A Kernel SVM Algorithm to Detect Mislabeled Microarrays in Human Cancer Samples

Manuel Martín-Merino

Abstract— **DNA Microarrays have been successfully applied to the identification of different cancer types considering the gene expression profiles. However, previous studies have shown that labeling errors are not uncommon in microarray studies. In this case, the training set may contain mislabelled examples that may lead the classifier to poor performance.**

In this paper we propose a new filtering algorithm based on one-class SVM classification to detect mislabelled samples. To this aim, samples and labels are mapped together to feature space using the kernel of dissimilarities. Next, outliers are detected via one-class classification. Mislabeled samples and outliers in input space can be separated comparing the outliers obtained in input and feature spaces.

The algorithm proposed has been tested using several complex cancer microarray datasets in which some samples are mislabelled according to the literature. The experimental results suggest that our algorithm is effective detecting labeling errors and compares favorably with a standard technique such as simple SVM.

I. INTRODUCTION

DNA Microarrays allow us to monitor the expression level of thousands of genes simultaneously across a collection of related samples [1]. This technology can be applied to identify different types of cancer and potential gene markers considering the gene expression profiles [2]. However, previous studies have shown that labeling errors are not uncommon in microarrays studies [3]. In particular, [4] has reported that there are $10-15\%$ of mislabelled samples in microarrays due to similarity of different subtypes of diseases. In this case, the training set contains mislabelled examples that may deteriorate the classifier performance particularly when label noise is asymmetric [5].

To overcome this problem, several methods have been proposed in the literature. [6] identifies suspect samples when in a neighborhood defined by a geometrical graph the proportion of samples from the same class is significantly greater than in the database. However, the algorithm has been applied only to datasets with $n > p$, n the number of samples and p the dimensionality of input space. To avoid this problem, [5] proposed two algorithms based on a perturbed classification matrix. These algorithms identify potential mislabelled samples analyzing the predicted labels under small perturbations obtained by flipping the label of a single sample. [3] improved the algorithm by considering an index that reflects better the effect of small perturbations. The performance of previous algorithms depends crucially on some tuning parameters that remain difficult to estimate. Finally, other approaches [7] model the label noise using probabilistic models. However, they are sensitive to the small sample size problem.

Manuel Martín-Merino is with the Computer Science Department University Pontificia of Salamanca, C/Compañía 5, 37002 Salamanca, Spain mmartinmac@upsa.es. Paper submission date: 30/07/2013

Let $\mathcal{X}$ denotes the input space and $\mathcal{Y}$ the space of labels. Our approach is based on the idea that mislabelled examples can be detected as outliers in the $\mathcal{X} \times \mathcal{Y}$ space after a suitable transformation, First, samples and labels are mapped to feature space using the kernel of dissimilarities. Each sample is represented by the distances to the k-nearest neighbors computed in input space. Mislabelled examples can be identified as outliers in feature space considering the one-class SVM classification algorithm. The parameters such as the number of neighbors are computed using a cross-validation approach.

The algorithm has been tested using several complex cancer microarray datasets in which some samples are mislabelled according to the literature. The experimental results suggest that our algorithm is effective detecting labelling errors and compares favorably with a standard technique such as the simple SVM algorithm.

The paper is organized as follows. In section II we introduce the one-class SVM algorithm for outlier detection. In section III the method proposed to detect mislabelled samples is presented. Section IV reports some experimental results using several human cancer classification problems. Finally, section V outlines conclusions and future research trends.

II. BACKGROUND: ONE-CLASS CLASSIFICATION

In this section we introduce the one-class SVM classification algorithm proposed by [8]. This technique allow us to detect the support of a high dimensional distribution. It can be applied to detect outliers even when the data is not represented in a vectorial space [9], [10].

