Bioinformatics Advance Access published August 6, 2009

Gene expression

Methods for Labeling Error Detection in Microarrays Based on the Effect of Data Perturbation on the Regression Model

Chen Zhang[1], Chunguo Wu[1], Enrico Blanzieri[2,*], You Zhou[1], Yan Wang[1], Wei Du[1], and Yanchun Liang[1,*]

[1]College of Computer Science and Technology, Jilin University, 130012, China.

[2]Department of Information and Communication Technology, University of Trento, 38050 Povo, Italy.

Associate Editor: Dr. Joaquin Dopazo

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on January 24, 2014

ABSTRACT

Motivation: Mislabeled samples often appear in gene expression profile because of the similarity of different subtype of disease and the subjective misdiagnosis. The mislabeled samples deteriorate supervised learning procedures. The LOOE-Sensitivity algorithm is an approach for mislabeled sample detection for microarray based on data perturbation. However, the failure of measuring the perturbing effect makes the LOOE-Sensitivity algorithm a poor performance. The purpose of this paper is to design a novel detection method for mislabeled samples of microarray, which could take advantage of the measuring effect of data perturbations.

Results: To measure the perturbing influence value, we define an index named perturbing influence value, PIV for short, based on the support vector machine regression model. The Column Algorithm (CAPIV), Row Algorithm (RAPIV) and Progressive Row Algorithm (PRAPIV) based on the PIV value are proposed to detect the mislabeled samples. Experimental results obtained by using six artificial datasets and five microarray datasets demonstrate that all proposed methods in this paper are superior to LOOE-Sensitivity. Moreover, compared with the simple SVM and CL-Stability, the PRAPIV algorithm shows an increase in precision and high recall.

Availability: The program and source code (in JAVA) are publicly available at http://ccst.jlu.edu.cn/CSBG/PIVS/index.htm

Contact: blanzier@dit.unitn.it, ycliang@jlu.edu.cn

1 INTRODUCTION

Microarrays are a powerful tool for high-throughput measurement of gene expression and more and more groups employ microarrays in cancer research (Edwin *et al*., 2000; Schramm *et al*., 2005; Alon *et al*., 1998; West *et al*., 2001; Pomeroy *et al*., 2002; Wong *et al*., 2003; Welsh *et al*., 2001). A number of methods based on classification have been proposed to discover the relationship between genes and tumors from the gene expression profiles (Antonov *et al*., 2004; Bø *et al*., 2002; Dudoit *et al*., 2002; Tusher *et al*., 2001). However, just as Zhang *et al*. (2006) reported, there are 10-15% samples mislabeled in a microarray, which are usually incurred by the similarity of different subtype of disease (Khan *et al*., 2001) and subjective misdiagnosis. The potential mislabeled samples would deteriorate classification accuracy seriously, especially for supervised learning procedures. Consequently, effective methods for labeling errors detection are necessary to improve the analysis procedure of microarray data.

Researchers (Brodley *et al*., 1999; Sanchez *et al*., 2003; Muhlenbach *et al*., 2004; Venkataraman *et al*., 2004;) proposed many approaches for detecting labeling errors when the number of features is usually smaller than the size of the samples. But most of existing approaches are not suitable for microarray data due to the characteristics of high dimensionality and small sample size.

There are some studies trying to identify the wrong labeled samples from microarray datasets exclusively. Kadota *et al* (2003) proposed a method based on Akaike's Information Criterion to detect outlier samples in the colon microarray data. Zhang *et al*. (2006) developed an iterative method in which the misclassification possibility is estimated for each sample in the training set and applied it on the breast cancer dataset with the subtypes of estrogen receptor status (ER1/ER2). However, these methods were mainly applied on only one microarray dataset. Malossini *et al*. (2006) proposed two data perturbing methods, named as the CL-Stability algorithm and the LOOE-Sensitivity algorithm, respectively, for labeling error detection. Both of the methods are general for binary-class microarray datasets. Malossini's methods are based on the construction of a Leave-One-Out Perturbed Classification (LOOPC) matrix in which the element LOOPC[i, j] is the predicted label of the sample x_j obtained with a SVM classifier while the sample x_j is excluded from the training dataset and the label of the sample x_i is flipped (since the labels are either +1 or -1 for binary-class issues). The CL-Stability algorithm is similar to a voting procedure in which if the number of dissenting votes against the original label for a sample is bigger than a threshold this sample will be considered as a suspect. The LOOE-Sensitivity algorithm focuses on flipped samples and tries to identify the wrong labeled samples according to the results with these samples flipped. Malossini's experimental results showed that the CL-Stability algorithm dominates the LOOE-Sensitivity algorithm in almost all situations. LOOE-Sensitivity tries to discover the difference be-

[*]To whom correspondence should be addressed.

© The Author (2009). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

tween the correct samples and wrong labeled samples from the results of the classification which are either +1 or -1, but we argue here that the discrete values are not capable to reflect the effect of the flipping. In another word, the failure of measuring the effect of the perturbation on the classifier could cause the poor performance of the LOOE-Sensitivity algorithm.

