30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

Smoothing Blemished Gene Expression Microarray Data via Missing Value Imputation

Zhipeng Cai, Yi Shi, Meng Song, Randy Goebel, and Guohui Lin*

Abstract— **Gene expression microarray technology has enabled advanced biological and medical research, but the data are well-recognized noisy and must be used with caution, since they are greatly affected by many experimental factors such as RNA concentration, spot typing, hybridization condition, and image analysis. It is highly desirable that the inaccurate data entries ("stains") can be identified and subsequently curated. In this paper, we propose a novel computational method, based on feature gene selection and sample classification, to efficiently discover the stains and apply imputation methods to estimate their values. Extensive experimental results on three Affymetrix platforms for human cancer diagnosis showed that by picking only 1–4% data entries as the most likely stains, the smoothed datasets could be used for better downstream data analyses such as robust biomarker identification and disease diagnosis.**

I. INTRODUCTION

Microarrays, typically high-density oligonucleotide arrays, such as Affymetrix GeneChip oligonucleotide arrays and Agilent Dual Mode whole genome gene expression arrays, can simultaneously assess expression levels of thousands of genes under a variety of conditions. This high-throughput technology provides a unique tool for systems biology, and has important applications in numerous biological and medical studies. One of the most common and important tasks in these applications is to compare the gene expression levels in tissues under different conditions, such as healthy versus diseased, for effective genetic profiling. Such a task can only be accomplished with complete and accurate expression data, which however is often challenging to obtain in practice due to a number of artifacts in the experiments [10].

Noise in gene expression microarray data comes from many sources, some of which is caused by experimental setup, such as insufficient resolution, image corruption, or even dust and scratches on the slide [20]. Other noise could be caused by the chip design itself. For example, in general, probes can over- or under-estimate gene activity [12]. A probe set in an Affymetrix oligonucleotide array normally consists of multiple pairs of oligonucleotide probes [8]. Using pre-specified mapping criteria, the expression value of the probe set can be obtained from the hybridization levels of these probe pairs. The large number of probe pairs guarantees a substantially low probability of missing all the hybridization levels, thus ensuring reading of the expression value for a probe set. However, although criteria have been set for assessing the overall quality of a chip, probes have not been designed to be specific to gene splice variants and little sensitivity is promised for detecting localized artifacts, such as "harshlight" in microarray image. So far there are no safeguards to signal potential physical blemishes. Another confounding factor for getting accurate expression data is cross-hybridization. Oligonucleotide probes often relate not only to gene products that exactly match the sequence, but also those with near matches.

On the other hand, all the downstream computational data analyses require the gene expression dataset to be complete and accurate. Therefore, it is desirable to identify the inaccurate data entries (called *stains*), if any, and adjust them. Several ideas have focused on discovering inaccurate data entries. For example, due to ozone degradation, one channel in a two-channel microarray experiment would produce poor quality data. There are two possible solutions: one to exclude one channel, and the other to discard only the affected arrays. Lynch *et al.* [11] proposed to combine these two methods with a linear model to detect affected positions. Due to variations in experimental conditions, it is difficult to combine the data from different arrays. Barenco *et al.* [3] proposed a simple recursive algorithm to correct the mismatches in oligonucleotide microarray data, by using constant genes to rescale the datasets such that expression data are normalized and consistent. Tran *et al.* [19] presented an approach to identify accurate signals and used a simple correlation between mean and median to adjust those inaccurate signals. Blemished data are usually outliers in the dataset, and those caused by different reasons will have different outlier patterns. Suarez-Farinas *et al.* [17], [18] proposed a method to find "harshlight" blemishes in chips due to physical or chemical problems. Using statistics on a number of the same type of arrays under similar experimental conditions, they devised a pattern recognition algorithm to identify and eliminate a variety of defects.

