspace \( B \) is defined as \( \delta_B(s_i) = \{ s_j | s_j \in S, \Delta_B(s_i, s_j) \leq \delta \} \), where \( \delta \) is a threshold value, and \( \Delta_B(s_i, s_j) \) is a metric function in subspace \( B \). Let \( s_1 \) and \( s_2 \) be two samples in \( n \)-dimensional space \( G = \{ g_1, \ldots, g_n \} \), \( f(s_i, g_i) \) denotes the value of sample \( s \) in the \( i \)-th dimension \( g_i \). Then Minkowsky distance is defined as

\[
\Delta_p(s_1, s_2) = \left( \sum_{i=1}^{n} |f(s_1, g_i) - f(s_2, g_i)|^p \right)^{1/p},
\]

where (1) if \( p = 1 \), it is called Manhattan distance \( \Delta_1 \); (2) if \( p = 2 \), it is called Euclidean distance \( \Delta_2 \); (3) if \( p = \infty \), it is called Chebychev distance.

Given a neighborhood decision table \( NDT, X_1, X_2, \ldots, X_c \) are the sample subsets with decisions 1 to \( c \), \( \delta_B(x_i) \) is the neighborhood information granules including \( x_i \), and generated by gene subset \( B \subseteq G \), then the lower and upper approximations of the decision \( D \) with respect to gene subset \( B \) are respectively defined as

\[
\text{Lower}(D, B) = \bigcup_{i=1}^{c} \text{Lower}(X_i, B),
\]

\[
\text{Upper}(D, B) = \bigcup_{i=1}^{c} \text{Upper}(X_i, B),
\]

where \( \text{Lower}(X, B) = \{ x_i | \delta_B(x_i) \subseteq X, x_i \in S \} \) is the lower approximations of the sample subset \( X \) with respect to gene subset \( B \), and is also called the positive region denoted by \( \text{Pos}(D, B) \), and \( \text{Upper}(X, B) = \{ x_i | \delta_B(x_i) \cap X \neq \emptyset, x_i \in S \} \) is the upper approximations of the sample subset \( X \) with respect to gene subset \( B \). The decision boundary region of \( D \) to \( B \) is defined as

\[
\text{BN}(D, B) = \text{Upper}(D, B) - \text{Lower}(D, B).
\]

The dependency degree of \( D \) to \( B \) is defined as the ratio of consistent objects \( \gamma(D, B) = \text{Card}(\text{Pos}(D, B)) / \text{Card}(S) \), here we define \( \gamma(D, \emptyset) = 0 \), and \( \text{Card}(S) \) denotes the cardinal number of sample set \( S \). Let \( a \in B \), then the significance of a gene is defined as

\[
\text{SIG}(a, D, B) = \gamma(D, B) - \gamma(D, B - a)
\]

or as \( \text{SIG}(a, D, B) = \gamma(D, B \cup a) - \gamma(D, B) \).

2.3. Improved neighborhood classifier

Hu\textsuperscript{18,19} designed a uniform theoretic framework for neighborhood-based classifier (NEC) based on the neighborhood rough set model, which integrates feature selection with classifier construction, and classifies an unknown sample \( s \) in the selected subspace based on the majority subclass in the neighborhood of the unknown sample \( s \). Their experimental results on UCI datasets show that NEC classifier outperforms
the popular CART learning algorithm and \(k\)-NN classifier, and a little weaker than SVM based on three norms. The NEC classifier algorithm can be described in Algo. 1.

**Algorithm 1**: Neighborhood classifier (NEC)

**Input**: Training set \(\{S, G \cup \{L\}\}\), unknown sample \(s\), threshold \(\delta\), the selected norm.

**Output**: Class label of sample \(s\).

1: Compute the distance between \(s\) and \(s_i \in S\) with the selected norm.
2: Find the samples in the neighborhood \(\delta(s)\) of sample \(s\).
3: Find the class \(\omega_i\) with the majority training samples in \(\delta(s)\), i.e. \(\omega^* = \arg \max_{\omega_i \in \Omega} (\text{card}(\omega_i \cap \delta(s)))\), where \(i \in \{1, \ldots, c\}\) and \(\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}\).
4: Label the unknown sample \(s\) with the label \(l^* \in L\) corresponding to the class \(\omega^*\).

