ABSTRACT

Motivation: Mislabeled samples often appear in gene expression profile because of the similarity of different sub-type 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 article 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 effect of data perturbation, we define an index named perturbing influence value (PIV), based on the support vector machine (SVM) 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 article 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://cost.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 (Alon et al., 1999; Edwin et al., 2000; Pomeroy et al., 2002; Schramm et al., 2005; Welsh et al., 2001; West et al., 2001; Wong et al., 2003). 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 and Jonassen et al., 2002; Dudoni 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 sub-type 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 (Brody and Friedl, 1999; Muhlenbach et al., 2004; Sanchez et al., 2003; Venkataraman et al., 2004;) proposed many approaches for detecting labeling errors when the number of features (FN) is usually smaller than the number of the samples (SN). 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 (WLS) 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 sub-types 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 xi obtained with a SVM classifier while the sample xj is excluded from the training dataset and the label of the sample xj 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 WLS 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 between the correct samples and WLS 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 article, 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 PIVs (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 first introduce the datasets used in the experiments, and then define the PIV to measure the effect of data perturbation on the regression model. At last, the proposed algorithms are described based on the PIV, respectively.

2.1 Datasets

Our goal is to detect the WLS in microarray data, so several 2-class microarray datasets on cancers are selected to evaluate the algorithms proposed in this article. 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 article. 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 WLS.

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 FN, the SN, the number of discriminating features (DFN) and the number of WLS (WLN) are given in Table 2.

2.2 PIVs

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 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 article 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 expresses the expression vector of the sample i, and y_i is the label value of the sample i where y_i € {-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

\[ y_i = f(x_i) + \epsilon, \]

where \( f \) is a real-value function and \( \epsilon \) is noise. The aim is to find the regression model \( f \) that is an estimate of \( f \). Although there are many well-studied regression models, SVM regression model is used in this article to construct the approximation \( f \) due to its good theoretical basis and application performance in a number of fields (Smola and Scholkopf, 1998).

Instead of the LOOPC matrix defined by Malossini et al. (2006), a Leave-One-Out Perturbed Regression (Loop) matrix is defined, where the element Loop(i,j) is the regression value \( f(x_j) \) while the sample \( x_i \) 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 \( y_i \) 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 Loop, the element Loop(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 WLS can make the regression value closer to its true label (−1 or +1).

Figure 1 shows the perturbing regression values (Loop(i,j), with the label value of sample \( x_j \) is flipped). The effect of the flipping on the classifier could cause the poor performance of the LOOE-sensitivity algorithm.

In this article, 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 PIVs (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.

| Table 1. The 2-class microarray datasets |
| Datasets | Number of genes | Number of samples (SN) | References |
| 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 |

| Table 2. The artificial datasets |
| FN | SN | DFN | WLN |
| Test 1 | 2000 | 30 | 5 | 4 |
| Test 2 | 2000 | 30 | 5 | 6 |
| Test 3 | 2000 | 30 | 5 | 10 |
| Test 4 | 2000 | 30 | 10 | 6 |
| Test 5 | 2000 | 30 | 5 | 6 |
| Test 6 | 2000 | 50 | 5 | 6 |
According to Hypothesis 1, if the label of the sample is sensitive to flipping the label of a single sample, the total influence value (TIV) of the sample results, we define the PIV, which is the difference between regression values of some samples violate Hypothesis 1. 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.

2.3 CAPIV

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 is denoted by \(Q\), defined as follows:

\[
Q = \sum_{i=1}^{n} (\text{Loop}(i,j) - \text{Loop}\{i\})
\]

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, \(Q\) is positive; otherwise, \(Q\) 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 WLS is much less than the correctly 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 \(\beta\) and the sample \(x_j\), if its label \(y_j\) is \(-1 < \beta\), or if \(y_j\) is \(+1\) and \(Q\) is smaller than \(\beta\), then \(x_j\) is a suspect of WLS.

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

2.4 RAPIV

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

\[
F_i = \sum_{j=1}^{n} (y_j \times q_{ij}) = \sum_{j=1}^{n} (\text{Loop}(j,j) - \text{Loop}(j,j))
\]

Assuming that most PIVs follow Hypothesis 1, it can be inferred from the definition of the PIV that most \(y_j \times q_{ij}\) should be positive since the label \(y_j\) and \(q_{ij}\) should have the same sign when the label of \(x_j\) is right, i.e. if the label \(y_j\) of sample \(x_j\) is right-labeled, the product \(y_j \times q_{ij}\) is expected to be positive; similarly, if sample \(x_j\) 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 \(x_j\) 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 four samples (with index 1, 2, 3 and 4) and the last four samples (with index 50, 51, 52 and 53) in Colon-p dataset to simulate the WLS. The IV value of every sample is shown in Figure 3. It can be seen that the IV values of all the intentionally flipped samples are negative, which means that Hypothesis 3 is followed well. It is noticed that there are five ‘correct-labeled’ samples having the negative IVs. 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.
2.5 PRAPIV

The IIV value of each sample.

