Predicting Viral Infection by Selecting Informative Biomarkers From Temporal High-Dimensional Gene Expression Data

Qiang Lou Zoran Obradovic*

Center for Data Analytics and Biomedical Informatics
Temple University
Philadelphia, USA
* corresponding author: zoran.obradovic@temple.edu

Abstract—**In order to more accurately predict an individual's health status, in clinical applications it is often important to perform analysis of high-dimensional gene expression data that varies with time. A major challenge in predicting from such temporal microarray data is that the number of biomarkers used as features is typically much larger than the number of labeled subjects. One way to address this challenge is to perform feature selection as a preprocessing step and then apply a classification method on selected features. However, traditional feature selection methods cannot handle multivariate temporal data without applying techniques that flatten temporal data into a single matrix in advance. In this study, a feature selection filter that can directly select informative features from temporal gene expression data is proposed. In our approach we measure the distance between multivariate temporal data from two subjects. Based on this distance, we define the objective function of temporal margin based feature selection to maximize each subject's temporal margin in its own relevant subspace. The experimental results on two real flu data sets provide evidence that our method outperforms the alternatives, which flatten the temporal data in advance.**

Keywords-high dimensional; temporal data; feature selection; margin; multivariate functional data

I. INTRODUCTION

Microarray technology has the ability to simultaneously measure expression levels of thousands of genes for a given biological sample. There is often interest in the analysis of dynamic biological processes with data from DNA gene expression microarray chips. In order to predict an individual's health status, it is very helpful to analyze such high dimensional gene expression data that varies with time. There are two major challenges in prediction from such temporal microarray data. One is dealing with small-sample high-dimensional data where the number of biomarkers used as features is typically much larger than the number of labeled subjects. A common way to address this problem is to perform feature selection methods as a preprocessing step, followed by a classification method on selected features to predict the health status of an individual.

Another challenge of analyzing dynamic biological processes is that the data gathered is temporal. For example, in the two real flu data sets we used in experiments section, the data records for each individual are multivariate time series. The whole data set consists of many such multivariate time series from different individuals. However, traditional feature selection methods cannot handle such multivariate time series data. The most straightforward method of handling this is to apply some techniques to flatten the temporal data, and then perform traditional feature selection methods in the flattened data. Obviously, the flattening process may result in loss of some information among temporal data. Such straightforward methods tend to select features that are not informative enough.

In this study, we proposed a feature selection filter that can directly select informative features from temporal high-dimensional biomarkers. We defined a temporal margin for each subject based on a measure of distance between two multivariate time series data from two different subjects. The objective function of the proposed selection method is to maximize each subject's temporal margin in its own relevant subspace. We applied stochastic gradient ascent to solve the optimization problem and get the optimal weight for each feature. Features with large weights are selected to build the prediction model to predict the health status of each individual. The experimental results show that our method outperforms the alternatives, which apply traditional feature selection methods after flattening the temporal multivariate gene expression data.

II. RELATED WORK

Feature selection methods can be broadly categorized into filtering models [1] and wrapper models [2]. Filtering methods separate the feature selection from the learning process, whereas wrapper methods combine them. The main drawback of wrapper methods is their computational inefficiency.

There are three widely used kinds of filtering methods. In [3, 5] a margin-based method is proposed as a feature-weighting algorithm that is a new interpretation of a RELIEF-based method [4]. The method in [5] is an online algorithm that solves a convex optimization problem with a margin-based objective function. Markov Blanket-based methods [1, 6, 7] perform feature selection by searching an optimal set of features using Markov Blanket approximation. Dependence estimation-based methods use the Hilbert-Schmide Independence Criterion as a measure of dependence between the features and the labels [8]. However, all these methods assume that the data is static without varying on time. They cannot be applied in temporal gene expression data that is the main problem of this study.

Several feature learning methods [9, 10] have recently been proposed to handle the temporal gene expression data, without flatting the data in advance. However, those two methods are different from the proposed method in this study. First, those methods treat the records for an individual at different time steps independently, which will result in loss of temporal information among the data. Secondly, all those works project the data to another space and learn features from the new space (factors or principal component). Those methods are actually methods for dimension reduction, rather than feature selection. Due to this, we will not compare our method with them in this study.