In this paper, the perturbing influence value (PIV) is defined to measure the effect of data perturbation on the regression model. Based on the PIV value, the Column Algorithm (CAPIV) and the Row Algorithm (RAPIV) are proposed adopting different perspectives on the effect of perturbing influence. In order to improve the RAPIV algorithm, the Progressive Row Algorithm based on the Perturbing Influence Values (PRAPIV) is proposed with a progressive correction procedure. We apply the proposed methods together with the simple SVM method and the CL-Stability algorithm on six artificial datasets and five microarray datasets. Experimental results show that the PRAPIV algorithm can increase precision and achieve high recall.

2 MATERIALS AND METHODS

In this section, we will firstly introduce the datasets used in the experiments, and then define the perturbing influence value to measure the effect of data perturbation on the regression model. At last, the proposed algorithms are described based on the perturbing influence value, respectively.

2.1 Datasets

Our goal is to detect the wrong-labeled samples in microarray data, so several 2-class microarray datasets on cancers are selected to evaluate the algorithms proposed in this paper. As a kind of supplement, some artificial datasets are designed to test the algorithms for different situations.

2.1.1 Microarray datasets Five 2-class microarray datasets listed in table 1 will be used in this paper,

Table 1. The 2-class microarray datasets

Datasets	Number of genes	Number of samples		Reference
		Class 1	Class 2	
Colon	2000	40(T)	22(N)	Alon *et al*., 1999
Breast	7129	25(ER+)	24(ER-)	West *et al*., 2001
CNS	7129	25(C)	9(D)	Pomeroy *et al*., 2002
Cervix	10692	25(T)	8(N)	Wong *et al*., 2003
Prostate	12626	24(T)	9(N)	Welsh *et al*., 2001

According to Alon *et al*. (1999) and West *et al*. (2001), the samples T2, T30, T33, T36, T37, N8, N12, N34, N36 in the Colon dataset and the samples 11, 14, 16, 31, 33, 45, 46, 40, 43 in the Breast dataset are identified as outliers with biological evidences. These two datasets can be used as real benchmark datasets to test the methods for labeling errors detection. And in order to enhance the reliability of the data source, these outliers are removed from Colon dataset and Breast dataset to make two pure datasets which are denoted by Colon-p and Breast-p respectively. We consider the other three datasets as pure datasets in which there is no wrong labeled sample.

2.1.2 Artificial datasets Six artificial datasets are constructed for providing more controlled conditions to evaluate the algorithms. In the artificial datasets, samples are labeled as either +1 or -1. Features in every sample are generated randomly. Some features are selected to be discriminating features which follow the Gaussian distributions. The mean μ and the standard deviation σ of the discriminating features are different depending on the sample labels. The other features are generated as white Gaussian noise. For samples labeled as +1, we take μ=3 and σ=1, and for samples labeled as -1, we take μ=-3 and σ=3. The number of features (FN), the number of samples (SN), the number of discriminating features (DFN), and the number of wrong-labeled samples (WLN) are given in Table 2

Table 2. The artificial datasets

	FN	SN	DFN	WLN
Test1	2000	30	5	4
Test2	2000	30	5	6
Test3	2000	30	5	10
Test4	2000	30	10	6
Test5	200	30	5	6
Test6	2000	50	5	6

2.2 Perturbing Influence Values

The LOOE-Sensitivity algorithm relies on the idea that the flipping will definitely affect the result of the SVM classification. The problem of LOOE-Sensitivity algorithm is that the discrete values (either +1 or -1) of the classification results are not capable to reflect the effect of flipping. It is easy to improve the algorithm using a regression model. To make the algorithm more sensitive to data perturbation, we introduce here an index called perturbing influence value (PIV) based on function regression models.

In order to describe our methods clearly, we only consider 2-class datasets here, and the idea in this paper can be generalized to solve the labeling error detection problem in multi-class datasets. Supposed that a 2-class microarray consists of p probes and n samples, x_i denotes the expression vector of the sample i, and y_i is the label value of the sample i where $y_i \in \{+1, -1\}$. We define a regression problem to describe the relationship between x_i and y_i. We assume x_i and y_i which is considered as continuous value here, are related by an unknown function f such that

$$y_i = f(x_i) + \varepsilon \tag{1}$$

Where f is a real-value function and ε is noise. The aim is to find the regression model $\hat{f}$ that is an estimate of f . Although there are many well-studied regression models, SVM regression model is used in this paper to construct the approximation $\hat{f}$ due to its good theoretical basis and application performance in a number of fields. (Smola *et al*., 1998).

Instead of the Leave-One-Out Perturbed Classification matrix defined by Malossini *et al.* (2006), a Leave-One-Out Perturbed Regression matrix (denoted by *Loopr*) is defined, where the element *Loopr*[*i*,*j*] is the regression value $\hat{f}(x_j)$ while the sample x_j is treated as testing sample and excluded from the training dataset, and the label y_i of sample x_i (which is included in training dataset) is flipped. In a binary classification problem, the label value y_i often takes value in {+1, -1}. Hence, if the label of sample x_i is flipped, y_i is multiplied by -1. According to the definition of *Loopr*, the element *Loopr*[*i*, *i*] is equivalent to the regression value without flipping.