Here we assume complete microarray datasets and present a **computational** method to discover the expression outliers as inaccurate data entries, then re-estimate them. We evaluate the quality of the resultant, called *smoothed*, datasets through a downstream application — feature gene selection and the sample classification accuracy based on the selected feature genes. The rationale supporting such an evaluation is that only an accurate prediction of sample conditions can eventually demonstrate the value of the gene expression microarrays [12]. We have included three real human cancer gene expression microarray datasets, each obtained using a common platform, in the experiments to demonstrate the success of our method. These three human datasets are the Carcinomas dataset [16], the Ovarian dataset [15], and

*Corresponding author. Department of Computing Science, University of Alberta. Edmonton, Alberta T6G 2E8, Canada. Emails: zhipeng, ys3, mengs, goebel, ghlin@cs.ualberta.ca

978-1-4244-1815-2/08/$25.00 ©2008 IEEE.

the Gliomas dataset [13]. We calculated the 5-fold cross validation classification accuracies for the KNN-classifier and the SVM-classifier, which are built on a number of genes selected from the smoothed datasets and the original datasets. The achieved classification accuracies before and after the data smoothing on all these datasets are statistically different, indicating that smoothing is able to adjust at least partially data blemishes.

II. Methods

The gene expression data generated from microarray experiments is presented as a matrix $A_{p\times n}$, in which there are p genes, n samples, and a_{ij} denotes the expression level of the i-th gene in the j-th sample.

In the 5-fold cross validation scheme, $\frac{4}{5}$ of the samples are used as the training dataset, in which every sample is labeled by its class membership. To build a sample classifier for testing sample prediction, a gene selection method is used to identify a number of discriminatory genes. In this study, we adopted two existing gene selection methods: F-test [2] and Scatter [5]. Essentially, each gene selection method assigns a score to every gene, where a bigger score indicates a higher class discrimination strength. This scoring is done on the training dataset. Let s_i be the score F-test assigns to gene i. Also, for each entry a_{ij}, F-test ignores the j-th sample to assign gene i another score s'_i. The value $s_{ij} = |s_i - s'_i|$ measures the *abnormality* of data entry a_{ij}, and a higher value indicates a more problematic entry. This way, F-test assigns an abnormality value to each data entry in the training dataset. Let r denote the percentage of data entries which our computational method will regard as inaccurate and subsequently to re-estimate their values. We call this pre-specified percentage r the *inaccurate rate* of the dataset, which ranged from 1% to 30% in our experiments. Given an r, the top r of the data entries, ranked by the abnormality values, s_{ij}, and under three separate inaccurate data entry distributions, are identified as inaccurate. We note that such a process of inaccurate entry identification relies on the detailed gene selection method, and different methods might identify different sets of inaccurate entries. Nevertheless, F-test and Scatter performed quite consistently in our experiments. Also, the dataset should be large enough, that is, a sufficient number of genes and a sufficient number of samples in each class; for otherwise the identification result could be biased.

The discovered inaccurate data entries are then erased from the training dataset, i.e., treated as missing. And subsequently a missing value imputation method is called to impute their values. Right now there are more than a dozen imputation methods available. In this study, we employed the weighted k-Nearest Neighbor imputation (KNNimpute) method [20] (three other methods, SKNNimpute [9], BPCAimpute [14], and ILLSimpute [4], were also used, whose performance were similar but slightly worse than KNNimpute, data not shown). KNNimpute is shown to be a simple yet competitive missing value imputation method. The imputed (or *smoothed*) training dataset is again complete and is ready for the next step of feature gene selection.

The two gene selection methods, F-test and Scatter, are re-used to select a number of feature genes on the smoothed datasets, for sample classifier construction. We included in this study two classifiers: the k-Nearest Neighbor (KNN) classifier [6] (we set the default value of k to be 5, after testing k from 3 to 10) and a linear kernel Support Vector Machine (SVM) classifier [7].

III. Results and Discussion

A. Dataset descriptions and inaccurate entry discovery

Three real human cancer gene expression microarray datasets from three platforms are used in this study. The Ovarian cancer dataset [15] contains a total of 104 samples in four classes. *serous*, *endometrioid*, *mucinous*, and *clear cell*, which have $53, 10, 33, 8$ samples respectively. This dataset is obtained from Affymetrix GeneChip Hu6800. Besides Ovarian cancer dataset, the other two cancer datasets are the Carcinomas dataset [16] and the Gliomas dataset [13]. All computational results are available at Supplementary Materials [1].