III. PROPOSED METHOD

Let $\mathbf{D} = \{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1,\ldots,I} \subset \Re^{n \times Ti} \times \pm 1$ be the data set with I individuals. $\mathbf{X}_i \in \Re^{n \times Ti}$ represents n observed biomarkers (e.g. gene expression data) for individual i measured at T_i time steps. $\mathbf{Y}_i \in \{1,-1\}$ represents the class label (e.g. health status) for individual i. Let $\mathbf{X}_i^{(r)}$ be the r^{th} column of $\mathbf{X}_\text{i}$ that corresponds n biomarkers measured at time t_r.

A. Measure Distance Among Multivairte Time Series

Given $\mathbf{X}_\text{i}$, $\mathbf{X}_\text{j}$ corresponding to the observed biomarkers measured at different time steps for individual i and individual j, respectively, the distance (we call Temporal distance, represented as *Tdist*) between two multivariate time series $\mathbf{X}_\text{i}$ and $\mathbf{X}_\text{j}$ is defined as:

$$Tdist(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{T_i \times T_j} \sum_{r=1}^{T_i} \sum_{s=1}^{T_j} d(\mathbf{X}_i^{(r)}, \mathbf{X}_j^{(s)})$$

where T_i and T_j are the number of time steps of individual i and individual j, respectively; $\mathbf{X}_i^{(r)}$ is the vector consists of biomarkers measured at time steps r for individual i; $\mathbf{X}_i^{(s)}$ is the vector of biomarkers measured at time steps s for individual j; for any two vectors $\mathbf{v}$ and $\mathbf{z}$, function $d(\mathbf{v}, \mathbf{z})$ is defined as $d(\mathbf{v}, \mathbf{z}) = \sqrt{\sum_p (\mathbf{v}_p - \mathbf{z}_p)^2}$.

B. Maximize Temporal Margin

Given an instance, the margin of a hypothesis is the distance between the hypothesis and the closest hypothesis that assigns an alternative label. For a given instance $\mathbf{X}_\text{i}$, we find two nearest neighbors for $\mathbf{X}_\text{i}$, one with the same class label (called ***nearhit***), and the other with different class label (called ***nearmiss***). The hypothesis-margin of a given instance $\mathbf{X}_\text{i}$ in data set D is defined as:

$$L_D(\mathbf{X}_i) = \frac{1}{2}(Tdist(\mathbf{X}_i, nearmiss(\mathbf{X}_i)) - Tdist(\mathbf{X}_i, nearhit(\mathbf{X}_i)))$$

In margin-based feature selection, we scale the feature by assigning a non-negative weight vector ***w***, and then choose the features with large weights that maximize the margin. One idea is to then calculate the margin in weighted feature space rather than the original feature space, since the nearest neighbor in the original feature space can be completely different from the one in the weighted feature space. Therefore, we define the instance margin for each instance $\mathbf{X}_\text{i}$ from D in a weighted feature space as:

$$\rho_D(\mathbf{X}_i \mid \mathbf{w}) = \frac{1}{2}(Tdist(\mathbf{X}_i, nearmiss(\mathbf{X}_i) \mid \mathbf{w}) - Tdist(\mathbf{X}_i, nearhit(\mathbf{X}_i) \mid \mathbf{w}))$$

which is equivalent to:

$$\rho_D(\mathbf{X}_i \mid \mathbf{w}) = \frac{1}{2T_i \times T_j} \sum_{r=1}^{T_i} \sum_{s=1}^{T_M} d(\mathbf{X}_i^{(r)}, nearmiss(\mathbf{X}_i)^{(s)} \mid \mathbf{w}) - \frac{1}{2T_i \times T_j} \sum_{r=1}^{T_i} \sum_{s=1}^{T_H} d(\mathbf{X}_i^{(r)}, nearhit(\mathbf{X}_i)^{(s)} \mid \mathbf{w})$$

where T_M and T_H are the number of time steps of nearmiss($\mathbf{X}_\text{i}$) and nearhit($\mathbf{X}_\text{i}$), respectively; for any two vectors $\mathbf{v}$ and $\mathbf{z}$, function $d(\mathbf{v}, \mathbf{z}|\mathbf{w})$ is defined as:

$$d(\mathbf{v}, \mathbf{z} \mid w) = \sqrt{\sum_p (\mathbf{v}_p - \mathbf{z}_p)^2 w_p^2}.$$

