2012 IEEE 12th International Conference on Data Mining

Analysis of Temporal High-Dimensional Gene Expression Data for Identifying Informative Biomarker Candidates

Qiang Lou
Department of Computer and Information Science
Center for Data Analytics and Biomedical Informatics
Temple University
Philadelphia, USA
qianglou@temple.edu

Zoran Obradovic
Department of Computer and Information Science
Center for Data Analytics and Biomedical Informatics
Temple University
Philadelphia, USA
zoran.obradovic@temple.edu

***Abstract*—Identifying informative biomarkers from a large pool of candidates is the key step for accurate prediction of an individual's health status. In clinical applications traditional static feature selection methods that flatten the temporal data cannot be directly applied since the patient's observed clinical condition is a temporal multivariate time series where different variables can capture various stages of temporal change in the patient's health status . In this study, in order to identify informative genes in temporal microarray data, a margin based feature selection filter is proposed. The proposed method is based on well-established machine learning techniques without any assumptions about the data distribution. The objective function of temporal margin-based feature selection is defined to maximize each subject's temporal margin in its own relevant subspace. In the objective function, the uncertainty in calculating nearest neighbors is taken into account by considering the change in feature weights in each iteration. A fixed-point gradient descent method is proposed to solve the formulated objective function. The experimental results on both synthetic and real data provide evidence that the proposed method can identify more informative features than the alternatives that flatten the temporal data in advance.**

Keywords- high dimensional; temporal data; feature selection; margin; multivariate time series data

I. INTRODUCTION

The major challenges in analyzing microarray data 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. Performing feature selection methods as a preprocessing step to identify informative biomarkers is a common way to address this problem. It is often followed by a classification method on selected genes to predict the health status of an individual.

However, there is often interest in the analysis of dynamic biological processes with data from DNA gene expression microarray chips instead of analyzing static gene expression data. 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. Besides the traditional challenge of curse of dimensionality, another challenge of analyzing dynamic biological processes is that the data gathered is temporal. Therefore, the data records for each individual are multivariate time series. However, traditional feature selection methods cannot handle such multivariate time series data. The most straightforward method is to apply some techniques to flatten the temporal data, and then perform traditional feature selection methods in the flattened data.

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 also take into account the uncertainty in calculating nearest neighbors because the feature weights change in each iteration, and it is hard to calculate nearest neighbors for a multivariate time series data. We applied fixed-point gradient ascent to solve the optimization problem and get the optimal weight for each gene. Genes 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. Convergence theorem of the proposed method is also presented.

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. The method proposed at [6]

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.92

approximates the Markov Blanket by applying a Grow-Shrink process and then removing the feature whose Markov Blanket can be found in the rest of features. Method [7] searches Markov Blanket after learning the structure of the Bayesian network.

Dependence estimation-based methods use the Hilbert-Schmidt Independence Criterion as a measure of dependence between the features and the labels [8]. The key idea in this method is that good features should maximize such dependence. 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 imputing missing values in advance. However, those two methods are different from the proposed method in this study, since those methods treat the records for an individual at different time steps independently, which will result in loss of temporal information among the data. 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.

The method proposed in this study extends our feature ranking method addressing the similar problem [13]. Previously, we measured nearest neighbors in the original space and did not update them while updating the feature weights. Both aspects are generalized here. In addition, our method proposed in [13] directly applies gradient descent optimization, which cannot guarantee convergence to a global solution, whereas a globally optimized solution is guaranteed when using the methods proposed here.

III. PROPOSED METHOD

Let $\mathbf{D} = \{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1,\ldots,N} \subset \Re^{n \times T_i} \times \pm 1$ be the data set with N 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.

We will first define the measure of distance between multivariate time series data of two subjects, and then present the temporal margin based on the distance measure as well as the objective function of proposed feature selection method and algorithm for solving the corresponding optimization problem.