We tested three separate assumptions on the distribution of the inaccurate data entries inside the expression matrices. Given a pre-specified inaccurate rate r, in the first assumption, for each gene, the top nr among its n entries are treated as inaccurate; in the second assumption, for each sample, the top pr among its p entries are treated as inaccurate; in the last assumption, the top pnr among all the $p \times n$ entries are treated as inaccurate. For simplicity, we call them the assumptions on genes, on samples, and on the whole dataset, respectively. We note that when combining multiple microarray chips into one single dataset for data analysis, one has to take into consideration that the individual chips were exposed in possibly, might only slightly, different experimental conditions. Therefore, even after the proper data normalization to tune the multiple chips into a common setup, there could be cases where one gene behave more abnormally than the other, or one chip behave more abnormally than the other. These three assumptions on the distribution of inaccurate entries were proposed to examine all these possibilities.

B. Sample classification accuracies

On each gene expression microarray dataset, for each inaccurate rate r, we need to collect 4 classification accuracies: on the original dataset, and on the smoothed dataset based on uniform distribution assumptions on genes, on samples, and on the whole dataset, respectively. We use *Original*, *Gene*, *Sample*, and *Whole* to denote these 4 classification accuracies, respectively.

In the 5-fold cross validation scheme, the other $\frac{1}{5}$ of the samples form the *testing dataset* in which the sample class labels are blinded to the sample classifiers, built on the associated training dataset, for prediction. Every $\frac{1}{5}$ of the samples are rotated to be the testing dataset. Note that testing samples have nothing to do with inaccurate entry discovery. The percentage of correctly predicted samples (true positives)

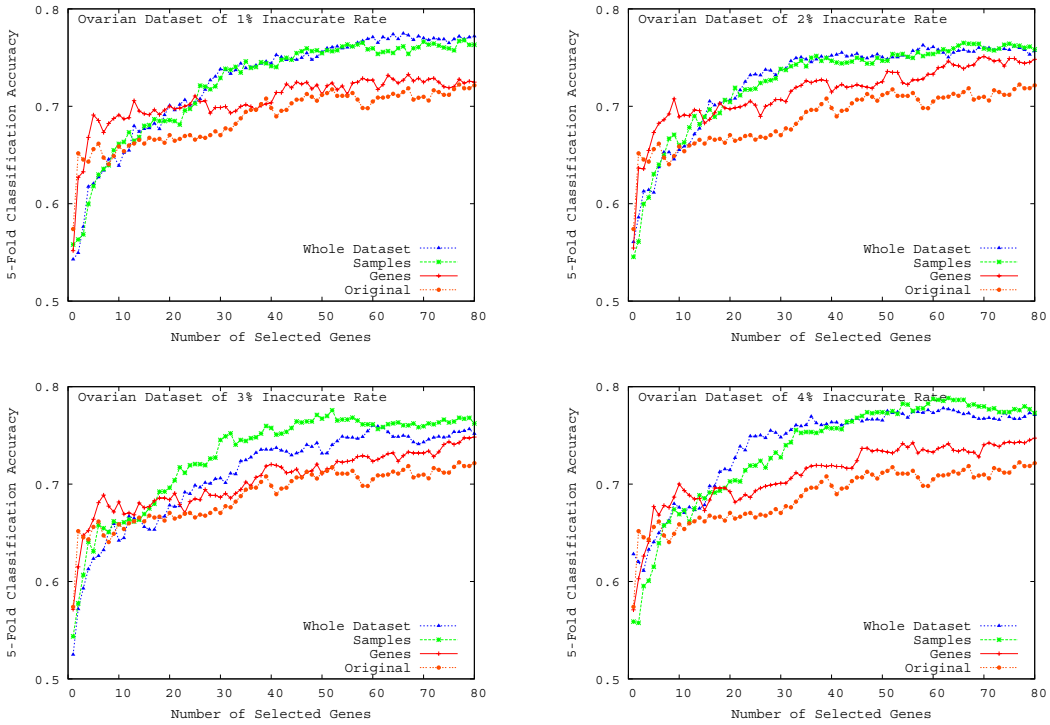

Fig. 1. 5-fold cross validation classification accuracies of F-test-KNNimpute-KNN on the Ovarian dataset, assuming 1–4% inaccurate entries under all three distribution assumptions, where 1–80 genes were selected.

is the classification accuracy associated with this partition. The cross validation process is repeated 20 times in our experiments, and the average classification accuracy is the final classification accuracy. Note that we have used two gene scoring methods, the KNNimpute method, and two classifiers. We concatenate their names to label the classification accuracies. For instance, "Gene-1%-F-test-KNNimpute-KNN" denotes the classification accuracy that is achieved by applying F-test to discover 1% inaccurate entries for every gene, using KNNimpute to re-fill their entries, again applying F-test to select feature genes to build a KNN-classifier, and running the KNN-classifier for sample class membership prediction.