In order to assess the behavior of the perturbation effect on the regression model, we state the following hypothesis:

Hypothesis 1: Flipping a correct sample can make the regression value further away from its true label (-1 or +1) and flipping a wrong labeled sample can make the regression value closer to its true label (-1 or +1).

Figure 1 shows the perturbing regression values (*Loopr*[*i*,T10], abusing of the index notation) of the tumor sample T10 labeled with +1 and its regression value without perturbation (*Loopr*[T10,T10], dashed line) in Colon-p dataset. If a sample without labeling error is flipped, the regression value of T10 will be further from +1 and closer to -1 than the regression value without flipping. It means that *Loopr*[i,T10] falls under the dashed line. It can be seen from Figure 1 that there are 35 regressed values falling

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on January 24, 2014

under and 17 ones above the dashed line. Hence, we could conclude that most of the samples satisfy Hypothesis 1.

In order to measure how much the flipping will affect the regression results, we define the perturbing influence value which is the difference between regression values before and after flipping. Because the elements in the *Loopr* matrix are continuous values instead of binary values (-1 or +1), they are much more sensitive to flipping the label of a single sample.

The perturbing influence value (PIV) of sample x_j under the flipping of sample x_i, which is denoted by q_{ij}, is defined as follow:

$$q_{ij} = Loopr[j, j] - Loopr[i, j] \tag{2}$$

Suppose that a correctly labeled sample x_i is flipped, the regression value of the sample x_j will be away from its original regression value before flipping. According to Hypothesis 1, if the label of x_j is +1, we expect that q_{ij} is positive; otherwise, q_{ij} is negative.

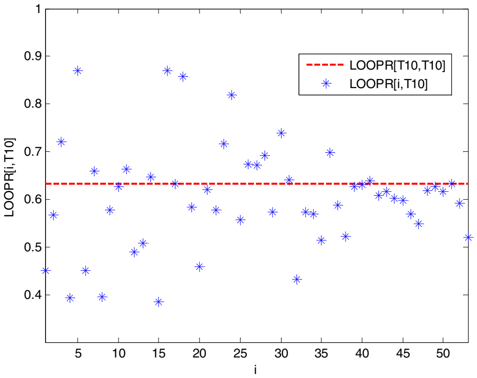


Fig. 1. The regression value of the sample T10 (*Loopr*[*i*, T10])

2.3 Column Algorithm based on the Perturbing Influence Value

In Figure 1, there are some values that are above the dashed line, which means that the regression values of some samples violate Hypothesis 1. However, we suppose that the sum of PIVs would follow a certain rule more uniformly. The total influence value (TIV) of the sample x_j denoted by Q_j is defined as follow:

$$Q_j = \sum_{i=1}^{n} q_{ij} = \sum_{i=1}^{n} \left(Loopr[j, j] - Loopr[i, j] \right) \tag{3}$$

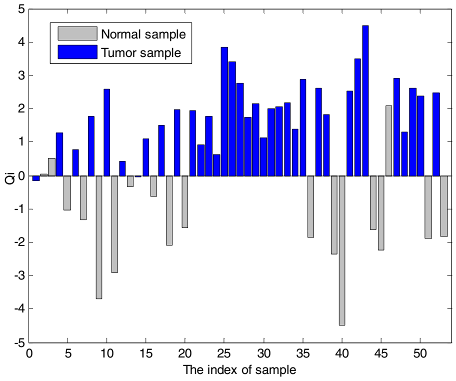


Fig. 2. The TIV value of each sample

If the majority of PIVs follow hypothesis 1, the sum of all PIVs of a sample should be positive for a sample with real label +1, and negative for a sample with real label -1. It is expected that the TIV value conforms to the following hypothesis.

Hypothesis 2: If the real label of the sample x_j is +1, Q_j is positive; otherwise, Q_j is negative.

Taking the Colon-p dataset as an example, Figure 2 shows the TIV value of each sample in the dataset.

As shown in Figure 2, the TIV value can clearly make a distinction between normal samples and tumor samples. It means the regression value can reflect the influence of flipping. Suppose that the wrong-labeled samples is much less than the correct-labeled ones and hence the minor ones do not predominate the sign of the TIV value. Consequently, we expected that the datasets containing label errors also satisfy Hypothesis 2.

Given an empirically determined threshold β and the sample x_j, if its label y_i is -1 and Q_i is larger than β, or if y_i is +1 and Q_i is smaller than β, then x_j is a suspect of wrong labeled sample.

Based on the discussion above, a column algorithm is proposed as follows.