An example with an imbalanced two-subclass dataset for NEC is shown in Fig. 1, from which we can see that the unknown sample is assigned to Class 1 label because within \(\delta\) neighborhood of the unknown sample represented by square there are nine samples of Class 1 and only four samples of Class 2, while the unknown sample may belong to Class 2. In NEC algorithm, threshold \(\delta\) can be computed as Eq. (6).

\[
\delta = \min(\Delta(x_i, s)) + w \cdot \text{range}(\Delta(x_i, s)), \quad 0 < w < 1,
\]

where \(x_i (i = 1, \ldots, m)\) is the training sample set, \(\min(\Delta(x_i, s))\) denotes the minimal value of distance between \(x_i\) and the unknown sample \(s\); \(\text{range}(\Delta(x_i, s))\) denotes the value range of \(\Delta(x_i, s)\), and \(w\) is a weight value.

However, the NEC classifier overlooks sample imbalance problem. When classifying the imbalanced tumor dataset, NEC classifier tends to assign the subclass label of majority samples in the neighborhood set \(\delta(s)\) to each unknown sample \(s\). As a result, the more samples a subclass has, the higher the classification accuracy for this subclass is. This can lead to the decrease of classification accuracy in the case of that the number of samples in a subclass is very small. Therefore, based on the idea in

![Fig. 1. A diagram with two imbalanced subclasses for neighborhood classifier.](image-url)
Ref. 20 that \( k \)-NN classifier is improved to deal with imbalanced dataset, we proposed an improved NEC classifier to alleviate the effect of sample imbalance in dataset on classification performance. We called this improved NEC classifier as weighted NEC (W-NEC).

Furthermore, we designed two weighting methods to evaluate subclass weight for each subclass \( \omega \in \Omega \) in dataset. The two methods are represented by Eqs. (7) and (8) called \( \text{Weight}_1 \) and \( \text{Weight}_2 \) methods, respectively.

\[
\text{Weight}_1(i) = \frac{1}{1/\left(\frac{\text{card}(\omega_i)}{\min\{\text{card}(\omega_j) | j = 1, \ldots, c\}\}^{1/\text{Exponent}}\right)}, \tag{7}
\]

\[
\text{Weight}_2(i) = 1 - \log \text{sigm}(\text{card}(\omega_i)/\min\{\text{card}(\omega_j) | j = 1, \ldots, c\}) \tag{8}
\]

where \( i \in L \) and \( \text{sigm}(x) = 1/(1 + \exp(-x)) \) is sigmoid function. W-NEC classifier can be obtained by only revising Step 3 in the NEC algorithm as follows.

\[
\omega^* = \arg \max_{\omega_j \in \Omega} (\text{Weight}(j) \times \text{card}(\omega_j \cap b(s))). \tag{9}
\]

Assuming that all subclasses in tumor dataset are evenly balanced, \( \text{Weight}_1(i) \), \( i \in L \) is always equal to 1 and \( \text{Weight}_2(i), i \in L \) is always approximately equal to a constant 0.2689. In this case, W-NEC classifier is the same as the NEC classifier. Therefore, applying \( \text{Weight}_1 \) or \( \text{Weight}_2 \) method to tumor dataset, we assigned small weight to the majority subclass and vice versa. Intuitively, the goal of alleviating the effect of sample imbalance problem on the tumor classification is attained. However, we did not theoretically determine which weighting method of the two is the best for tumor classification and whether the W-NEC classifier is superior to the NEC classifier or not.

