T-Relief: Feature Selection for Temporal High-Dimensional Gene Expression Data

Milos Radovic[1,2*], Milos Jordanski[3], Nenad Filipovic[4] and Zoran Obradovic[1]

[1]Center for Data Analytics and Biomedical Informatics, College of Science and Technology, Temple University, North 12th Street, 19122 Philadelphia, PA, USA,
{zoran.obradovic,milos.radovic}@temple.edu

[2]Bioengineering Research and Development Center - BioIRC,
Prvoslava Stojanovica 6, 34000 Kragujevac, Serbia
milos.radovic@temple.edu

[3]Faculty of Mathematics, University of Belgrade,
Studentski trg 16, 11000 Belgrade, Serbia
jordanski90@hotmail.com

[4]Faculty of Engineering, University of Kragujevac,
Sestre Janjic 6, 34000 Kragujevac, Serbia
fica@kg.ac.rs

Abstract. The high-dimensionality of microarray datasets with a small number of samples presents a challenge for the microarray data classification task, especially when features vary through time. Feature selection has been widely used in data mining and machine learning as a preprocessing step to reduce number of features and to enhance classification performance. Even though various feature selection methods for microarray data classification have been proposed, most are not able to cope with temporal data. We proposed the temporal Relief (T-Relief) algorithm, which follows the main idea of the Relief algorithm, but is able to handle temporal nature of data. T-Relief uses dynamical time warping to calculate distance between two features varying through time. The proposed method is evaluated on an H3N2 virus gene expression dataset and obtained results provide evidence that the T-Relief algorithm outperforms alternatives widely used in gene expression studies.

Keywords: Feature selection; Gene expression; Temporal data

1 Introduction

In recent years, with the rapid advances of science and technology, datasets with large numbers of features and relatively small numbers of samples were produced. For example, biology researchers are able to measure in a single experiment thousands of gene expressions simultaneously. However, some of these experiments are expensive and, consequently, small numbers of samples are usually available. A classification in

such a high-dimensional dataset with a small number of samples is a challenging task. High-dimensional nature of a dataset may lead to increased computational cost and decreased accuracy of machine learning models. However, usually among a large number of features just a small fraction is relevant for the particular classification task. For this purpose, application of methods that can reduce dimensionality of the dataset by extracting relevant features as a preprocessing step to classification can be of great importance.

Feature selection has been widely used as a preprocessing step in machine learning to reduce dimensionality of data. There are three main categories of feature selection methods: filters [1], wrappers [2] and embedded methods [3]. Filter based methods evaluate the quality of feature subsets by observing only the intrinsic data characteristics, i.e. statistical measures, independently of a prediction model. Wrapper based methods employ learning algorithm to determine the goodness of a selected feature subset and usually the accuracy of the learning algorithm is used to guide the search process. Even though these methods are superior in terms of predictive accuracy, they are not suitable for high dimensional datasets since they suffer from high computational cost. Finally, the embedded methods merge feature selection with a specific machine learning model and then an optimal subset of features is obtained by the learning algorithm.

Microarray data classification poses a serious task for machine learning researchers. A challenge in microarray data classification is the identification of discriminative genes for a specific problem, which can enhance the accuracy of the classification task. However, this task is a challenging one, especially when features are measured in time, i.e. the data records for each sample are multivariate time series. Even though various feature selection methods for microarray data classification have been proposed [4], most are not able to cope with temporal data without data flattening, i.e. transforming a temporal data into a single matrix, which results in a loss of temporal information in gene expression data.

Recently, several feature selection approaches for temporal data have been proposed. For instance, in [5]-[6] authors proposed an approach where they project data to another space in order to learn new features. However, these methods represent dimensionality reduction rather than feature selection. In [7], authors proposed a margin-based feature selection where the objective is to maximize each subject's temporal margin in its own relevant subspace. However, the measure of distance between two multivariate time series defined in [7] is not able to capture similarities between two sequences. Multi-task Lasso method [8] is the embedded feature selection method, which employs the group lasso regularization using the $L_{2,1}$ - norm penalty. In this way, all classification models at different time point share a common set of features.

Relief is a well-known feature selection algorithm, widely used for microarray data, developed by Kira and Rendel [9] and further improved by Kononenko et al. [10], which estimates the quality of attributes, but cannot deal with temporal data. In this paper, we propose a Relief inspired feature selection approach, temporal Relief (T-Relief), which is able to handle multivariate temporal data without flattening. We preserved the original idea of Relief algorithm, but we employ dynamical time warping (DTW) as a distance measure between gene expression time series in order to find the nearest neighbors.