We already define the instance margin for each subject $\mathbf{X}_\text{i}$. Therefore, we can define the temporal margin of the entire data D that has I subjects as the sum of all instance margins, which can be written as:

$$\rho_{D|\mathbf{w}} = \sum_{i=1}^{I} \rho_D(\mathbf{X}_i \mid \mathbf{w})$$

The feature weights can be learned by solving an optimization problem that maximizes the temporal margin of entire data D. Therefore the most informative features can be chosen based on the feature weights learned. The bigger weight a feature has, the more important the feature is. This optimization problem can be represented as:

$$\max_{\mathbf{w}} \sum_{i=1}^{I} \rho_D(\mathbf{X}_i \mid \mathbf{w})$$

which is equivalent to:

$$\max_{\mathbf{w}} \sum_{i=1}^{I} \left(\frac{1}{2T_i \times T_j} \sum_{r=1}^{T_i} \sum_{s=1}^{T_M} d(\mathbf{X}_i^{(r)}, nearmiss(\mathbf{X}_i)^{(s)} \mid \mathbf{w}) - \frac{1}{2T_i \times T_j} \sum_{r=1}^{T_i} \sum_{s=1}^{T_H} d(\mathbf{X}_i^{(r)}, nearhit(\mathbf{X}_i)^{(s)} \mid \mathbf{w})\right)$$

We solve this optimization problem to get the optimal weights $\mathbf{w}$ by applying stochastic gradient ascent. The gradient of $\rho_{D|\mathbf{w}}$ when evaluated on a data D is:

$$(\nabla \rho_{D|\mathbf{w}})_i = \frac{\partial \rho_{D|\mathbf{w}}}{\partial \mathbf{w}_i} = \sum_{j=1}^{I} \frac{\partial \rho_D(\mathbf{X}_j)}{\partial \mathbf{w}_i}$$

$$= \frac{1}{2} \sum_{j=1}^{I} \left(\frac{\frac{1}{T_i \times T_M} \sum_{r=1}^{T_i} \sum_{s=1}^{T_M} |\, X_i^{(r)} - nearmiss(X_i)^{(r)} \,|_j^2}{Tdist(\mathbf{X}_i, nearmiss(\mathbf{X}_i) \mid \mathbf{w})} - \frac{\frac{1}{T_i \times T_H} \sum_{r=1}^{T_i} \sum_{s=1}^{T_H} |\, X_i^{(r)} - nearhit(X_i)^{(r)} \,|_j^2}{Tdist(\mathbf{X}_i, nearhit(\mathbf{X}_i) \mid \mathbf{w})}\right)$$

where, for two vectors **v** and **z**, the function $|\mathbf{v}-\mathbf{z}|_j^2$ is defined as: $|\mathbf{v}-\mathbf{z}|_j^2=(\mathbf{v}_j-\mathbf{z}_j)^2$.

C. Feature Selection Algorithm

In this section we will introduce our feature selection method, which solves the optimization problem introduced in previous section.

The proposed algorithm for **Feature Selection** in **Temporal** microarray data (we call it **FST**) is shown in Table 1. The **FST** algorithm starts with initializing the values of **w** to be 1. With such initialization we can estimate the instance margin for each instance $\mathbf{X}_i$. Then, in each iteration, the weights vector **w** is updated by solving the optimization problem introduced in previous section. We repeat the iteration until convergence or using all instances to update the weights.

TABLE I. **FST** FEATURE SELECTION METHOD

Input: data set D = $\{\mathbf{X}_i, y_i\}_{i=1,...,I}$
Output: feature weights *w*
Initialization: set $w^{(0)}=1$, t = 1
For each subject $\mathbf{X}_i$, Do
Calculate Tdist($\mathbf{X}_i$, $\mathbf{X}_r$|$\mathbf{w}^{(t-1)}$) when r ≠i
Calculate nearmiss($\mathbf{X}_i$) and nearhit($\mathbf{X}_i$)
For each dimensionality j, Do
Calculate ∇_j
End For
t = t + 1;
$\mathbf{w}(t) = \mathbf{w}(t-1) + \nabla$
End For
$\mathbf{w} \leftarrow \mathbf{w}^2/\|\mathbf{w}^2\|$

IV. EXPREIMENTS

To characterize the proposed algorithm, we conducted large-scale experiments on 2 real flu data sets [9, 10]. In summary, H3N2 data consists of records of 17 subjects collected at 16 different time steps. H1N1 data consists of records of 24 subjects collected at 16 different time steps. For H3N2 and H1N1 gene expression data, the same 12,023 genes are considered for analysis for each subject at each time step.