### 2.4. Algorithm framework

Although the classic rough set theory based feature reduction is effective, its classification performance is not obviously improved compared with other gene selection methods due to the information loss from the discretization of gene expression values before tumor classification. Therefore, the neighborhood rough set model is introduced to tumor classification, in which the discretization procedure can be omitted, so no information loss occurs before gene reduction. The framework of our algorithm consists of three steps.

**Step 1:** Adopt Kruskal–Wallis rank sum test\(^{21}\) to calculate the \( p \)-value for each gene, and then rank all genes according to their \( p \)-values by ascending order, and lastly simply selecting \( P \) top-ranked genes.

**Step 2:** Adopt the forward attribute reduction based on neighborhood model (FARNeM)\(^{18,19}\) as shown in Algo. 2 to conduct gene reduction using the \( P \) top-ranked genes, and then the informative gene subsets are obtained.
Algorithm 2: FARNeM

Input: \( NDT = \langle S, G \cup D, V, f \rangle \) and \( \delta \) neighborhood

Output: \( \text{red} \) //reduction of gene set

1: \( \forall a \in A \), compute neighborhood relation \( N_a \);
2: \( \text{red} = \emptyset \); //\( \text{red} \) is the pool to contain the selected genes
3: for each \( \forall a_i \in A - \text{red} \)
   Compute \( SIG(a_i, D, \text{red}) = \gamma(D, \text{red} \cup a_i) - \gamma(D, \text{red}) \);
4: Select \( a_k \), and satisfy \( SIG(a_k, D, \text{red}) = \max_i(SIG(a_i, B, \text{red})) \);
5: if \( SIG(a_k, \text{red}, D) > 0 \)
   \( \text{red} = \text{red} \cup a_k \);
   go to 3
else return \( \text{red} \).

In FARNeM algorithm, \( \gamma(D, B) = \text{Card(Lower}(D, B))/\text{Card}(S) \) denotes the dependency of decision feature \( D \) to condition feature \( B \), and \( SIG(a, B, D) = \gamma(D, B \cup a) - \gamma(D, B) \) denotes the significance of feature \( a \) with respect to \( B \). The FARNeM appends the optimal informative gene into the reduction set in each loop until the dependence does not increase. Given a \( \delta \) value, one optimal gene subset can be selected by running the FARNeM only once. Usually, different optimal gene subset corresponds to different \( \delta \) value (it is possible that the same gene subsets correspond to different \( \delta \) value in some cases), so the selected gene subsets are diverse.

Step 3: Classify tumor samples by W-NEC and evaluate with unknown samples to obtain the classification accuracy and gene subsets with the highest accuracy.

3. Experiments

3.1. Tumor datasets

In our experiments, we adopted three tumor datasets: Leukemia dataset, Colon dataset and Small Round Blue Cell Tumor (SRBCT) dataset. The leukemia and colon datasets that contain only two subclasses are described in Table 1.

SRBCT dataset which contains 88 samples and 2,308 genes for each sample were downloaded from http://research.nhgri.nih.gov/microarray/Supplement. According to the suggestion in original paper, the datasets were divided into 63 training samples and 25 test samples. The 63 training samples contain 23 Ewing family of tumors.

<table>
<thead>
<tr>
<th>Tumor dataset</th>
<th>#Gene</th>
<th>#Sample</th>
<th>Subtype 1</th>
<th>Subtype 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>7129</td>
<td>72</td>
<td>47 (ALL)</td>
<td>25 (AML)</td>
</tr>
<tr>
<td>Colon</td>
<td>2000</td>
<td>62</td>
<td>40 (Tumor)</td>
<td>22 (Normal)</td>
</tr>
</tbody>
</table>
(EWS), 20 rhabdomyosarcomas (RMS), 12 neuroblastomas (NB), and eight Burkitt lymphomas (BL). The 25 test samples contain six EWSs, five RMSs, six NBs, three BLs. Five non tumor-related samples are removed in our experiments. The imbalance of these three datasets is obvious, i.e., one subtype’s number is two to three times of the other one, even though it is not the extreme of imbalance in dataset.