2 Methods

Relief is a feature selection algorithm proposed for binary classification [9], which seems to be simple, efficient and very powerful in estimating the quality of attributes. An extension of original Relief (ReliefF) is proposed, which can deal with noisy, incomplete, and multi-class datasets [10].

Let us denote by $D = \{(x_n, y_n)\}_{n=1}^{N} \in \mathbb{R}^{p \times t} \times \{-1,1\}$ a training set. The main idea of Relief is to iteratively estimate the feature weights based on their power to discriminate between neighboring patterns. In each iteration of the Relief algorithm, an instance R is randomly chosen and its two nearest neighbors are selected: one from the same class (H) and one from the opposite class (M). The weight of *j*th feature is updated as: $W_j = W_j - diff(R_j, H_j) + diff(R_j, M_j)$, where $diff(a,b) = |a-b|/r$ and r is range for that particular feature. In the updated version ReliefF [10], the K nearest neighbors from each class are selected and the average distance is used for updating feature weights.

As might be noticed the Relief is not able to deal with features that have dynamical characteristics. One way to overcome this problem is data flattening, i.e. transforming a temporal data into a single matrix, which results in loss of temporal information. In order to overcome this problem, we proposed a new feature selection algorithm T-Relief for extracting informative attributes from datasets where features vary through time. T-Relief preserves the original idea of Relief algorithm, but it employs DTW as a distance measure between time series data.

DTW is a well-known algorithm for measuring similarity between time series sequences and has been used in various patter recognition applications, such as handwriting recognition [11], signature recognition [12] and elsewhere. Given two time series sequences $a = a_1 a_2 ... a_r$ and $b = b_1 b_2 ... b_s$, DTW finds the optimal path between a and b using dynamical programming to calculate the minimal cumulative distance $dtw(a,b) = c(r,s)$, where $c(i,j)$ is recursively defined as:

$$c(i,j) = d(a_i, b_j) + \min(c(i-1, j-1), c(i-1, j), c(i, j-1)). \quad (1)$$

A data point in a time series sequence is usually a numerical value and $d(a_i, b_j)$ can be calculated as $(a_i - b_j)^2$. Let $x_i, x_j \in \mathbb{R}^{p \times t}$ be two samples from D, where p is the dimensionality of the feature space and each feature is measured in t time steps. The distance between samples x_i and x_j can be calculated as:

$$dist(x_i, x_j) = \sum_{l=1}^{p} dtw(x_{il}, x_{jl}) \quad (2)$$

Advantage of the proposed distance is that it is able to deal with missing values, since DTW is able to find the distance between two sequences of different size. In addition, DTW uses "elastic" alignment and is able to capture similarity between signals even if they are out of phase in time (in such cases Euclidean distance measure, which align corresponding time points, would fail to detect similarity). The T-Relief method follows the main idea of the Relief algorithm, but uses DTW distance to calculate the nearest neighbors. The T-Relief algorithm is given in Fig. 1.

Algorithm: T-Relief
Input: dataset $D = \{(x_n, y_n)\}_{n=1}^{N} \in \mathbb{R}^{p \times t} \times \{-1, 1\}$
Output: the vector W of estimations of the qualities of attributes
1. $W = (0, 0, ..., 0)$;
2. for i = 1 to m
3. Pick at random an instance x_i from D ;
4. Find K nearest neighbors H_k from the same class using distance measure given in (2);
5. Find K nearest neighbors M_k from the opposite class using distance measure given in (2);
6. for j = 1 to p
7. $W_j = W_j - \sum_{k=1}^{K} dtw(x_{ij}, H_{kj}) + \sum_{k=1}^{K} dtw(x_{ij}, M_{kj})$;
8. end
9. end

Fig. 1. Pseudo code of the T-Relief algorithm.