A. Measure Distance Among Multivariate Time Series

Given $\mathbf{X}_\text{i}$, and $\mathbf{X}_\text{j}$ 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:

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

where $\mathrm{T_i}$ and $\mathrm{T_j}$ 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 $d(\mathbf{v}, \mathbf{z})$ can be any kind of distance function. To keep the notation simple, we defined $d(\mathbf{v}, \mathbf{z})$ as the Manhattan distance between two vectors.

B. Maximize Temporal Margin With Uncertainty

Given an instance, the margin of a hypothesis is the distance between the hypothesis and the closest hypothesis that assigns an alternative label [4]. 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:

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

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:

$$\rho_D(\mathbf{X}_i \mid \mathbf{w}) = \frac{1}{2}(Tdist(\mathbf{X}_i, nearmiss(\mathbf{X}_i) \mid \mathbf{w}) - Tdist(\mathbf{X}_i, nearhit(\mathbf{X}_i) \mid \mathbf{w})) \quad (3)$$

which is equivalent to:

$$\rho_D(\mathbf{X}_i \mid \mathbf{w}) = \frac{1}{2T_i \times T_M} \sum_{r=1}^{T_i} \sum_{s=1}^{T_M} d(\mathbf{X}_i^{(r)}, nearmiss(\mathbf{X}_i)^{(s)} \mid \mathbf{w}) - \frac{1}{2T_i \times T_H} \sum_{r=1}^{T_i} \sum_{s=1}^{T_H} d(\mathbf{X}_i^{(r)}, nearhit(\mathbf{X}_i)^{(s)} \mid \mathbf{w}) \quad (4)$$

$$= \mathbf{w}^T \boldsymbol{\beta}_i$$

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 each instance $\mathrm{X_i}$, the corresponding $\boldsymbol{\beta}_i$ is defined as:

$$\boldsymbol{\beta}_i = \frac{1}{2T_i \times T_M} \sum_{r=1}^{T_i} \sum_{s=1}^{T_M} | \mathbf{X}_i^{(r)} - nearmiss(\mathbf{X}_i)^{(s)} | - \frac{1}{2T_i \times T_H} \sum_{r=1}^{T_i} \sum_{s=1}^{T_H} | \mathbf{X}_i^{(r)} - nearhit(\mathbf{X}_i)^{(s)} | \quad (5)$$

where $|\cdot|$ is the element-wise absolute operator.

One possible problem may exist in the current definition of instance margin. The nearest neighbors we calculate for each instance might not be the real nearest neighbors, since we calculate the nearest neighbor for each instance in the weighted space that changes each time when the weights get updated. To solve this problem, we take into account the uncertainty of calculating nearest neighbors when calculating instance margins. We calculate the uncertainty of each instance being the nearest neighbor of $\mathbf{x}_n$. The uncertainty is evaluated by standard Gaussian kernel estimation with kernel width of σ. Specifically, we define the uncertainty that an instance $\mathbf{x}_i$ with the same class label as $\mathbf{x}_n$ can be the nearest hit neighbor of $\mathbf{x}_n$ as:

$$U_{\text{nearhit}}(\mathbf{x}_i \mid \mathbf{x}_n, \mathbf{w}) = \frac{\exp(\frac{1}{2T_i \times T_M}\sum_{r=1}^{T_i}\sum_{s=1}^{T_M} d(\mathbf{X}_i^{(r)}, (\mathbf{X}_n)^{(s)} \mid \mathbf{w})/\sigma)}{\sum_j \exp(\frac{1}{2T_i \times T_M}\sum_{r=1}^{T_i}\sum_{s=1}^{T_M} d(\mathbf{X}_j^{(r)}, (\mathbf{X}_n)^{(s)} \mid \mathbf{w})/\sigma)}$$
$$\text{where } 1 \le i \le N, i \ne n, y_i = y_n$$
$$\text{and } 1 \le j \le N, y_j = y_n \quad (6)$$