Column Algorithm based on the Perturbing Influence Value (CAPIV)

```
Function CAPIV (Loopr, y)
    Begin
        $S = \{\}$  //initialize the entry list of the suspects
        For $j$=1 to n do
            calculate $Q_j$ for $x_j$
            If $y_j \times Q_j < \beta$ Then
                Sample $x_j$ is a suspect
                $S = S \cup x_j$
            Else
                Sample $x_j$ is not a suspect
            End If
        End For
        Return $S$
    End
```

2.4 Row Algorithm based on the Perturbing Influence Value

The TIV depends on the label value of the sample, but it is better to find a direct relationship between the wrong labeled sample and PIVs. We focus on the PIV values which are computed under the flipping of the same sample. The integrated influence value (IIV) of sample x_i, denoted by F_i is defined as follows:

$$F_i = \frac{1}{n} \sum_{j=1}^{n} \left(y_j \times q_{ij} \right) = \frac{1}{n} \sum_{j=1} y_j \left(Loopr[j, j] - Loopr[i, j] \right) \tag{4}$$

Assuming that most PIVs follow hypothesis 1, it can be inferred from the definition of the perturbing influence value that most $y_i \times q_{ij}$ should be positive since the label y_i and q_{ij} should have the same sign when the label of x_i is right, that is, if the label y_i of sample x_i is right-labeled, the product $y_i \times q_{ij}$ is expected to be positive; similarly, if sample x_i is wrong-labeled, it is expected to be negative.

Based on the discussion above, we state hypothesis 3 as follows:

Hypothesis 3: If the sample x_i is wrong labeled, F_i is negative; otherwise, F_i is positive.

In order to make a preliminary checking of what hypothesis 3 is promising, we intentionally flip the first 4 samples (with index 1, 2, 3 and 4) and the last 4 samples (with index 50, 51, 52 and 53) in Colon-p dataset to simulate the wrong labeled samples. The IIV value of every sample is shown in Figure 3. It can be seen that the IIV values of all the intentionally flipped samples are negative, which means that Hypothesis 3 is followed well. It is noticed that there are 5 "correct labeled" samples having the

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on January 24, 2014

negative IIVs. However, it is obvious that the absolute values of the "outliers" are relatively small except for the 11th sample. Hence, it is reasonable to infer that these outliers result from data noise.

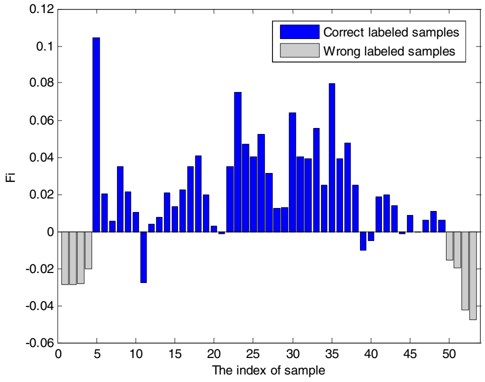


Fig. 3. The IIV value of each sample

Based on Hypothesis 3, a row algorithm is proposed as follows. Given a threshold γ and a sample x_i, if F_i is smaller than γ, x_i is a suspect sample.

Row Algorithm based on the Perturbing Influence Value (RAPIV)

```
Function RAPIV(Loopr, y)
    Begin
        $S = \{\}$   //initialize the entry list of the suspects
        For $i$=1 to n do
            calculate $F_i$ for $x_i$
            If $F_i < \gamma$ Then
                Sample $x_i$ is a suspect
                $S = S \cup x_i$
            Else
                Sample $x_i$ is not a suspect
            End If
        End For
        Return $S$
    End
```

2.5 Progressive Row algorithm based on the Perturbing Influence Value

A successful algorithm for mislabeling detection should take advantage of its performance by fixing the potential wrong-labeled sample. We expect that more samples would obey Hypothesis 3, if the IIV values are calculated by using pure labels according to Eq 4. Here, the pure labels mean that all labels do not have labeling error. It can be easily explained theoretically. In Eq 4, because of wrong-labeled samples, F_i cannot be calculated precisely. The fixing of wrong labels would be helpful to detect more suspected samples with IIVs around the threshold γ.

There is no need to rebuild *Loopr* after certain wrong labels are corrected, which would definitely enhance the implementation efficiency.

It is extremely hard to fix all wrong labels at one time because it is exactly our final goal. But we can correct labels progressively and just fix one or two of the wrong labels at each time rather than fix all of them at one time. According to the above discussion, we develop a progressive correction procedure based on the RAPIV algorithm proposed in Section 2.4.

(1) Firstly, define a variable V_{min} to save the minimum evaluation value in the progressive correction procedure and let $V_{min} = n$. Run RAPIV to obtain an initial suspect list S. Let the new label $y^{'}$=y, and the set of flipped samples T={}.

(2) Check every sample in the suspect list S which is not contained in T. Suppose that sample x_i is going to be checked, let $y^{'}_i = - y_i$, and the new suspect list $S^{'}_i$=RAPIV($Loopr$, y'). Then let $y^{'}_i = y_i$.

(3) Then an estimate method is performed to estimate the number of errors which includes false positives and false negatives in $S^{'}_i$. In the estimate method, each element of $S^{'}_i$ is flipped in y to get a new label vector $y^{''}$, and a mislabeling error detection method is used to detect the wrong labels in $y^{''}$. The number of the wrong labels denoted by D is the output of the estimate method. $S^{'}_i$ is a better suspect list when D is smaller.