3 Results and discussion

3.1 Dataset description

In this study, we evaluated the T-Relief feature selected approach by comparing it with alternatives on the Influenza A virus gene expression dataset from human viral challenge study [13]. This dataset contains gene expression data for 17 human volunteers infected with H3N2 virus and then labeled based on severity of reaction to infection as "symptomatic" (9 subjects) or "asymptomatic" (8 subjects). In particular, symptoms were recorded twice daily and quantified using the modified Jackson score [14]. Thereafter, all patients with a modified Jackson score larger than or equal to 6 over the quarantine period were labeled as "symptomatic", while the other were labeled as "asymptomatic". Gene expression values for 12023 genes are available at baseline (24h prior to inoculation), and then at 15 more time points after the virus was injected (at 8 hr intervals). To summarize, the H3N2 dataset is balanced and it contains gene expression values for 12023 genes for 17 subjects at 16 time points which makes it a good candidate for evaluation of the T-Relief approach.

3.2 Comparison methods

We compared the proposed T-Relief method with five feature selection approaches commonly used in gene expression studies:

(1) **mRMR**: This method ranks features according to the minimal-redundancy-maximal-relevance criterion [15], meaning that it tends to select features which are highly correlated with the class and uncorrelated between themselves.
(2) **ANOVA**: A method that selects features based on F-statistic values.

(3) **ReliefF**: An improved version of the Relief [9] algorithm, robust to incomplete data and generalized to work on multi-class problems [10]. According to this algorithm, a good feature should have similar values in nearest neighbors from the same class and different values in nearest neighbors from different classes.
(4) Multi-task Lasso (**MT-Lasso**): One of the state-of-the-art methods for feature selection from temporal multivariate data [8] which employs the $L_{2,1}$ regularization term and thus ensures that all regression models at different time points (tasks) share a common set of features.
(5) Feature Selection Temporal (**FST**): A feature selection approach for temporal multivariate data that transforms the original feature space into a weighted feature space where it performs optimization to maximize temporal margin [7].

Two of the baseline approaches are designed to work on temporal data (MT-Lasso and FST), whereas remaining three (mRMR, ANOVA and ReliefF) require data flattening prior to feature selection.

3.3 Performance evaluation procedure

In this study, we evaluated the feature selection approaches on the H3N2 dataset by calculating the classification accuracy (which may serve as an appropriate metric since the dataset is balanced) of three classifiers: K-nearest neighbors (KNN), Naive Bayes classifier (NB) and Random Forest (RF). We tested classifiers by using leave-one-out cross-validation (LOOCV) procedure where in each iteration, the left-out observation was used for testing purposes, while the remaining observations were used for feature selection followed by classifier training (training set). In each iteration of the LOOCV procedure we tuned parameters of the feature selection methods by applying nested 4-fold cross validation procedure on the training set. In this way we found optimal values of $K_1 \in \{1,2,3\}$ - parameter, which defines the number of nearest hits/misses in the ReliefF algorithm, $\lambda \in \{0.1,1,5\}$ - regularization parameter in the MT-Lasso, and $K_2 \in \{1,2,3\}$ - parameter, which defines the number of nearest hits/misses in the T-Relief algorithm. In addition, we also tuned parameters of the classifiers by the internal 3-fold nested cross validation procedure applied on the same data on which feature selection is performed. In this way, we tuned the number of nearest neighbors in the KNN algorithm $k \in \{1,3,5\}$ and the number of trees in the RF algorithm $N_{tree} \in \{10,50,100\}$. Here, it should be noted that the testing observation was never used neither for feature selection nor for classifiers training (including parameter tuning of the feature selection methods and classifiers).

3.4 Classification accuracy on gene expression data

The proposed T-Relief feature selection approach was compared with five baseline feature selection algorithms (mRMR, ANOVA, ReliefF, MT-Lasso and FST) according to the evaluation procedure described in the previous section. By using the LOOCV procedure, the accuracy of KNN, NB and RF classifiers was calculated for

the top $m \in \{1,10,20,30,40,50,100\}$ genes selected by different feature selection methods.

Table 1 summarizes the results for the H3N2 dataset. It shows that the proposed T-Relief approach outperformed alternatives in most cases. In particular, the T-Relief approach achieved the highest accuracy in 16 out 21 cases, while the other methods outperformed it in no more than 3 experiments. These results indicate that the proposed T-Relief method has selected the most discriminative features (genes).

Table 1. Evaluation of feature selection methods on H3N2 dataset using the top m genes. Values represent classification accuracy (bold represents the best accuracy).