Figure 1 plots the classification accuracies of the F-test-KNNimpute-KNN method on the Ovarian dataset, each assumed 1–4% inaccurate entries under all three distribution assumptions. The standard deviations of those 20 5-fold cross validation classification accuracies on each dataset were all very small compared to the average classification accuracies, almost always less than 0.05 and decreasing with the number of selected genes. These deviations on the Ovarian dataset, with inaccurate entry distribution assumption on the whole dataset, are supplied in Supplementary Materials [1]. The classification accuracies achieved on the smoothed datasets are almost always higher than those achieved on the original datasets, where the number of selected feature genes ranges from 1 to 80. More specifically, under the assumptions on the whole dataset and on samples, the achieved classification accuracies intertwine a bit, but they are clearly higher than the classification accuracies achieved on the original datasets, particularly when the number of selected genes is large (> 25). Under the third assumption on genes, the achieved classification accuracies on the smoothed datasets are mostly lower than those on the first two smoothed datasets, yet still slightly higher than those achieved on the original datasets. It is worth pointing out that, the differences between the achieved classification accuracies on the original dataset and the smoothed one vary from dataset to dataset, which might be due to the quality of the original datasets. Overall, from these results on the three datasets obtained by Affymetrix genechips, we may conclude that the inaccurate entry uniform distributions on the whole dataset and on samples might be better than the third assumption on genes. This conclusion might largely correlate with imperfect experimental conditions for collecting the gene expression data.

C. *Difficult samples now correctly predicted*

From Figure 1, we see that when assuming 4% whole dataset inaccurate rate in the Ovarian dataset, F-test-KNNimpute-KNN reached the highest classification accuracy (78.75%) when the KNN-classifier was built on 60 selected genes. Using the same smoothing procedure, in the leave-one-out cross validation (LOOCV) to select 60 genes, the achieved classification accuracy by F-test-KNN-classifier is also 78.75%. We collected the detailed confusion matrix on the smoothed datasets, and compared it with the LOOCV sample class prediction confusion matrix on the original dataset by F-test-KNN-classifier (78.85% versus 72.12%),

in the following Table 1. Note that there are 9 mucinous

TABLE I

THE LOOCV SAMPLE PREDICTION RESULTS ON THE OVARIAN DATASET

Class	Original				Smoothed @4%			
serous	**51**		2		**47**		6	
endometrioid	1	**5**	4			**5**	5	
mucinous	18		**15**		9		**24**	
clear cell		3	1	**4**		2		**6**

samples and 2 clear cell samples, which were difficult for prediction using the original dataset, now correctly predicted using the smoothing technique. One also sees that there are 4 serous samples now mis-classified. One possible cause is the gene selection method F-test. Note that each gene has its own power to differentiate some specific pairs of classes. In our experiments, only a limited number of feature genes were picked and combined with classifiers to perform the classification. Consequently, when some feature genes that have more discriminative power to identify the mucinous and clear cell samples were selected to construct the sample classifier, the genes that have more discriminative power to identify the serous samples might be kicked out. In this case, the prediction on serous samples became worse.

D. Experiments on simulated datasets

In these experiments, we further demonstrate that our smoothing method can indeed discover the inaccurate data entries and smooth them to improve dataset quality. We do this through a simulation study, where a good quality gene expression microarray dataset is artificially perturbed with random noise, and examine the performance of a classifier on the original dataset, the perturbed dataset, the smoothed dataset based on the original dataset, and the smoothed dataset based on the perturbed dataset. We use three datasets: the Carcinomas dataset, the Lung dataset, and the SRBCT dataset, in this simulation study. We only show the results on the SRBCT dataset because of page limit. All the other results can be found in Supplementary Materials [1]. Note that the 5-fold cross validation classification accuracies of an F-test-KNN-classifier on these three original datasets are all higher than 90%, and therefore considered as of good quality.