One class-classification estimates a binary function that takes the value of $+1$ for regions of high density than contains most of the points and -1 else where. The algorithm maps the data points to a feature space determined by a kernel function [10] and compute a hyperplane that separates with largest margin the mapped data $\{\mathbf{\Phi}(\boldsymbol{x}_i)\}_{i=1}^n$ from the origin, where $\mathbf{\Phi}$ is the mapping induced by the kernel function. Figure 1 shows the separating hyperplane determined by the normal vector $\boldsymbol{w}$. The outlier $\mathbf{\Phi}(\boldsymbol{x})$ is associated with a slack variable ξ. The distance from the outlier to the hyperplane is $\xi/\|\boldsymbol{w}\|$. The distance from the hyperplane to the origin is $\rho/\|\boldsymbol{w}\|$ such that a small $\|\boldsymbol{w}\|$ corresponds to large margin of separation from the origin.

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

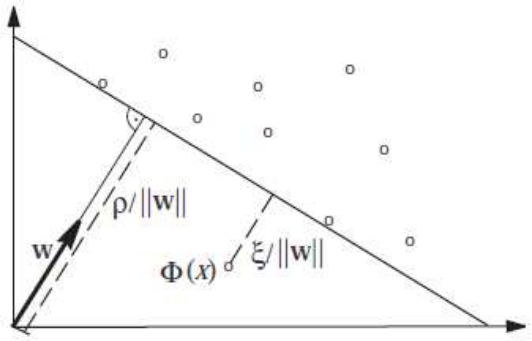

Fig. 1. Diagram of linear one-class SVM for outlier detection

The separating hyperplane that maximizes the margin is obtained solving the following quadratic optimization problem:

$$\begin{aligned} \min_{w,\{\xi_i\},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 - b + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\ \text{s. t.} \quad & \boldsymbol{w}^T\boldsymbol{\Phi}(\boldsymbol{x}_i) \geq b - \xi_i \quad i = 1,\ldots,n \\ & \xi_i \geq 0, \quad i = 1,\ldots,n \end{aligned} \tag{1}$$

where ξ_i are the slack variables, $\nu \in [0,1]$ is a regularization parameter that represents the fraction of outliers and b is the decision value that determines if a point belongs to the high density region.

The optimization problem can be solved efficiently in the dual space where it is quadratic.

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \sum_{i,j}\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j) - \sum_i \alpha_i k(\boldsymbol{x}_i,\boldsymbol{x}_i) \\ \text{s. t.} \quad & 0 \leq \alpha_i \leq \frac{1}{\nu m} \\ & \sum_i \alpha_i = 1 \end{aligned} \tag{2}$$

(3)

The discriminant function can be expressed exclusively in terms of scalar products [11]. Therefore, the support of any density function can be estimated considering an appropriate kernel that induces a non-linear mapping function to feature space [10].

The one class-classification algorithm exhibits several interesting properties for the problems considered in this paper.

- The regularization term allow us to handle high dimensional and noisy problems overcoming the 'curse of dimensionality'.
- The optimization problem may be written exclusively in terms of kernel evaluations. Therefore, one-class classification may be applied to non-vectorial datasets provided a kernel is defined.
- The optimization problem can be solved efficiently and will converge to a global minimum.

III. AN ALGORITHM TO DETECT MISLABELED SAMPLES

In this section we present the algorithm proposed to detect wrong labels. First samples are transformed to a feature space in an appropriate manner such that mislabeled samples can be detected as outliers. Next, the one-class classification algorithm is applied. Finally, unreliable labels are reported.

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathcal{X}\times\mathcal{Y}$ be the training set, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and d is the dimension of input space. For simplicity, in this paper we consider binary classification problems where the labels $y_i \in \{-1, 1\}$. Figure 2 shows how mislabelled samples can be identified as outliers.