Similarly, the uncertainty that an instance x_i with a different class label from $\mathbf{x}_n$ can be the nearest miss neighbor of $\mathbf{x}_n$ is defined as:

$$U_{\text{nearmiss}}(\mathbf{x}_i \mid \mathbf{x}_n, \mathbf{w}) = \frac{\exp(\frac{1}{2T_i \times T_H}\sum_{r=1}^{T_i}\sum_{s=1}^{T_H} d(\mathbf{X}_i^{(r)}, (\mathbf{X}_n)^{(s)} \mid \mathbf{w})/\sigma)}{\sum_j \exp(\frac{1}{2T_i \times T_H}\sum_{r=1}^{T_i}\sum_{s=1}^{T_H} d(\mathbf{X}_j^{(r)}, (\mathbf{X}_n)^{(s)} \mid \mathbf{w})/\sigma)}$$
$$\text{where } 1 \le i \le N, y_i \ne y_n$$
$$\text{and } 1 \le j \le N, y_j \ne y_n \quad (7)$$

Please note that distance in equations (6) and (7) denotes the distance between $\mathbf{x}_n$ and $\mathbf{x}_i$ in weighted space determined by weight vector $\mathbf{w}$. Finally, by checking the uncertainty of each instance to be the nearest neighbor of $\mathbf{x}_n$, we define our final **temporal margin with uncertainty** as the expectation of the instance margin of $\mathbf{x}_n$, which can be written as:

$$E_{\rho_D}(\mathbf{x}_n \mid \mathbf{w}) = \mathbf{w}^T \mathbf{E}_{\beta_n}$$
where
$$\mathbf{E}_{\beta_n} = \sum_{i,when\ y_i \ne y_n} U_{\text{nearmiss}}(\mathbf{x}_i \mid \mathbf{x}_n, \mathbf{w}) \cdot \frac{1}{2T_i \times T_M}\sum_{r=1}^{T_i}\sum_{s=1}^{T_M} \mid \mathbf{X}_i^{(r)} - nearmiss(\mathbf{X}_i)^{(s)} \mid$$
$$- \sum_{i,when\ y_i = y_n} U_{\text{nearhit}}(\mathbf{x}_i \mid \mathbf{x}_n, \mathbf{w}) \cdot \frac{1}{2T_i \times T_H}\sum_{r=1}^{T_i}\sum_{s=1}^{T_H} \mid \mathbf{X}_i^{(r)} - nearhit(\mathbf{X}_i)^{(s)} \mid \quad (8)$$

As we mentioned before, our temporal margin incorporates the uncertainty in calculating two nearest neighbors ($E_{\beta n}$).

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

$$\rho_{D|\mathbf{w}} = \sum_{n=1}^{N} \mathbf{E}_{\rho_D}(\mathbf{x}_n \mid \mathbf{w}) \quad (9)$$

The feature weights can be learned by solving an optimization problem that maximizes the uncertainty margin of data D. This optimization problem can be represented as:

$$\max_{\mathbf{w}} \sum_{n=1}^{N} E_{\rho_D}(\mathbf{x}_n \mid \mathbf{w}) \quad \text{subject to } \mathbf{w} \ge 0 \quad (10)$$

We followed logistic regression formulation framework. In order to avoid huge values in weight vector $\boldsymbol{w}$, we add a normalization condition $\| \mathbf{w} \|_1 \le \theta$. Therefore, we can rewrite the optimization problem as:

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \log(1 + \exp(-E_{\rho_D}(\mathbf{x}_n \mid \mathbf{w})) \quad \text{subject to } \mathbf{w} \ge 0, \| \mathbf{w} \|_1 \le \theta \quad (11)$$

The above formulation is called nonnegative garrote. We can rewrite the formulation as:

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \log(1 + \exp(-\mathbf{w}\mathbf{E}_{\beta_n}) + \lambda \| \mathbf{w} \|_1 \quad \text{subject to } \mathbf{w} \ge 0 \quad (12)$$

For each solution to (12), there is a parameter θ, corresponding to the obtained λ in (12), which gives the same solution in (11). Formulation (12) is actually the optimization problem with ℓ_1 regularization. The benefits of adding the ℓ_1 penalty have been well studied [12] and it is shown that the ℓ_1 penalty can effectively handle sparse data and huge amounts of irrelevant features.