On each of the three datasets, we randomly selected 10% data entries from the whole dataset and perturbed them by adding to them a 0-mean uniformly distributed noise with the standard deviation equal to the absolute expression value. The subsequent smoothing method was used to identify the same percentage of data entries from the whole dataset, and treated them as inaccurate. Figure 2 plots the 5-fold cross validation classification accuracies on the four datasets. To summarize, though varying a little, the performance of F-test-KNNimpute-KNN on the two smoothed datasets is slightly better than on the original dataset, and the performance on the perturbed dataset is clearly worse. Note that the original SRBCT dataset is considered as of good quality, and therefore our smoothing method may only contribute a little

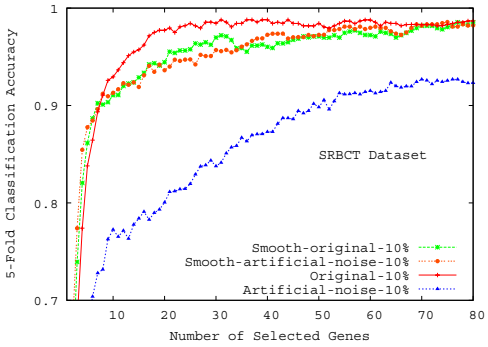

Fig. 2. 5-fold classification accuracies of F-test-KNNimpute-KNN on the original, the perturbed with 10% whole dataset inaccurate rate, the smoothed based on the original, and the perturbed and smoothed SRBCT datasets.

(original versus smoothed original). However, when the noise is obvious, such as the perturbed datasets, our smoothing method worked effectively in discovering the inaccurate data entries and curating them to improve the data quality.

IV. CONCLUSIONS AND FUTURE WORKS

Gene expression microarray datasets are in general noisy. We proposed a novel computational method to detect those inaccurate data entries based on their effects on feature gene selection and the subsequent sample classification. The extensive experiments showed that the proposed smoothing method reduced the noise level in the original datasets, and that the smoothed datasets had significantly better quality in terms of feature gene selection and sample classification. Note that for Affymetrix platform, the expression level of a gene is derived from hybridization values of multiple probes and the hybridization value of a probe affects the expression levels of multiple genes (i.e., many-to-many mapping rules). A possibly more effective smoothing approach is to detect the inaccurate probe hybridization values, adjust them, and subsequently re-calculate the affected gene expression values.

REFERENCES

[1] Supplementary materials: [http://www.cs.ualberta.ca/~ghlin/src/WebTools/smoothing.php]
[2] P. Baldi and A. D. Long. *Bioinformatics*, 17:509–519, 2001.
[3] M. Barenco *et al. BMC Bioinformatics*, 7:251, 2006.
[4] Z. Cai *et al. J. Bioinfo. Comput. Biol.*, 4:935–957, 2006.
[5] H. Chai and C. Domeniconi. In *Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics*, 2004.
[6] S. Dudoit *et al. J. the Amer. Stat. Assoc.*, 97:77–87, 2002.
[7] I. Guyon *et al. Machine Learning*, 46:389–422, 2002.
[8] R. A. Irizarry *et al. Nucleic Acids Research*, 31:4e15, 2003.
[9] K. Y. Kim *et al. BMC Bioinformatics*, 5:160–169, 2004.
[10] L. Klebanov and A. Yakovlev. *Biology Direct*, 2:9, 2007.
[11] A. G. Lynch *et al. BMC Bioinformatics*, 8:26, 2007.
[12] E. Marshall. *Science*, 306:630–631, 2004.
[13] C. L. Nutt *et al. Cancer Research*, 63:1602–1607, 2003.
[14] S. Oba *et al. Bioinformatics*, 19:2088–2096, 2003.
[15] D. R. Schwartz *et al. Cancer Research*, 62:4722–C472, 2002.
[16] A. I. Su *et al. Cancer Research*, 61:7388–7393, 2001.
[17] M. Suarez-Farinas *et al. BMC Bioinformatics*, 6:65, 2005.
[18] M. Suarez-Farinas *et al. BMC Bioinformatics*, 6:294, 2005.
[19] P. H. Tran *et al. Nucleic Acids Research*, 30:e54, 2002.
[20] O. Troyanskaya *et al. Bioinformatics*, 17:520–525, 2001.