First, for each $\boldsymbol{x}_i \in \mathbb{R}^d$ we compute the k-nearest neighbors in $\mathbb{R}^d$. If $\boldsymbol{x}_i$ is a mislabelled sample, the first nearest neighbors in input space will have different labels. Now $(\boldsymbol{x}_i, y_i)$ is represented by the distances to the nearest neighbors in $\mathbb{R}^d$,

$$\begin{aligned} \phi(\boldsymbol{x}_i, y_i) = (d(\boldsymbol{x}_i, \boldsymbol{x}_{\sigma_1(\boldsymbol{x}_i)}), , \ldots, d((\boldsymbol{x}_i, \boldsymbol{x}_{\sigma_k(\boldsymbol{x}_i)}), \\ d(y_i, y_{\sigma_1(\boldsymbol{x}_i)}), \ldots d(y_i, y_{\sigma_k(\boldsymbol{x}_i)})) \in \mathbb{R}^{2k} \end{aligned} \tag{4}$$

where d is the Euclidean distance and $\sigma_j(\boldsymbol{x})$ is a function that provides the index of the j nearest neighbor of $\boldsymbol{x}$. Considering this representation, mislabelled samples will have large values in $\mathcal{Y}$ while the other samples will map close to the origin. Therefore, in $\mathbb{R}^{2k}$ detecting mislabelled samples is equivalent to outlier detection.

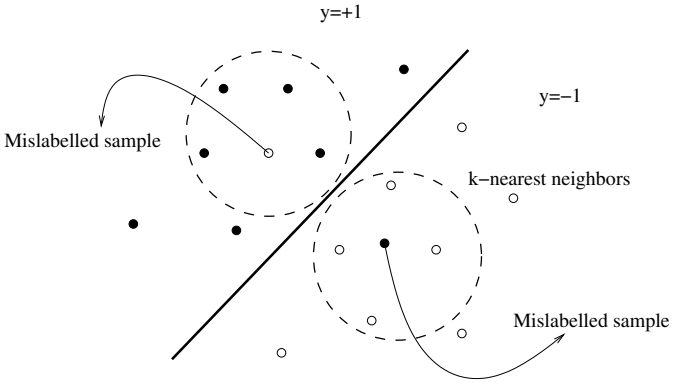

Fig. 2. Diagram representing how mislabelled samples can be detected as outliers.

One-class classification can be applied to detect outliers using the dissimilarity representation via the *empirical kernel map* proposed by [12], [13]. Now we define briefly this kernel:

Let d: $\mathcal{X}\times\mathcal{X} \rightarrow \mathbb{R}$ be a dissimilarity and $R = \{p_1,\ldots,p_k\}$ a subset of representatives drawn from the training set. Define the mapping $\phi : \mathcal{F} \rightarrow \mathbb{R}^n$ as:

$$\phi(z) = D(z,R) = [d(z,p_1), d(z,p_2), \ldots, d(z,p_k)] \tag{5}$$

This mapping defines a dissimilarity space where feature i is given by $d(.,p_i)$. The set of representatives R determines the dimensionality of the feature space.

TABLE I

TWO CLASS MICROARRAYS DATASETS.

Dataset	Number of genes	Class 1	Class 2	Wrong labels	Reference
Colon	2000	$40(T)$	$22(N)$	9	[14]
Breast	7129	25(ER+)	24(ER-)	9	[15]
Pure Colon	2000	$35(T)$	$18(N)$	6	[14]
Pure Breast	7129	21(ER+)	19(ER-)	6	[15]

The kernel of dissimilarities can be defined as the dot product of two dissimilarity vectors in feature space.

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle = \sum_{i=1}^{n} d(\boldsymbol{x}, x_i) d(\boldsymbol{x}', x_i) \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X} . \quad (6)$$

The kernel of dissimilarities maps the samples to a feature space where misclassified objects will be far away from the others. Moreover, wrong labeled samples can be separated in two groups. The first one corresponds to samples that are outliers in feature space but not in input space. They are mislabelled samples with a high degree of confidence. The second group is given by samples that are outliers also in input space. They may correspond to mislabelled examples or may be originated by errors in the data acquisition and experimental protocols.