C. *Feature Selection Algorithm*

In this section we will introduce our feature selection method, which solves the optimization problem introduced in Section 3.2. As we can see from (12), the optimization problem is convex if $E_{\beta n}$ is fixed. Fox a fixed $E_{\beta n}$, (12) is a constrained convex optimization problem. However, it cannot be directly solved by gradient descent because of the nonnegative constraints on $\boldsymbol{w}$. To handle this problem, we introduce a mapping function:

$$f: \mathbf{w} \to \mathbf{u}, \text{ where } \mathbf{w}(i) = \mathbf{u}(i)^2, \ \forall i = 1, 2, ...M \quad (13)$$

Therefore, the formulation (12) can be rewritten as:

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \log(1 + \exp(-\mathbf{w}\mathbf{E}_{\beta_n}) + \lambda \| \mathbf{u} \|_2^2 \quad (14)$$

By taking the derivative with respect to $\boldsymbol{u}$, we obtain the following updated rule for $\boldsymbol{u}$:

$$\mathbf{u}^{(new)} = \mathbf{u}^{(old)} - \alpha(\lambda - \frac{\sum_{n=1}^{N} \exp(-\sum_{j=1}^{M} \mathbf{u}_j^2 \mathbf{E}_{\beta_n})}{1 + \sum_{n=1}^{N} \exp(-\sum_{j=1}^{M} \mathbf{u}_j^2 \mathbf{E}_{\beta_n})}) \otimes \mathbf{u} \quad (15)$$

where α is learning rate, ⊗ is the Hadamard product, and $\mathbf{E}_{\beta n}$ is defined as:

$$\mathbf{E}_{\beta_n} = \sum_{i,when\ y_i \neq y_n} U_{\text{nearmiss}}(\mathbf{x}_i \mid \mathbf{x}_n, \mathbf{w}\ ) \cdot \frac{1}{2T_i \times T_M} \sum_{r=1}^{T_i} \sum_{s=1}^{T_M} | \mathbf{X}_i^{(r)} - nearmiss(\mathbf{X}_i)^{(s)} |$$

$$- \sum_{i,when\ y_i = y_n} U_{\text{nearhit}}(\mathbf{x}_i \mid \mathbf{x}_n, \mathbf{w}\ ) \cdot \frac{1}{2T_i \times T_H} \sum_{r=1}^{T_i} \sum_{s=1}^{T_H} | \mathbf{X}_i^{(r)} - nearhit(\mathbf{X}_i)^{(s)} |$$

However, $E_{\beta n}$ is determined by $\boldsymbol{w}$ so that (14) is not a convex problem. We use a fixed-point EM algorithm to find the optimal $\boldsymbol{w}$. The proposed algorithm for **Margin-based Feature Selection in Temporal Microarray** data (we call it **MSTM**) is shown in Table 1.

The MSTM algorithm starts by initializing the values of w to be 1. With such initialization, we can estimate the $\mathbf{s}_n$ and $E_{\beta n}$ for each instance $\mathbf{x}_n$. Then, in each iteration, the weights vector w is updated by solving the optimization problem (14) with estimated values of $\mathbf{s}_n$ and $E_{\beta n}$ in the previous iteration. We repeat the iteration until convergence. The MSTM algorithm requires pre-defined kernel width σ and a regularization parameter λ. We applied cross validation to select the values of parameters.

