Time Series Gene Expression Data Classification via L_1-norm Temporal SVM

Carlotta Orsenigo and Carlo Vercellis

Dept. of Management, Economics and Industrial Engineering,
Politecnico di Milano Via Lambruschini 4b, 20156 Milano, Italy
{carlotta.orsenigo,carlo.vercellis}@polimi.it

Abstract. Machine learning methods have been successfully applied to the phenotype classification of many diseases based on static gene expression measurements. More recently microarray data have been collected over time, making available datasets composed by time series of expression gene profiles. In this paper we propose a new method for time series classification, based on a temporal extension of L_1-norm support vector machines, that uses dynamic time warping distance for measuring time series similarity. This results in a mixed-integer optimization model which is solved by a sequential approximation algorithm. Computational tests performed on two benchmark datasets indicate the effectiveness of the proposed method compared to other techniques, and the general usefulness of the approaches based on dynamic time warping for labeling time series gene expression data.

Keywords: Time series classification, microarray data, L_1-norm support vector machines, dynamic time warping.

1 Introduction

In the last decade several machine learning methods have been proposed for the classification of gene expression data based on static datasets [1–4]. These datasets are usually composed by a huge number of features (genes) and a relatively few number of examples, and their values represent gene expression levels observed in a snapshot under precise experimental conditions.

The analysis of microarray expression levels recorded at a single time frame has proven to be effective for several biomedical tasks, among which the most prominent one is the phenotype classification in the early stages of a disease. However, it may appear inadequate to properly grasp the complex evolving interactions steering the biological processes. For example, in functional genomics studies the automatic categorization of genes based on their temporal evolution in the cell cycle plays a primary role, since genes with similar expression profiles are supposed to be functionally related or co-regulated [5]. As another example consider the prediction of the clinical response to a drug [6], where patients may exhibit different rates of disease development or treatment response. In this case, the overall profiles of the expression levels of two patients may be similar but not

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 264–274, 2010.

aligned, since individuals may progress at different speed [7]. In both scenarios it is required to analyze gene expression profiles as they evolve over time and, consequently, to develop classification methods able to consider also the temporal dimension. Over recent years a growing number of microarray experiments have been performed in order to collect and analyze time series gene expression profiles. The resulting datasets then provide examples of labeled time series that can be useful for classifying new temporal sequences whose label is unknown, and for identifying hidden explanatory biological patterns.

More generally, time series classification is a supervised learning problem aimed at labeling temporally structured univariate or multivariate sequences. Several alternative paradigms for time series classification have been proposed in the literature; see the review [8]. A common approach is based on a two-stage procedure that first derives a rectangular representation of the time series and then applies a classification method for labeling the data. An alternative approach relies on the notion of dynamic time warping (DTW) distance, an effective measure of similarity between pairs of time series. This distance allows to detect clusters and to predict with high accuracy the class of new temporal sequences by using distance-based methods, such as the k-nearest neighbor classifier [9, 10]. Furthermore, kernels based on DTW have been devised and incorporated within traditional support vector machines in [11–13].

In this paper we propose a new classification method based on a temporal variant of L_1-norm support vector machines (SVM), denoted as L_1-TSVM. The resulting mixed-integer optimization model, solved by a sequential approximation algorithm, takes into account the similarity among time series assigned to the same class, by including into the objective function a term that depends on the warping distances. A first research contribution along these lines is presented in [14], in which authors propose a temporal extension of discrete SVM, a variant of SVM based on the idea of accurately evaluating the number of misclassified examples instead of measuring their distance from the separating hyperplane [15, 16]. In this paper L_1-norm SVM [17–19] have been preferred as the base classifier for incorporating the temporal extension since they are efficient and well suited to deal with datasets with a high number of attributes, particularly in presence of redundant noisy features.

A second aim of the paper is to investigate whether DTW distance can be generally beneficial to different classifiers for labeling time series gene expression data. To this purpose, we comparatively evaluated the performances of five alternative methods beside L_1-TSVM: these are L_1-norm SVM, L_2-norm SVM with radial basis function and DTW as kernels, and the k-nearest neighbor (k-NN) classifier either based on Euclidean or DTW distances. Computational tests performed on two datasets seem to indicate that the proposed method L_1-TSVM has a great potential to perform an accurate classification of time series gene expression profiles and that, in general, SVM techniques based upon DTW perform rather well with respect to their non-DTW-based counterparts.

The paper is organized as follows. Section 2 defines time series classification problems and the concept of warping distance. In section 3 a new classification

model based on L_1-norm temporal SVM is presented. In section 4 computational experiences are illustrated. Finally, section 5 discusses some future extensions.

2 Time Series Classification and Warping Distance

In a time series classification problem we are given a set of multivariate time series $\{\mathbf{A}_i\}$, $i \in \mathcal{M} = \{1, 2, \ldots, m\}$, where each $\mathbf{A}_i = [a_{ilt}]$ is a rectangular matrix of size $L \times T_i$ of real numbers. Here $l \in \mathcal{L} = \{1, 2, \ldots, L\}$ is the index associated to the *attributes* of the time series, whereas $t \in \mathcal{T}_i = \{1, 2, \ldots, T_i\}$ is the temporal index, that may vary in a different range for each $\mathbf{A}_i$. Every time series is also associated with a *class label* $y_i \in \mathcal{D}$. Let $\mathcal{H}$ denote a set of functions $f : \Re^n \mapsto \mathcal{D}$ that represent hypothetical relationships between $\{\mathbf{A}_i\}$ and y_i. The *time series classification problem* consists of defining an appropriate hypotheses space $\mathcal{H}$ and a function $f^* \in \mathcal{H}$ which optimally describes the relationship between the time series $\{\mathbf{A}_i\}$ and their labels $\{y_i\}$, in the sense of minimizing some measure of misclassification. When there are only two classes, i.e. $D = 2$ and $y_i \in \{-1, 1\}$ without loss of generality, we obtain a *binary* classification problem, while the general case is termed *multicategory* classification.