IV. EXPERIMENTAL RESULTS

The algorithm proposed has been applied to the detection of mislabelled samples in two real microarray classification datasets. Next we comment the properties of the datasets considered and the experimental results obtained.

A. Datasets

Table I shows the main properties for the two microarray datasets considered. Both are binary classification problems with different number of features.

For both datasets several samples have been identified as wrong labeled in the literature. According to [14] the samples T2, T30, T33, T36, T37, N8, N12, N34, N36 are identified as mislabeled in the colon dataset with biological evidences. Similarly, [15] have identified as suspect the samples $11, 14, 16, 31, 33, 45, 46, 40, 43$ in the breast cancer dataset. Both datasets have been used previously by [3], [7] as real benchmarks to test methods for labeling errors.

In order to enhance the reliability of the data source we have built two more data sets removing the outliers reported in the literature from colon and breast. They are called in this section pure colon and pure breast. Next, six labels randomly chosen have been flipped. These datasets will allow to improve the evaluation of the labeling error detection methods.

B. Results

The method proposed has been compared with a standard labeling detection algorithm introduced in [3] and denoted in this section as simple SVM. 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.

Regarding the parameters for the algorithm proposed they are estimated using a ten fold cross-validation strategy. To this aim, we divide the dataset in ten subsets and flip alternatively the labels of the test set. The number of neighbors k and the regularization parameter of one-class classification are estimated to maximize the average number of flipped labels detected. For all the experiments the optimal value for $k = 3$ while for the regularization parameter ν depends on the dataset. Therefore, it is recommended to take $k = 3$ and to optimize only the ν parameter considering a grid centered around 0.5.

No feature selection method has been applied because when the labels noise is high they perform poorly and will not help to improve the detection rate. Finally, the kernel considered for the one-class classification algorithm is linear.

In order to evaluate the algorithms we have computed two standard measures considered by other authors [3], [7], [5], *recall* and *precision*. The first one determines the fraction of mislabeled samples identified by the algorithms. *Precision* gives the fraction of wrong labels reported erroneously by the algorithms. For the application at hand, recall is more relevant because potential wrong labels will be analyzed in depth by human experts before taking a decision.

Table II shows the *recall* and *precision* measures for the two real microarrays datasets in which 9 mislabeled samples have been identified in the literature. According to the *recall* index, our method performs similarly to simple SVM in colon cancer but improves significantly the simple SVM method in breast cancer data. In particular, it identifies 7 out of 9 wrong labels while simple SVM detects only 5 out of 9. Simple SVM provides higher precision values for both datasets but, as we have mentioned this is not relevant for the application at hand.

Table III shows *recall* and *precision* for pure colon and breast cancer datasets. Six sample labels have been flipped randomly and the process is repeated 20 times. Thus, *recall* and *precision* values are averages over 20 independent runs. Although other authors such as [3] have averaged over 50 independent runs, for the data sets considered in this paper the experimental results are the same that for 20 independent runs. Thus, we have chosen the smaller number in order to reduce the computational burden of the experiments.

Table III supports the empirical results mentioned earlier. The *recall* suggests that our method improves significantly

TABLE II

RECALL AND *PRECISION* FOR ONE-CLASS SVM VERSUS SIMPLE SVM.

	Recall		Precision	
Data set	Simple SVM	One-class SVM	Simple SVM	One-class SVM
Colon	0.77	0.66	0.63	0.26
Breast	0.55	0.77	0.45	0.28

9 samples have been identified as mislabelled for Colon and Breast cancer datasets in the literature.

TABLE III

AVERAGE *RECALL* AND *PRECISION* FOR ARTIFICIALLY GENERATED MICROARRAY DATASETS.