TABLE I. MSTM FEATURE SELECTION METHOD

Input:	data set $D = \{(x_i, y_i)\}_{i=1,...N}$ kernel width σ regularization parameter λ
Output:	feature weights *w*
Initialization:	set $w^{(0)}=1$, t = 1
Do	
	Calculate $E_{\beta n}^{(t)}$ using $w^{(t-1)}$ and equation (8)
	Update $\boldsymbol{u}^{(t)}$ using updated rule in equation (15)
	Update $w^{(t)}$ using $\boldsymbol{u}^{(t)}$ using equation (13)
	t = t + 1
Until convergence	

To prove convergence of MSTM algorithm we will use the following theorem.

Theorem 1 (Contraction Mapping Theorem). *Let T: X→X be a contraction mapping on a complete metric space X. The sequence generated by $x_n=T(x_{n-1})$ for n = 1, 2,3,...converges to unique limit x*, where x* is the fixed point of T (T(x*)=x*). In other words, there is a nonnegative real number r<1 such that*

$$d(x^*, x_{n+1}) \leq \frac{r^n}{1-r} d(x_1, x_0)$$

Proof: See [12].

Based on this theorem we prove the following:

Theorem 2. *There exists σ_0 such that for any $\sigma > \sigma_0$ the MSTM algorithm converges to a fixed unique solution $\boldsymbol{w}^*$ when initial feature weights $\boldsymbol{w}^{(0)}$ are nonnegative.*

The proof is following a similar schema as in [3]. Details are omitted for lack of space.

The complexity of the MSTM algorithm is $O(TN^2M)$ where T is the total number of iterations, *N* is the number of instances, and *M* is the number of features. Our experimental results show that the algorithm converges in a small number of iterations (less than 40). Therefore, the complexity of MSTM algorithm in real application is about $O(N^2M)$. Note that the MSTM algorithm is linear to the number of features, so the proposed method can handle a huge number of features.

IV. EXPERIMENTS

To characterize the proposed algorithm, we conducted large-scale experiments on both synthetic and 2 real flu data sets [9, 10]. All experiments of this study were performed on a PC with 3 GB of memory. We compared our proposed **MSTM** algorithm in temporal gene expression data with four traditional feature selection methods (the method proposed in [12] that we call **BAHSIC**, **SIMBA** [4], **Relief** [11] and **FST** [13]) after flattening temporal multivariate data into one single matrix.

For the prediction method, we apply Nearest Neighbor classifier on all features and select features by different feature selection methods. We compare results on both synthetic data and real data.

A. Results on Synthetic Data

We generate synthetic data simulating 20 subjects. Each subject has 50-dimensional records at 20 different time steps. Each subject *i* is generated according the following process. We first generate 50-dimensional random data $\mathbf{X}_i$ for subject *i* at time step 1. Label Y_i is complete decided by the first four features following $Y_i = (\mathbf{X}_{i1} \vee \mathbf{X}_{i2}) \wedge (\mathbf{X}_{i3} \vee \mathbf{X}_{i4})$. We then generate records for subject *i* at other time steps using formula: $\mathbf{X}_i^{(t+1)} = \mathbf{X}_i^{(t+1)} + \varepsilon$, where $\varepsilon \sim N(0, \frac{t}{10})$ is the Gaussian noise that is also a function of time steps.

The results on synthetic data are shown in Figure 1 and Table II. Figure 1 shows the feature weights for each feature learned by our proposed **MSTM** and three alternatives. It clearly shows that our method assigns significantly larger weights to the first four features used to decide the Label than to most other features. Moreover, our method applied L1 regularization so that the feature weights learned are sparse (most of feature weights are tend to zero).

Table II shows the results comparing our method to three alternatives (three alternatives are applied after flattening the temporal data). We choose the top 4 features selected by each method, and compare the number of features correctly selected among these top 4 features. Top 4 features means 4 features with biggest feature weight. We can see from Table II that our method included all 4 informative features in the top 4 features, whereas **SIMBA** hits 3, **FST** and **Relief** hits only 2. Our method outperforms alternatives on this synthetic data. We didn't apply **BAHSIC** on the synthetic data because **BAHSIC** is not feature weighting method.