(4) The evaluation value V_i=D+F_i where F_i is the IIV value of the sample x_i. If all the samples in S have been checked, go to step 5, otherwise, jump to step 2.

(5) Let $V_{i^*} = \min(V_i)$ where $x_{i^*} \in S$ and $x_{i^*} \notin T$. If $V_{i^*} > V_{\min}$ go to step 6, otherwise, put x_{i^*} in T, and let $y^{'}_{i^*} = -y_{i^*}$, $S = S^{'}_{i^*}$, and $V_{\min} = V_{i^*}$. If V_{min}>0, jump to step 2.

(6) Return S as the final result.

Note that in the estimate method, a mislabeling error detection method is performed. Considering the high precision of the CL-Stability algorithm (Malossini *et al*. 2006), we use it as the estimate method in this paper.

The pseudo code of PRAPIV is shown below:

Progressive Row Algorithm based on the Perturbing Influence Value (PRAPIV)

```
Function PRAPIV(Loopr, y)
    Begin
        $V_{min} = n$
        $S$ = RAPIV($Loopr$, $y$)
        $y^{'}$=$y$
        $T$={}
        While $V_{min}$>0 Do
            $k$=0
            For each $x_i$ in $S$
                If $x_i$ is not in $T$
                    $y^{'}(i) = -y^{'}(i)$
                    $S^{'}$=RAPIV($Loopr$, $y^{'}$)
                    $D$=Estimate($S^{'}$)
                    $V_i$=$D$+$F_i$
                    If $V_i$<=$V_{min}$ Then
                        $V_{min} = V_i$
                        $k = i$
                    End
                    $y^{'}(i) = -y^{'}(i)$
                End If
            End For
            If $k$=0 Then
                Return $S$
            Else
                $y^{'}(k) = -y(k)$
                $S$ = RAPIV($Loopr$, $y^{'}$)
                $T = T$ □ $x_k$
            End If
        End While
        Return $S$
    End
```

3 RESULTS AND DISCUSSION

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on January 24, 2014

Artificial datasets and microarray datasets are used to evaluate the proposed algorithms, including CAPIV, RAPIV and PRAPIV. We also perform simple SVM, CL-Stability and LOOE-Sensitivity algorithm in order to compare them against the methods presented in this paper. The simple SVM method takes all the samples except for the test sample as training set and use SVM to classify the test sample; if the result is not equal to the original label then it is a suspect of being a wrong-labeled sample.

3.1 Artificial Datasets

Artificial datasets are more reliable because their wrong labeled samples are known exactly. The experimental results on these datasets can reflect the true performance of the proposed methods.

We construct 6 kinds of artificial datasets mentioned in Section 2.1 and perform each algorithm with them. For each kind of artificial dataset, the experiments are performed independently for 50 times. The mean precision and recall values are listed in Table 3 and Table 4, respectively. In the experiments on artificial datasets, both β in CAPIV and γ in RAPIV are set equal to 0.

For the 6 kinds of datasets, the precision values of PRAPIV are all higher than those of the other methods, and CL-Stability always gives the second highest precision value. CAPIV and RAPIV have similar precision values as the simple SVM method. The recall values of CAPIV, RAPIV, PRAPIV and simple SVM are all higher than those of CL-Stability. The LOOE-Sensitivity algorithm performs worst in both precision values and recall values.

Table 3. Mean precision values on the artificial datasets

	Test1	Test2	Test3	Test4	Test5	Test6
Simple SVM	0.6127	0.5812	0.4481	0.7857	0.6292	0.7531
CL-Stability	0.7462	0.6827	0.5182	0.8613	0.7229	0.8226
LOOE-Sensitivity	0.2543	0.1436	0.04	0.3879	0.2259	0.3417
CAPIV	0.6523	0.5205	0.4463	0.7985	0.6499	0.7300
RAPIV	0.6418	0.5302	0.4537	0.8064	0.5512	0.7062
PRAPIV	**0.8341**	**0.7644**	**0.5484**	**0.9771**	**0.8681**	**0.9342**

Table 4. Mean recall values on the artificial datasets

	Test1	Test2	Test3	Test4	Test5	Test6
Simple SVM	0.925	0.8281	**0.6300**	0.9567	0.9133	0.95
CL-Stability	0.855	0.75	0.56	0.8933	0.82	0.9167
LOOE-Sensitivity	0.4928	0.3395	0.1047	0.2855	0.1436	0.3831
CAPIV	0.9134	0.8281	0.6291	0.9503	0.9082	0.9323
RAPIV	**0.935**	**0.8385**	0.62	0.9600	0.8267	0.9233
PRAPIV	0.915	0.8333	0.59	**0.9767**	**0.9167**	**1.0**