Feature selection method	KNN							NB							RF						
	Number of features							Number of features							Number of features						
	1	10	20	30	40	50	100	1	10	20	30	40	50	100	1	10	20	30	40	50	100
mRMR	**0.88**	0.59	0.65	0.76	0.82	0.71	0.82	**0.88**	0.71	0.59	0.65	0.76	0.88	**1.00**	**0.76**	0.59	0.71	0.59	0.88	0.82	0.82
ANOVA	**0.88**	0.82	0.76	0.88	0.88	0.94	0.94	**0.88**	0.88	0.88	0.94	0.94	0.94	0.88	**0.76**	0.82	0.88	0.94	0.82	0.88	0.88
ReliefF	0.71	0.47	0.47	0.71	0.71	0.88	0.94	0.71	0.53	0.88	**1.00**	**1.00**	**1.00**	**1.00**	0.71	0.47	0.82	**1.00**	0.94	**1.00**	**1.00**
MT-Lasso	0.53	0.82	0.88	0.94	**1.00**	**1.00**	**1.00**	0.41	0.94	0.94	0.94	0.94	**1.00**	**1.00**	0.35	0.88	0.94	0.94	0.94	**1.00**	0.94
FST	0.29	0.76	0.88	0.94	0.88	0.82	0.88	0.18	0.76	0.88	0.88	0.88	0.94	**1.00**	0.24	0.82	0.88	**1.00**	0.94	0.94	0.88
T-Relief	0.82	**0.94**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.76	**0.94**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.65	**0.94**	**0.94**	**1.00**	**1.00**	0.94	0.94

The results are graphically depicted in the Fig. 2 where the accuracy of different classifiers is plotted as a function of m (number of selected genes). This figure shows that classifiers benefit from the features selected by the T-Relief approach. In addition, Fig. 2 shows that temporal methods (MT-Lasso, FST and T-Relief) lead to monotonic improvement in accuracy of classifiers with the increase of m which is a desirable property. This is not the case with other methods (mRMR, ANOVA and ReliefF) and thus, we can conclude that flattening of temporal data might lead to the selection of irrelevant genes which further cause unstable behavior of classifiers.

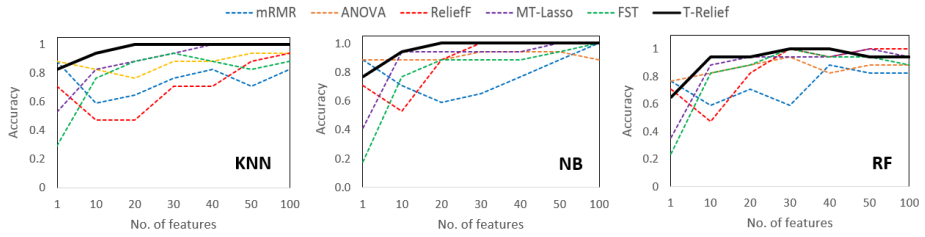

Fig. 2. Classification accuracy obtained by using LOOCV procedure on the H3N2 dataset. Results are given for the three classifiers: KNN (left), NB (middle) and RF (right).

3.5 Gene ontology enrichment analysis

By analyzing the list of genes selected by the T-Relief method we found out that 62 genes were selected among the top 100 genes in all folds of the LOOCV procedure. For this set of genes, we have performed enrichment analysis to find over-represented gene ontology (GO) terms. Annotations for the 62 genes were submitted to the

PANTHER classification system (http://www.pantherdb.org/), which extracted significantly over-represented biological processes. The top 20 GO terms are reported in Table 2, where the last column represents the p-values corrected based on the Bonferroni procedure. This table shows that GO terms related to the immune system and response to viruses are dominant among the top 20 enriched GO terms. This is consistent with the fact that the H3N2 dataset originates from human viral challenge study where human volunteers were infected with H3N2 influenza virus.

Table 2. Top 20 GO terms enriched in the 62 genes selected by the T-Relief algorithm.