All experiments of this study were performed on a PC with 3 GB of memory. We compared our proposed **FST** algorithm in temporal gene expression data with three traditional feature selection methods (the method proposed in [1] that we call **FCBF**, **HSMB** [6] and **Relief** [11]) after flattening temporal multivariate data into one single matrix. For the prediction method, we apply a Nearest Neighbor classifier on all features and select features by different feature selection methods.

Since we don't know in advance which genes among these two datasets are deciding an individual's health status, we evaluate our method and three alternatives in a different way than that applied to Synthetic data. We apply all methods on both data sets, and build the prediction model on selected genes. We compare the accuracy of the prediction models built from different methods. We believe that the selected features tend to be more correct if the prediction model built on these features is more accurate.

For the feature selection and learning-prediction process, we apply leave-one-out schema because of the low number of subjects in both two data set. To avoid overfitting, in each iteration of leave-one-out schema, the training set is used to perform feature selection and learn the prediction model, and the one test subject is only touched in prediction process. We applied a Nearest Neighbor classifier to build the prediction model because it is easy to perform on multivariate temporal gene expression data sets.

The results on H3N3 and H1N1 data sets are listed in Table II and Table III. Since H1N1 data set is imbalanced data (8 negative subjects and 16 positive subjects). We report sensitivity, specificity, and balanced accuracy to evaluate the results from all methods.

The classification results on H3N3 and H1N1 are shown at the top sub-table of Table II and Table III. The number of selected features from different methods are shown at the bottom sub-table of Table II and Table III. **FCBF** and **HSMB** can automatically select the optimal set of features, whereas **Relief** and **FST** are feature ranking features. For comparison, we let **Relief** and **FST** selects the same number of features as the bigger one among the number of features **FCBF** and **HSMB** returns automatically. We repeat experiments 20 times and report the mean ± std values for classification results (sensitivity, specificity, and balanced accuracy).

Table II shows the results on H3N3 data. We can see there that the accuracy of predictor built on the features selected by out proposed **FST** method outperforms all alternatives including the predictor built on all features. This proves that our **FST** method selects more accurate features. The bottom sub table of Table II shows that **FCBF** selects the smallest number of features among all methods, which is consistent to the one of widely know drawbacks of **FCBF**: **FCBF** tend to remove features too aggressively.

We got similar results, shown in Table III, on H1N1 to the results on H3N2. Moreover, H1N1 is an unbalanced dataset (with large fraction of positive subjects). We can see from Table III that if we build a predictor on all features, we will tend to predict most negative subjects as positive subjects. The specificity results from **FCBF**, **HSMB** and **Relief** are also small, because they didn't select most informative features. The predictors built on these selected features suffered from imbalanced data, and treated most negative subjects as positive subjects.

V. CONCLUSION

There are two major challenges in predicting an individual's health status from multivariate temporal gene expression data. One is small-sample high-dimensional microarray data where the number of biomarkers used as features is typically much larger than the number of labeled subjects, and the other is the temporal property of the data, from which traditional feature selection cannot be applied directly. To address these two challenges, in this study, we proposed a feature selection filter that can directly select informative genes from temporal high-dimensional

TABLE II. RESULTS ON H3N2 DATA

	All feature	FCBF	HSMB	Relief	FST
Sensitivity	0.667 ± 0	0.755 ± 0.242	0.750 ± 0.175	0.875 ± 0.063	**1.000** ± 0
Specificity	0.811 ± 0	0.556 ± 0.046	0.667 ± 0.135	0.778 ± 0.118	**0.889** ± 0.130
Balanced_Accuracy	0.771 ± 0	0.653 ± 0.149	0.708 ± 0.162	0.826 ± 0.065	**0.944** ± 0.064