Because of the different conditions in the artificial datasets, different results are obtained with each method. For Test1, all the methods perform well, since the noise is low (only 4 wrong labeled samples in 30 samples). For Test2, the effect of every method turns worse because there are 2 more wrong labeled samples than Test1. For Test3, the number of the wrong labeled samples is up to 10 in all 30 samples. SVM produces poor results under such conditions, and so do the other 5 methods because they are based on SVM. It should be noticed that in Test3 every method performs worse than it does in the other datasets, especially for PRAPIV. The precision value of PRAPIV is around only 3% larger than that of the CL-Stability while in the other datasets it will be much larger. The PRAPIV algorithm uses the CL-Stability and the RAPIV algorithms, so the effect of PRAPIV depends on the results of those two methods. If both of them do not work well, it cannot be expected that PRAPIV could provide good results. Even though, the precision value of PRAPIV in Test3 is the largest among the 5 methods. The larger number of discriminate features in the Test4 and the smaller number of features in Test5 will actually increase the proportion of the discriminate features, so the results of all methods become better. Especially for the PRAPIV algorithm, its precision values and recall values are both the highest in Test4 and Test5, and its precision values are more than 10% higher than those of CL-Stability. Test6 has 20 more samples than the others, which makes the proportion of wrong labeled samples low. Hence, the situation in Test6 is similar to that in Test1.

PRAPIV presents a better balance between precision and recall than other methods according to the results in Table 3 and Table 4. The recall values of the simple SVM are large, but its precision values are small. The SVM method is good at classification, but the precision of classification cannot reach 100%. The samples misclassified by SVM become false positives in labeling error detection. Compared to the SVM method, the advantage of CL-Stability is that there are more classification results generated by Leave-One-Out method. Only when those classification results show some statistical significance, a sample will be detected as a wrong-labeled suspect. This advantage can help to limit the number of false positive samples, but it also makes some wrong labeled samples not be detected. Actually, the high precision values of CL-Stability are at the expense of recall. The CAPIV and RAPIV algorithms can keep the advantage of SVM providing high recall values, but their precision values are still very small. The reason is that the wrong labeled samples cause the imprecise calculations of the TIV values and the IIV values. The PRAPIV algorithm can overcome this deficiency of CAPIV and RAPIV by progressively correcting the suspects, so it has both the high precision and the high recall. The LOOE-Sensitivity algorithm always gives the worst results, because it uses SVM classification which cannot reflect the effect of perturbation. Considering the bad performance of the LOOE-Sensitivity algorithm, its result would not be shown in the experiments of microarray datasets.

3.2 Microarray Datasets

3.1.1 The original microarray datasets As we mentioned in 2.1.1, some evidences show that there may be some wrong labeled samples in the Colon dataset and the Breast datasets. Firstly we use the five methods mentioned above to test the Colon dataset and the Breast dataset. The results are shown in Table 5 and Table 6. The LOOE-Sensitivity is excluded in this section due to its bad performance in artificial dataset experiments.

For the Colon dataset, the CL-Stability detects correctly 6 suspects out of 9, and it produces only 2 false positives. The other methods all detect 7 suspects, but the Simple SVM, CAPIV and RAPIV give more false positives, especially for CAPIV. The

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on January 24, 2014

PRAPIV performs better than the others because it detects the most suspects correctly and produces the least false positives.

For the Breast dataset, the simple SVM, the CL-Stability, the CAPIV and the RAPIV all identify 5 suspect samples, and there is no false positive in the results of CL-Stability. The CAPIV and the RAPIV produce 1 false positives and the simple SVM gives 2. The PRAPIV detects 4 suspects with 1 false positive.

Table 5. List of suspects on original Colon dataset

	Simple SVM	CL-Stability	CAPIV	RAPIV	PRAPIV
T2	√	√	√		
T30	√	√	√	√	√
T33	√	√	√	√	√
T36	√	√	√	√	√
T37				√	√
N8				√	√
N12	√		√		
N34	√	√	√	√	√
N36	√	√	√	√	√
others	N2,N7, N27,N39	N2, N28	N2,T8,T9, T12,T25, N28,N40	N28, N29, N40	N2, N28

Table 6. List of suspects on original Breast dataset

	Simple SVM	CL-Stability	CAPIV	RAPIV	PRAPIV
11					
14	√	√	√	√	
16	√	√	√	√	√
31	√	√	√	√	√
33					
40	√	√	√	√	√
43	√	√	√	√	√
45					
46					
others	19, 47		47	47	19

3.1.2 The artificially flipped datasets Because in the real microarray datasets we may not truly know which samples are wrong labeled and the datasets containing the wrong labeled samples with biological evidences are hard to find, we choose some microarray datasets and artificially flip some samples to make them wrong labeled. Those artificially flipped datasets is reliable, so they can be used to examine the methods for labeling errors detection. Note that for the Colon dataset and the Breast dataset, we use the purified datasets (denoted by Colon-p and Breast-p) in which the suspect samples are eliminated instead of original ones.

Then we use artificially flipped datasets to test the 5 methods. The datasets used here are the Colon dataset (Alon *et al*., 1998), the Breast dataset (West *et al*., 2001), the CNS dataset (Pomeroy *et al*., 2002), the Cervix dataset (Wong *et al*., 2003) and the Prostate dataset (Welsh *et al*., 2001). We randomly choose six samples for each dataset and run the five methods. With 50 independent experiments, the mean precision values and the mean recall values are shown in Table 7 and Table 8 respectively for every dataset.