Fig. 3. The IIV value of each sample.

Based on Hypothesis 3, a Row Algorithm is proposed as follows. Given a threshold \(\gamma\) and a sample \(x_i\), if \(F_i < \gamma\), \(x_i\) is a suspect sample.

Row Algorithm based on the Perturbing Influence Value (RAPIV)

Function RAPIV(\(Loop_r\), \(y\))

1. Begin
2. \(S = \{\}\) //initialize the entry list of the suspects
3. For \(i = 1\) to \(n\) do
4. \(\text{calculate } F_i\) for \(x_i\)
5. If \(F_i < \gamma\) Then
6. Sample \(x_i\) is a suspect
7. \(S = S \cup x_i\)
8. Else
9. Sample \(x_i\) is not a suspect
10. End If
11. End For
12. Return \(S\)
13. End

2.5 PRAPIV

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

There is no need to rebuild \(\text{Loop}\) 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) First, define a variable \(V_{\text{max}}\) to save the minimum evaluation value in the progressive correction procedure and let \(V_{\text{max}} = 0\). Run RAPIV to obtain an initial suspect list \(S\). Let the new label \(y'_i = y_i\), 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 \(S' = \text{RAPIV}(\text{Loop}, \gamma')\). 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 \(S'\). In the estimate method, each element of \(S'\) 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. \(S''\) 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 \(V_{\text{max}} = \min(V_i)\) where \(x_i \in S\) and \(x_i \notin T\). If \(V_{\text{max}} > V_{\text{max}}\) go to step 6, otherwise, put \(x_i\) in \(T\) and let \(y'_i = -y'_i\), \(S = S'\), and \(V_{\text{max}} = V_i\). If \(V_{\text{max}} > 0\), 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 article.

The pseudo code of PRAPIV is shown below.

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

Function PRAPIV(\(Loop_r\), \(y\))

1. Begin
2. \(V_{\text{max}} = 0\)
3. \(S = \text{RAPIV}(\text{Loop}, \gamma)\)
4. \(y' = y\)
5. \(T = \{\}\)
6. While \(V_{\text{max}} > 0\) Do
7. \(k = 0\)
8. For each \(x_i\) in \(S\) do
9. If \(x_i\) is not in \(T\)
10. \(y'_i = -y'_i\)
11. \(S' = \text{RAPIV}(\text{Loop}, \gamma')\)
12. \(D = \text{Estimate}(S')\)
13. \(V_i = D + F_i\)
14. If \(V_i > V_{\text{max}}\) Then
15. \(V_{\text{max}} = V_i\)
16. \(S = S'\)
17. \(k = i\)
18. \(y'_k = -y'_k\)
19. End If
20. End For
21. If \(k > 0\) Then
22. Return \(S\)
23. Else
24. \(y' = y' \cup x_i\)
25. \(S = T \cup x_i\)
26. \(T = T \cup x_i\)
27. End If
28. End While
29. Return \(S\)
30. End
Artificial datasets are more reliable because their WLS are known. Mean precision values on the artificial datasets are listed in Table 3. The bold value in each column indicates the highest mean precision value for each dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
<th>Test 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple SVM</td>
<td>0.6127</td>
<td>0.5812</td>
<td>0.4481</td>
<td>0.7857</td>
<td>0.6292</td>
<td>0.7531</td>
</tr>
<tr>
<td>CL-stability</td>
<td>0.7462</td>
<td>0.6827</td>
<td>0.5182</td>
<td>0.8613</td>
<td>0.7228</td>
<td>0.8226</td>
</tr>
<tr>
<td>LOOE-sensitivity</td>
<td>0.2543</td>
<td>0.1436</td>
<td>0.04</td>
<td>0.3879</td>
<td>0.2259</td>
<td>0.3417</td>
</tr>
<tr>
<td>CAPIV</td>
<td>0.6522</td>
<td>0.5205</td>
<td>0.4463</td>
<td>0.7985</td>
<td>0.6499</td>
<td>0.7300</td>
</tr>
<tr>
<td>RAPIV</td>
<td>0.6614</td>
<td>0.5302</td>
<td>0.4537</td>
<td>0.8064</td>
<td>0.5512</td>
<td>0.7062</td>
</tr>
<tr>
<td>PRAPIV</td>
<td>0.8341</td>
<td>0.7644</td>
<td>0.5484</td>
<td>0.9771</td>
<td>0.8681</td>
<td>0.9342</td>
</tr>
</tbody>
</table>

The bold value in each column indicates the highest mean precision value for each dataset.