3.2. Parameters optimization

In our experiments, four parameters need to be optimized. First, 4-fold cross-validation (CV) method was adopted to evaluate our classification model on the three datasets, respectively. Second, in forward attribute reduction algorithm, the range of \( \delta \) neighborhood is from 0.01 to 1 with the increment of 0.01, and for each \( \delta \) value we can obtain a gene subset. Third, the range of \( w \) in Eq. (6) is from 0 to 1 increased by 0.01. Last, the range of parameter \( \text{Exponent} \) in Eq. (7) is from 1 to 6 by 0.5. Our final goal is to find the best combination of these parameters.

4. Experimental Results and Analysis

4.1. Comparison of W-NEC with NEC and other related classifiers

To validate the effectiveness of Kruskal–Wallis rank sum test (KWRST) and to evaluate the classification ability of W-NEC, we respectively selected from 5 to 100 top-ranked genes to be used as the input of the two classifiers. Figures 2–4 show the classification accuracy of W-NEC with Weight\(_1\) and NEC with different numbers of top-ranked genes on the three tumor datasets, respectively. We found that the accuracy obtained by W-NEC on the three datasets is equal to or higher than that by NEC. Especially, for the colon dataset that is difficult to be classified, W-NEC
classifier is obviously superior to NEC in the classification accuracy rate. We concluded that our W-NEC obviously improves the classification performance on the dataset with complex boundary between its every two subclasses.

To compare the classification ability among the three classifiers: NEC, W-NEC with Weight1 and W-NEC with Weight2, we adopted principal component analysis (PCA) to extract principal components (PCs) from 25 or 50 top-ranked genes to be used as the input of the classifiers. The accuracy on the three classifiers is shown in Table 2, from which we can see that W-NEC with Weight1 is better than or as well as NEC classifier, but W-NEC with Weight2 is not consistently superior to NEC classifier, which indicates that W-NEC with Weight2 is not steady in performance.

After selecting the gene subsets with minimum cardinal number by using FARNeM, we adopted W-NEC with Weight1 to evaluate them. Table 3 shows

Fig. 3. The comparison of classification accuracy on leukemia dataset using W-NEC and NEC classifier.

Fig. 4. The comparison of classification accuracy on SRBCT dataset using W-NEC and NEC classifier.
various selected gene subsets and their corresponding 4-fold CV accuracy by using W-NEC with $Weight_1$ on each gene subset, from which we can see that seven genes can obtain 95.16% accuracy on the colon dataset, and two genes can obtain 100% accuracy on the leukemia dataset, and six genes can obtain 100% accuracy on SRBCT dataset.

Comparison with some other related tumor classification methods in accuracy is shown in Table 4, which indicates that our approach obtains almost the best results
among the listed approaches. For the colon dataset, our approach outperforms the other ones except the JCFO (linear kernel), and for the leukemia and SRBCT datasets our approach is also equal to or better than other methods in accuracy. Moreover, our gene selection method can obtain near minimum gene subset with a range of $2^{C_0}$ genes for classification (shown in Table 3). For example, Khan et al. selected 96 informative genes obtaining 100% accuracy and Deutsch adopted evolutionary algorithm and $k$-NN classifier to select 12 informative genes obtaining 100% classification accuracy. Experiments on the three datasets also show that gene selection by a neighborhood rough set model obviously outperforms that by a classic rough set theory. Therefore, compared with some other related works, our method performs well in imbalanced dataset and is simple and easy to interpret biomedically.