For the precision values, in all datasets except for the CNS dataset, PRAPIV dominates all other methods. CL-Stability always produces the second biggest precision value. For the recall values, the differences between the five methods are small, but usually the recall values of CL-Stability are the smallest. RAPIV and PRAPIV perform relatively better than the others for the recall values. Note that in the CNS dataset, the effect of PRAPIV is abnormally poor, where the precision value and the recall value are both the smallest. The performance of every method in the CNS dataset is worse than it is in other datasets, so this situation is similar to the artificial dataset Test3. Because PRAPIV depends on CL-Stability and RAPIV, if those two methods fail, it cannot be expected that PRAPIV will produce good results.

Table 7. Mean precision values on the microarray datasets

	Colon-p	Breast-p	CNS	Cervix	Prostate
Simple SVM	0.4570	0.6696	0.3904	0.4429	0.4549
CL-Stability	0.5040	0.7134	0.4271	0.4968	0.4972
CAPIV	0.4414	0.6716	0.4092	0.4142	0.4207
RAPIV	0.4346	0.6402	**0.4324**	0.4479	0.4515
PRAPIV	**0.7161**	**0.8444**	0.4024	**0.5571**	**0.8188**

Table 8. Mean recall values on the microarray datasets

	Colon-p	Breast-p	CNS	Cervix	Prostate
Simple SVM	0.8734	**0.9250**	0.7222	0.7889	0.8426
CL-Stability	0.8611	0.9106	0.6944	0.7333	0.7191
CAPIV	0.8732	0.9266	0.7222	**0.8111**	0.8605
RAPIV	0.8765	0.9268	**0.8056**	0.7778	0.8673
PRAPIV	**0.8765**	0.9146	0.5278	0.7111	**0.9105**

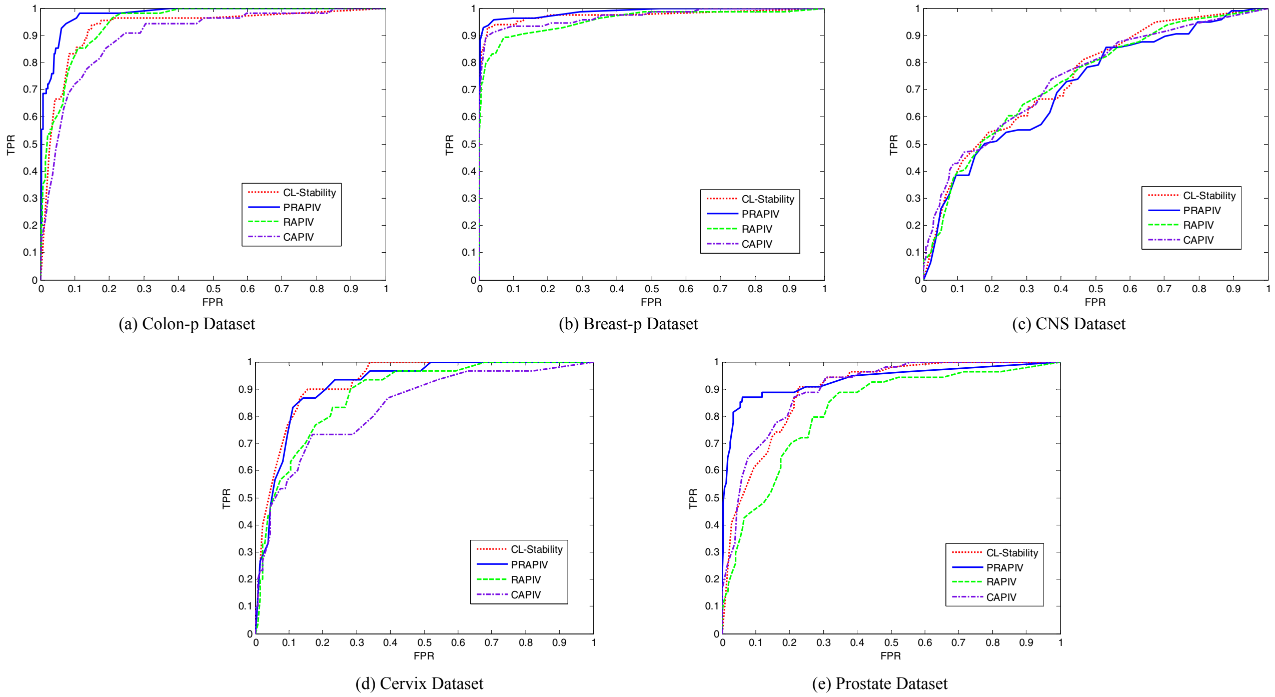
In order to compare the performance of the methods extensively, we plot the receiver operating characteristic (ROC) curves of CL-Stability as a function of the parameter α , CAPIV as a function of β, RAPIV and PRAPIV as functions of γ. For CL-Stability, the parameter α is varying from n+1 to 0 in step of 1. In CAPIV and RAPIV, the order of magnitude of the Q_i values and the F_i values depend on the specific datasets and they are hard to estimate beforehand, so the fixed values of the thresholds may not work in the ROC curve. Instead, we calculate $y_i \times Q_i$ and F_i for every sample x_i. For CAPIV the parameter β is varying from the largest $y_i \times Q_i$ to the smallest, and for RAPIV the parameter γ is varying from the largest F_i to the smallest. For PRAPIV, in the last time RAPIV is performed, the parameter γ is varying as it is stated above. Also, for every method, 50 replicates were performed and the average true positive rates and the average false positive rates are plotted in figure 4.