Mean recall values on the artificial datasets are listed in Table 4. The bold value in each column indicates the highest mean recall value for each dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
<th>Test 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple SVM</td>
<td>0.925</td>
<td>0.8281</td>
<td>0.6300</td>
<td>0.9567</td>
<td>0.9113</td>
<td>0.95</td>
</tr>
<tr>
<td>CL-stability</td>
<td>0.855</td>
<td>0.75</td>
<td>0.56</td>
<td>0.8933</td>
<td>0.82</td>
<td>0.9167</td>
</tr>
<tr>
<td>LOOE-sensitivity</td>
<td>0.4928</td>
<td>0.3395</td>
<td>0.1047</td>
<td>0.2855</td>
<td>0.1436</td>
<td>0.3831</td>
</tr>
<tr>
<td>CAPIV</td>
<td>0.9134</td>
<td>0.8281</td>
<td>0.6291</td>
<td>0.9503</td>
<td>0.9082</td>
<td>0.9323</td>
</tr>
<tr>
<td>RAPIV</td>
<td>0.935</td>
<td>0.8385</td>
<td>0.62</td>
<td>0.9600</td>
<td>0.8267</td>
<td>0.9233</td>
</tr>
<tr>
<td>PRAPIV</td>
<td>0.915</td>
<td>0.8333</td>
<td>0.59</td>
<td>0.9767</td>
<td>0.9167</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The bold value in each column indicates the highest mean recall value for each dataset.

3.1 Artificial datasets

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

We construct six 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 Tables 3 and 4, respectively. In the experiments on artificial datasets, both $\beta$ in CAPIV and $\gamma$ in RAPIV are set equal to 0.

For the six 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 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 with 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 WLS 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 WLS 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 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

3.2.1 The original microarray datasets

As we mentioned in Section 2.1.1, some evidences show that there may be some WLS in the Colon dataset and the Breast datasets. First, we use the five methods mentioned above to test the Colon dataset and the Breast dataset. The results are shown in Tables 5 and 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 six suspects out of nine, and it produces only two false positives. The other methods all detect seven suspects, but the Simple SVM, CAPIV and RAPIV give more false positives, especially for CAPIV. The 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 five suspect samples, and there is no false positive in the results of CL-stability. The CAPIV and the RAPIV produce one false positive and the simple SVM gives two. The PRAPIV detects four suspects with one false positive.

3.2.2 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 WLS 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 are 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 five methods. The datasets used here are the Colon dataset (Alon et al., 1999), the Breast dataset (West et al., 2001), the CNS dataset (Pomeroy et al., 2002), the Cervix dataset (Wong et al., 1999) 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 Tables 7 and 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 Test 3. Because PRAPIV depends on CL-stability and RAPIV, if those two methods fail, it cannot be expected that PRAPIV will produce good results.

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 $\alpha$, CAPIV as a function of $\beta$, RAPIV and PRAPIV as functions of $\gamma$. For CL-stability, the parameter $\alpha$ is varying from $n+1$ to 0 in step 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 $\beta$ is varying from the largest $y_i \times Q_i$ to the smallest, and for RAPIV the parameter $\gamma$ is varying from the largest $F_i$ to the smallest. For PRAPIV, in the last time RAPIV is performed, the parameter $\gamma$ 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 article, three methods are proposed based on the definition of the PIV 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 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.
The authors are grateful to Prof. 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.

Funding: NSFC (grants 60673023, 60703025, 60803052 and 10872077); European Commission (grants TH/Asia Link/010-111084, and 155776-EM-I-2009-1-IT-ERAMUNDUS-ECW-L12); National High-Tech R&D Program of China (863) (grants 2007AA042114 and 2009AA022307); Collaboration Project from Guangdong Province and MOE of China (grant 2007B090400031); Science-technology Development Project of Jilin Province of China (grant 20080172); Jilin University (project ‘985’, ‘211’ ); Jilin University (project 200810026, 20091021); 2007AA04Z114 and 2009AA02Z307); Collaboration Project from National High-Tech R&D Program of China (863) (grants: NSFC (grants 60673023, 60703025, 60803052 and 10872077); European Commission (grants TH/Asia Link/010-111084, and 155776-EM-I-2009-1-IT-ERAMUNDUS-ECW-L12); National High-Tech R&D Program of China (863) (grants 2007AA042114 and 2009AA022307); Collaboration Project from Guangdong Province and MOE of China (grant 2007B090400031); Science-technology Development Project of Jilin Province of China (grant 20080172); Jilin University (project ‘985’, ‘211’ ).

Conflict of Interest: none declared.

REFERENCES