4.2. Function analysis of the selected genes

Biologically experimental results also prove that most of the selected gene subsets are related to carcinogenesis and tumor histogenesis. For leukemia, Zyxin is a gene correlated to leukemia of ALL and Zyxin protein possesses LIM domain which is known to interact with leukemogenic bHLH proteins. CD33 is a marker of leukemia (ALL) and is also a member of the sialic acid-binding immunoglobulin-like lectin (Siglec) family of inhibitory receptors and a therapeutic target for acute myeloid leukemia. The NM23 gene is isolated as a metastasis suppressor gene that exhibits low expression in high-level metastatic cancer cells. Its gene is related to the
prognosis of acute myelogenous leukemia (AML) and non-Hodgkin’s lymphoma (NHL). For colon tumor, gene M36634 (Vasoactive Intestinal Peptide, VIP) is closely related to colon tumor. VIP can enhance the proliferation of colon cell lines LoVo and HT29 and AOM (adenosine 3’, 5’-monophosphate, ornithine decarboxylase) induced colon tumor growth, as well as the increase of ornithine decarboxylase (ODC) mRNA. GCAP-II gene has a high level of expression in human colon, which indicates a pivotal role in cGMP-mediated functions of the colon, and it stimulates cGMP generation in T84 cells (colonic carcinoma cell line). The heterogeneous nuclear ribonucleoprotein K (hnRNP K) protein is an RNA-binding protein involved in many processes that regulate gene expression. K protein is upregulated in the malignant processes and has been shown to modulate the expression of genes involved in mitogenic responses and tumorigenesis. A single G-to-A base substitution at position 274 in this gene was found in tumors and surrounding mucosa, but not in individuals that had no colorectal tumor.

4.3. The effects of different parameters on tumor classification

We studied the effects of different parameters on the performance of our classification models. With Exponent values ranging from 1 to 6.8 with an increment of 0.2, the classification accuracy do not vary much, which indicates that the classification accuracy is not sensitive to the Exponent value, so the Exponent value in W-NEC with Weight1 method can be fixed to 1 for the three tumor datasets.

For every δ neighborhood value in FARNeM algorithm, we can obtain an optimal gene subset. Therefore, when δ neighborhood takes values from 0.01 to 1 with the increment of 0.01, 100 gene subsets can be obtained. Then we adopted W-NEC classifier to evaluate the 100 gene subsets to obtain their 4-fold CV accuracy. Figures 5 and 6 show the classification accuracy with different δ neighborhood values on the three datasets respectively, from which we can see that the classification
accuracy rate is sensitive to δ neighborhood value. We found that the range of optimal δ neighborhood value is different on different datasets and that the optimal or near optimal δ neighborhood take values within the range [0.5–0.9] for the three datasets.

5. Conclusions and Future Work

Usually, complex methods are not necessarily better than the simplest one in performance and the loss of biomedical meaning deriving from the utilization of overcomplex methods may be not significantly be counterbalanced by a little improvement in prediction performance. This paper proposed a simple subclass-weighted classifier, W-NEC, that is improved from NEC to deal with the problem of sample imbalance in tumor dataset. Two subclass-weighted methods, Weight 1 and Weight 2 methods, are designed, respectively. The experiments on three tumor datasets show that W-NEC with Weight 1 is superior to NEC while W-NEC with Weight 2 is not always superior to NEC. Especially, W-NEC can improve the classification accuracy of imbalanced tumor datasets when the boundary between every two subtypes in the dataset is not clear.

The main advantage of the proposed method is its simplicity to biomedical interpretation, but its disadvantage is that there are two parameters to be set, and finding the best combination of the two parameters is a time-consuming task. Therefore, our future work is to find the best combination of the two parameters without increasing the computational cost. Another work is to design better subclass-weighted method to further improve the classification performance of imbalanced dataset.

Acknowledgment

This work was supported by the National Science Foundation of China (60973153 and 30700161), the Knowledge Innovation Program of the Chinese Academy of Sciences (0823A16121) and the China Postdoctoral Science Foundation (20090450707).
References


11. H. L. Huang, C. C. Lee and S. Y. Ho, Selecting a minimal number of relevant genes from microarray data to design accurate tissue classifiers, BioSystenms 90 (2007) 78–86.