GO biological process	GO ID	No.	Expected	Fold enrichment	P-value
defense response to virus	GO:0051607	25	0.46	54.19	4.04E-33
response to virus	GO:0009615	26	0.7	37.12	1.90E-30
innate immune response	GO:0045087	30	1.73	17.31	2.26E-26
type I interferon signaling pathway	GO:0060337	17	0.18	95.92	2.62E-25
cellular response to type I interferon	GO:0071357	17	0.18	95.92	2.62E-25
response to type I interferon	GO:0034340	17	0.19	90.19	7.41E-25
immune effector process	GO:0002252	26	1.34	19.38	2.92E-23
defense response	GO:0006952	34	3.45	9.84	7.07E-23
immune response	GO:0006955	32	3.1	10.32	1.09E-21
defense response to other organism	GO:0098542	25	1.38	18.06	2.04E-21
response to cytokine	GO:0034097	26	2.04	12.73	1.10E-18
response to other organism	GO:0051707	27	2.38	11.36	2.41E-18
response to external biotic stimulus	GO:0043207	27	2.38	11.36	2.41E-18
response to biotic stimulus	GO:0009607	27	2.46	10.96	6.08E-18
cytokine-mediated signaling pathway	GO:0019221	22	1.3	16.96	1.10E-17
negative regulation of viral process	GO:0048525	14	0.25	55.91	5.51E-17
regulation of multi-organism process	GO:0043900	21	1.3	16.12	3.11E-16
immune system process	GO:0002376	34	5.64	6.03	4.57E-16
regulation of viral process	GO:0050792	16	0.51	31.42	5.90E-16
negative regulation of viral life cycle	GO:1903901	13	0.24	53.73	2.54E-15

4 Conclusions

We proposed the filter-based feature selection method for temporal data, i.e. data in which features varies through time. The proposed method is mainly based on the Relief feature selection algorithm, but is adapted to deal with temporal data. In order to handle multivariate temporal data without data flattening, we modified the distance measure between samples. More specifically, the distance we proposed is based on dynamical time warping, which calculates the similarity between two features through time. The proposed method has been tested on H3N2 temporal gene expression dataset from viral study by three classification methods (KNN, NB and RF). We showed that in most cases the proposed T-Relief method outperforms alternatives. Our further research will be focused on testing the proposed method on other multivariate temporal datasets. In addition, other ideas for further research include using classifiers that are able to handle temporal data.

Acknowledgments. This material is based upon work partially supported by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under Contract No. W911NF-16-C-0050, and partially supported by DARPA grant No. 66001-11-1-4183 negotiated by SSC Pacific grant, and Serbian Ministry of Education, Science and Technological Development grants III41007 and ON174028.

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, pp. 856--863 (2003).
[2] Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics. **23**(19), 2507--2517 (2007).
[3] Maldonado, S., Weber, R., Basak, J.: Simultaneous feature selection and classification using kernel-penalized support vector machines. Information Science. **181**(1), 115--128 (2011).
[4] Canedo, V.B., Marono, N.S., Betanzos, A.A., Benítez, J.M., Herrera, F.: A review of microarray datasets and applied feature selection methods. Information Sciences. 282, 111--135 (2014).
[5] Chen, B., Chen, M., Paisley, J., Zass, A., Woods, C., Ginsburg, G.S., Lucas, J., Dunson, D., Carin, L.: Bayesian Inference of the Number of Factors in Gene-expression Analysis: Application to Human Virus Challenge Studies. BMC bioinformatics, 11(552), (2010).
[6] 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, 106(496), (2011).
[7] Lou, Q., Obradovic, Z.: Classifying Temporal Microarray Data by Selecting Informative Genes. Journal of Bioinformatics and Computational Biology. **11**(3), (2013)
[8] Argyriou, A., Evgeniou, T., and Pontil, M.: Multi-task feature learning. Advances in neural information processing systems. **19**, 41--48 (2007).
[9] Kira K., and Rendell L.: A Practical Approach to Feature Selection. Proc. Ninth Int'l Conf. Machine Learning. pp. 249--256 (1992).
[10] Kononenko, I., Simec, E., & Robnik-Sikonja, M.: Overcoming the myopia of inductive learning algorithms with RELIEFF. Applied Intelligence. **7**, 39--55 (1997).
[11] Bahlmann, C., Haasdonk, B., Burkhardt, H.: Online handwriting recognition with support vector machines - a kernel approach. Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 49--54 (2002).
[12] Faundez-Zanuy, M.: On-line signature recognition based on VQ-DTW. Pattern Recognition. 40(3), 981--992 (2007).
[13] Zaas, A.K. et al.: Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infection in Humans. Cell Host Microbe. **6**(3), 207--217 (2009).
[14] Jackson, G.G., Dowling, H.F., Spiesman I.G., and Boand, A.V.: Transmission of the common cold to volunteers under controlled conditions. I. The common cold as a clinical entity. AMA Arch Intern Med. 101, 267--278 (1958).
[15] Ding, C., and Peng, H.: Minimum redundancy feature selection from micro-array gene expression data. Journal of Bioinformatics and Computational Biology. **3**(2), 185--205 (2005).