TABLE II. NUMBER OF CORRECTLY SELECTED FEATURES AMONG TOP 4 FEATURES

	Relief	**FST**	**Simba**	**MSTM**
# correct features	2	2	3	4

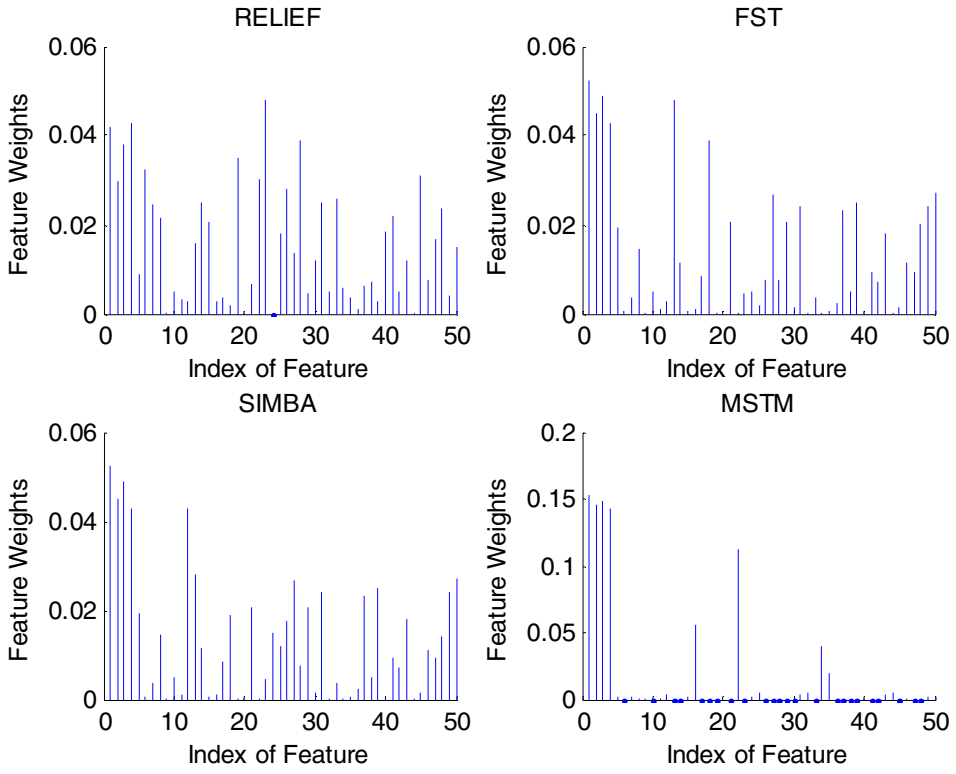

Figure 1. Feature weights learned on synthetic data.

B. Results on Two Real Flu Datasets

We first describe the data sets [9, 10] used in this section. Two challenge studies were performed with two groups of healthy human volunteers. One of these groups was exposed to H3N2 virus, and the other was exposed to H1N1 virus. The H1N1 and H3N2 studies were performed independently, with different subjects. In summary, H3N2 data consists of records for 17 subjects collected at 16 different time steps. H1N1 data consists of records for 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.

For the feature selection and learning-prediction process, we apply leave-one-out schema because of the low number of subjects in both data sets. 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 III and Table IV. Since the 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 balanced accuracy is the average of sensitivity and specificity. Balanced accuracy tends to decrease the chance that the classifier takes advantage of an imbalanced test set.

The classification results on H3N3 and H1N1 are shown at the top sub-table of Table III and Table IV. We repeat experiments 20 times and report the mean ± std values for classification results (sensitivity, specificity, and balanced accuracy). We can see there that the accuracy of the predictor built on the features selected by our proposed **MSTM** method outperforms all alternatives including the predictor built on all features. This proves that our **MSTM** method selects more accurate features.