4 CONCLUSION

In this paper, three methods are proposed based on the definition of the perturbing influence value which is used to measure the effect of the data perturbation on the regression model. CAPIV and RAPIV adopt different perspectives on considering the perturbing influence. Compared with the LOOE-Sensitivity algorithm which measures the perturbing effect to the classifier, CAPIV and RAPIV perform better in every situation. Based on RAPIV, we develop the

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on January 24, 2014

PRAPIV algorithm with a progressive correction procedure in which CL-Stability is used to test the suspect samples identified by RAPIV. Experimental results show that the PRAPIV algorithm can increase the precision ratio while maintaining the high recall ratio. Because PRAPIV depends on RAPIV and CL-Stability, when RAPIV and CL-Stability fail to detect the suspects precisely, PRAPIV cannot perform well.



(a) Colon-p Dataset

(b) Breast-p Dataset

(c) CNS Dataset

(d) Cervix Dataset

(e) Prostate Dataset

Fig. 4. ROC curve for CL-Stability, RAPIV, CAPIV and PRAPIV in the Colon-p dataset, Breast-p dataset, CNS dataset, Cervix dataset and the Prostate dataset

FUNDING

NSFC (grant 60673023, 60703025, 60803052, and 10872077); the European Commission (grant TH/Asia Link/010 – 111084, and 155776-EM-1-2009-1-IT-ERAMUNDUS-ECW-L12); the National High-Tech R&D Program of China (863) (grant 2007AA04Z114, 2009AA02Z307); the Collaboration Project from Guangdong Province and MOE of China (grant 2007B090400031); the science-technology development project of Jilin Province of China (20080172), the project 200810026, 20091021 support by Jilin Unversity and “985”, “211” project of Jilin University.

ACKNOWLEDGEMENTS

The authors are grateful to Professor Raymond T. Ng, Department of Computer Science at University of British Columbia, for his helpful suggestions. Special thanks to the anonymous reviewers for their constructive comments and suggestions.

REFERENCES

Alon,U. et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotides array. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Antonov,A.V. et al. (2004) Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics*, **20**, 644–652.

Bø,T.H. and Jonassen,I. (2002) New feature subset selection procedures for classification of expression profiles. *Genome Biol.*, **3**, research0017.1–0017.11.

Brodley,C.E. and Friedl,M.A. (1999) Identifying mislabeled training data. J. *Artif. Intell. Res.*, **11**, 131–166.

Dudoit,S. et al. (2002) Comparison of discrimination methods for classification of tumors using gene expression data. *J. Am. Statist Assoc.*, **97**, 77–87.

Edwin, A. C. et al. (2000) Genomic analysis of metastasis reveals an essential role for RhoC. *Nature*, **406**, 532–535.

Kadota,K. et al. (2003) Detecting outlying samples in microarray data: a critical assessment of the effect of outliers on sample classification. *Chem-Bio Inform. J.*, **3**, 30–45.

Khan,J. et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med.*, **7**, 673–679.

Malossini,A. et al. (2006) Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics*, **22**, 2114–2121.

Muhlenbach,F., Lallich,S. and Zighed,D.A. (2004) Identifying and handling mislabelled instances. *J. Intell. Inform. Syst.*, **22**, 89–109.

Pomeroy,S.L. et al. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.

Sanchez,J.S. et al. (2003) Analysis of new techniques to obtain quality training sets. *Patt. Recogn. Lett.*, **24**, 1015–1022.

Schramm,A. et al. (2005) Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. *Oncogene*, **24**, 7902-7912.

Smola,A., and Scholkopf,B. (1998) A Tutorial on Support Vector Regression (Tech. Rep. No. NeuroCOLT NC-TR-98-030). Royal Holloway College, University of London, UK.

Tusher,V.G. et al. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Venkataraman,S. et al. (2004) Distinguishing mislabeled data from correctly labeled data in classifier design. *In 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04)*, Boca Raton, FL, p. 668–672.

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on January 24, 2014

Welsh,J.B. et al. (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.*, **61**, 5974–5978.

West,M. et al. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.

Wong,Y.F. et al. (2003) Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by DNA microarray. *Clinical Cancer Research*, **9**, 5486–5492.

Zhang,W. et al. (2006) A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer. *Bioinformatics*, **22**, 317–325.

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on January 24, 2014