	Recall		Precision	
Data set	Simple SVM	One-class SVM	Simple SVM	One-class SVM
Pure colon	0.87	0.83	0.45	0.21
Pure breast	0.50	0.83	1	0.25

Pure colon and pure breast cancer datasets have been generated flipping randomly 6 sample labels in the original datasets.

simple SVM algorithm particularly for Breast cancer dataset. *Precision* is higher for simple SVM but this penalizes strongly the *recall* particularly for breast cancer data. This behavior is not recommended for the application considered in this paper.

V. CONCLUSIONS AND FUTURE RESEARCH

In this paper we have proposed a new algorithm to detect mislabeled samples in cancer classification using the gene expression profiles. The method proposed uses the kernel of dissimilarities to map the data to a feature space where mislabeled samples can be detected as outliers by the one-class classification algorithm. The optimal parameters are estimated using a cross-validation strategy.

The algorithm have been applied to two real standard microarrays datasets for which several samples are reported as suspect in the literature with biological evidences. The experimental results suggest that our method improves significantly a standard detection method such as simple SVM particularly for the *recall* index.

Future research trends will analyze other cancer microarrays datasets and will work to improve *precision* values.

ACKNOWLEDGMENTS

The authors would like to thank Javier De Las Rivas, supervisor of the "Functional Genomics and Bioinformatics" group at the Cancer Research Center of Salamanca (CiC-IBMCC, CSIC) by their useful comments and suggestions.

REFERENCES

[1] Y. Hoshida and et al., Subclass mapping: identifying common subtypes in independent disease data sets, *PLOS ONE*, vol. 11, 2007, pp. 1-8.

[2] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, vol. 16, no. 10, 2000, pp. 906-914.

[3] C. Zhang, C. Wu, E. Blanzieri, Y. Zhou, Y. Wang, W. Du, and Y. Liang, Methods for labeling Error detection in microarrays based on the effect of data perturbation on the regression model, *Bioinformatics*, vol. 25, no. 20, 2009, pp. 2708-2714.

[4] W. Zhang, R. Rekaya, and K. Bertrand, A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer, Bioinformatics, vol. 22, no. 3, pp. 317-325, 2006.

[5] A. Malossini, E. Blanzieri, and R. T. Ng, Detecting potential labeling errors in microarrays by data perturbation, *Bioinformatics*, vol. 22, no. 17, pp. 2114-2121, 2006.

[6] F. Muhlenbach, S. Lallich, and D. A. Zighed, Identifying and handling mislabelled instances, it Journal of Intelligent Information Systems, vol. 22, pp. 89-109, 2004.

[7] J. Bootkrajang, and A. Kabán, Classification of mislabelled microarrays using robust sparse logistic regression, *Bioinformatics*, vol. 0, no. 0, 2011, pp. 1-7.

[8] B. Schölkopff, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, Estimating the support of a high-dimensional distribution. *Neural Computation*, vol. 13, 2001, pp. 1443-1471.

[9] J. M. Moguerza, and A. Muñoz, Support Vector Machines with Applications, *Statistical Science*, vol. 21, no. 3, 2006, pp. 322-336.

[10] V. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, New York; 1998.

[11] B. Schölkopff, A. J. Smola, *Learning with kernels: Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press, Massachussetts; 2002.

[12] K. Tsuda, "Support Vector Classifier with Assymetric Kernel Function", *In Proceedings of ESANN*, Bruges, 1999, pp. 183-188.

[13] E. Pekalska, P. Paclick, and R. Duin, A generalized kernel approach to dissimilarity-based classification, *Journal of Machine Learning Research*, vol. 2, 2001, pp. 175-211.

[14] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *In Proc. Nat'l Acad Sci USA*, vol. 96, 1999, pp. 6745-6750.

[15] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. Olson, J. Marks, and J. Nevins, Predicting the clinical status of human breast cancer by using gene expression profiles, *In Proc. Nat'l Acad Sci USA*, vol. 98, no. 20, 2001, pp. 11462-11467.