Number of Selected features. The number of selected features from different methods is shown at the bottom sub-table of Table III and Table IV. Our **MSTM** method can automatically select the optimal feature set by eliminating features with weight zero. **MSTM** selected 55 genes out of 12,023 features on *H3N3*, and 27 genes out of 12,023 features on *H1N1*. However, **FST**, **Simba, BAHSIC** and **Relief** cannot select the optimal feature set automatically, since they are all feature ranking methods. We report the number of top features where we get the highest accuracy for these three methods. The number of selected features is listed at the bottom of Table III and Table IV. Our method forces the weights of most irrelevant features to be zero, and it therefore selects much fewer features than the alternatives.

TABLE III. RESULTS ON H3N2 DATA

	All feature	BAHSIC	Relief	Simba	FST	MSTM
Sensitivity	0.667 ± 0	0.735 ± 0.202	0.875 ± 0.063	0.882 ± 0.073	1.000 ± 0	**1.000 ± 0**
Specificity	0.811 ± 0	0.582 ± 0.056	0.778 ± 0.118	0.763 ± 0.053	0.889 ± 0.130	**0.922 ± 0.150**
Balanced_Accuracy	0.771 ± 0	0.659 ± 0.129	0.826 ± 0.065	0.823 ± 0.063	0.944 ± 0.064	**0.961 ± 0.084**

(a) Classification Accuracy (mean ± std)

BAHSIC	Relief	Simba	FST	MSTM
217	154	135	50	55

(b) Number of Selected Features

TABLE IV. RESULTS ON H1N1 DATA

	All feature	BAHSIC	Relief	Simba	FST	MSTM
Sensitivity	0.938 ± 0	0.806 ± 0.052	1.000 ± 0	0.948 ± 0.003	1.000 ± 0	**1.000 ± 0**
Specificity	0.125 ± 0	0.405 ± 0.131	0.500 ± 0.132	0.605 ± 0.163	0.750 ± 0.151	**0.801 ± 0.131**
Balanced_Accuracy	0.531 ± 0	0.606 ± 0.092	0.750 ± 0.074	0.777 ± 0.085	0.875 ± 0.101	**0.901 ± 0.065**

(a) Classification Accuracy (mean ± std)

BAHSIC	Relief	Simba	FST	MSTM
346	121	141	43	27

(b) Number of Selected Features

V. CONCLUSION

We proposed a margin-based feature selection filter that can directly select a few informative genes from temporal high-dimensional gene expressions. For each subject, we define a temporal margin based on a measure of distance between two multivariate time series from other subjects. We take into account the uncertainty in calculating nearest neighbors by considering the updated weights in each iteration. The objective function of the proposed selection method is to maximize each subject's temporal margin in its own relevant subspace. The optimal weight for each feature is learned by solving this optimization problem.

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. 2003. "Feature Selection for High-dimensional Data, A Fast Correlation-based Filter Solution." In 20th *International Conference on Machine Learning*, pp. 856-863

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

[3] Lou, Q, and Obradovic, Z. "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. 2004. Margin Based Feature Selecgtion - theory and algorithms. In 21st *International Conference on Machine Learning*, pp. 43-50

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

[6] Margaritis, D. and Thrun, S. "Bayesian Network Induction via Local Neighborhoods", In *Neural Information Processing System*, 1999.

[7] Shen, J., Li, L. and Wong, W. "Markov Blanket Feature Selection for Support Vector Machine" In Proc. of *AAAI Conference on Artificial Intelligence (AAAI-08)*. 2008,

[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, Pittsb*

[12] Song, L., Smola, A., Gretton, A. Borgwardt, K.M., and Bedo, J. "Supervised Feature Selection via Dependence Estimation", International Conference on Machine Learning. 2007, pp. 856-863.

[13] Lou, Q and Obradovic, Z. "Prediciting Viral Infection by Selecting Informative Biomarkers From Temporal High-Dimensional Gene Expression Data" *IEEE International Confenece on Bioinformatics and Biomedicine*, October 2012, Philadelphia, USA