(a) Classification Accuracy (mean ± std)

FCBF	HSMB	Relief	FST
15	50	Top 50	Top 50

(b) Number of Selected Features

TABLE III. RESULTS ON H1N1 DATA

	All feature	FCBF	HSMB	Relief	FST
Sensitivity	0.938 ± 0	0.813 ± 0.068	0.813 ± 0.136	1.000 ± 0	**1.000** ± 0
Specificity	0.125 ± 0	0.375 ± 0.146	0.500 ± 0.128	0.500 ± 0.132	**0.750** ± 0.151
Balanced_Accuracy	0.531 ± 0	0.594 ± 0.102	0.656 ± 0.065	0.750 ± 0.074	**0.875** ± 0.101

(a) Classification Accuracy(mean ± std)

FCBF	HSMB	Relief	FST
23	43	Top 43	Top 43

(b) Number of Selected Features

biomarkers. For each subject , we defined a temporal margin based on a measure of distance between two multivariate time series data from two different subjects. The objective function of the proposed selection method is to maximize each subject's temporal margin in its own relevant subspace. To solve the optimization problem and get the optimal weight for each feature, the stochastic gradient ascent is applied. Informative features are those with large weights after optimizing the objective function. The prediction model is build on selected informative genes to predict the health status of each individual. The experimental results show that our method outperforms the alternatives, which apply traditional feature selection methods after flattening the temporal multivariate gene expression data.

ACKNOWLEDGMENT

This work was supported in part by DARPA grant DARPA-N66001-11-1-4183 negotiated by SSC Pacific (to ZO).

REFERENCES

[1] Yu, L., and Liu, H.. Feature Selection for High-dimensional Data, A Fast Correlation-based Filter Solution. In 20th *International Conference on Machine Learning*, 2003, pp. 856-863

[2] Kohave, R. and John, G.. Wrappers for Feature Subset Selection. *Artificial Intelligence*, vol 1-2, 1997, pp. 273-324

[3] Lou, Q, and Obradovic, Z. (in press) "Margin-Based Feature Selection in Incomplete Data,". In Proc. Of 26th AAAI Conference on Artificial Intelligence (AAAI-12). July 2012, Toronto, Ontario, Canada

[4] Gilad-Bachrach, R., Freund, Y., Bartlett, P. L. and Lee, W. S.. Margin Based Feature Selecgtion - theory and algorithms. In 21st *International Conference on Machine Learning*, 2004, pp. 43-50

[5] Sun, Y., Todorovic, S. and Goodison, S. Local Learning Based Feature Selection for High Dimensional Data Analysis. *IEEE Transactions on Pattern Analysis and Machine Learning*, 2009.

[6] Lou, Q, and Obradovic, Z, "Feature Selection by Approximating the Markov Blanekt in a Kernel-Induced Space", Proc. 19th European Conference on Artificial Intelligence, Lisbon, Portugal, 2010,

[7] Lou, Q, parkman, H.P., Jacobs, M.R, Krynetskiy, E. and Obradovic, Z. "Exploring Genetic Variability in Drug Therapy by Selecting a Minimum Subset of the Most Informative Single Nucleotide Polymorphisms through Approximation of a Markov Blanket in a Kernel-induced Space," Proc. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, San Deigo, CA, May 2012.

[8] Song, L, Smola, A, Gretton, A. and Borgwardt, K.L.. "A dependence maximization view of clustering", *In Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

[9] Chen, B., Chen, M., Paisley, J., Zass, A., Woods, C., Ginsburg, G.S.,Lucas, J., Dunson, D., and Carin, L. "Bayesian Inference of the Number of Factors in Gene-expression Analysis: Application to Human Virus Challenge Studies,", BMC bioinformatics, 2010.

[10] Chen, M., Zaas, A., Woods, C., Ginsburg, G.S., Lucas, J., Dunson, D., and Carin, L. "Predicting Viral Infection From High-Dimensional Biomarkers Trajectories" Journal of the American Statistical Association, vol. 106, No. 496. December 2011.

[11] Sun, J. and Li, J. "Iterative RELIEF for Feature Weighting." In 23rd International Conference on Machine Learning, 2006.