The *warping distance* has proven to be an effective proximity measure for clustering and labeling univariate time series [9, 10]. Indeed, it appears more robust than the Euclidean metric as a similarity measure, since it can handle sequences of variable length and automatically align the time series to identify similar profiles with different phases.

In order to find the optimal alignment between two time series $\mathbf{A}_i$ and $\mathbf{A}_k$, let $G = (V, E)$ be a directed graph whose vertices in V correspond to the pair of time periods $(r, s), r \in \mathcal{T}_i, s \in \mathcal{T}_k$. A vertex $v = (r, s)$ indicates that the r-th value of the time series $\mathbf{A}_i$ is matched with the s-th value of $\mathbf{A}_k$. An oriented arc (u, v) connects vertex $u = (p, q)$ to vertex $v = (r, s)$ if and only if one of the following mutually exclusive conditions holds

$$\{r = p + 1, s = q\} \vee \{r = p + 1, s = q + 1\} \vee \{r = p, s = q + 1\}. \tag{1}$$

Consequently, each vertex $u \in G$ has at most three outgoing arcs, associated to the three conditions described in (1). The arc (u, v) connecting the vertices $u = (p, q)$ and $v = (r, s)$ has length

$$\gamma_{uv} = \sum_{l=1}^{L} (a_{ilr} - a_{kls})^2, \tag{2}$$

given by the sum over the attributes of the squared distances associated to the potential alignment of period r in $\mathbf{A}_i$ to period s in $\mathbf{A}_k$. Let also $v_f = (1, 1)$ and $v_l = (T_i, T_k)$ be the vertices corresponding to the alignment of the first and last periods in the two sequences, respectively.

A *warping path* in G is any path connecting the source vertex v_f to the destination vertex v_l. It identifies a phasing and alignment between two time series such that matched time periods are monotonically spaced in time and

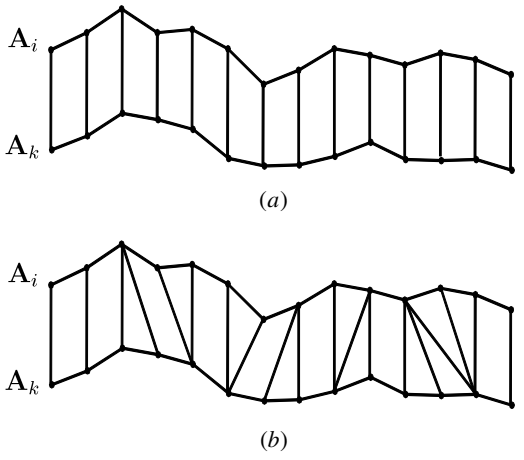

Fig. 1. Alignment of $\mathbf{A}_i$ and $\mathbf{A}_k$with Euclidean distance (a) and DTW distance (b)

contiguous. The *warping distance* between time series $\mathbf{A}_i$ and $\mathbf{A}_k$ is then defined as the length of the shortest warping path in G, and provides a measure of similarity between two temporal sequences which is often more effective than the Euclidean metric, as shown in Figure 1.

The warping distance between $\mathbf{A}_i$ and $\mathbf{A}_k$ can be evaluated by a dynamic optimization algorithm, with time complexity $O(T_{max}^2)$ ($T_{max} = \max\{T_i : i \in \mathcal{M}\}$), based on the following recursive equation

$$g(r,s) = \gamma_{uv} + \min\{g(r-1,s-1), g(r-1,s), g(r,s-1)\}, \tag{3}$$

where $g(r,s)$ denotes the cumulative distance of a warping path aligning the time series through the periods going from the pair $(1,1)$ to the pair (r,s).

3 L_1-norm Temporal Support Vector Machines

In this section we propose a new classification method based on a temporal variant of L_1-norm SVM, denoted as L_1-TSVM. The resulting mixed-integer optimization model, solved by a sequential approximation algorithm, takes into account the similarity among time series assigned to the same class, by including into the objective function a term that depends on the warping distances. We confine our attention to binary classification, since multicategory classification problems can be reduced to sequences of binary problems by means of *one-against-all* or *all-against-all* schemes [16, 20]. By applying an appropriate rectangularization preprocessing step, as described in section 4 for the time series considered in our tests, we may assume that the input dataset is represented by a $m \times n$ matrix, in which each row is a vector of real numbers $\mathbf{x}_i \in \Re^n$ which represents the corresponding time series $\mathbf{A}_i$.

A linear hypothesis for binary classification corresponds to a space $\mathcal{H}$ composed by separating hyperplanes taking the form $f(\mathbf{x}) = \text{sgn}(\mathbf{w}'\mathbf{x} - b)$. In order to choose the optimal parameters $\mathbf{w}$ and b, traditional SVM [21–23], hereafter denoted as L_2-norm SVM, resort to the solution of the quadratic minimization problem

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|_2 + C\sum_{i=1}^{m}\xi_i && (L_2\text{-SVM}) \\ \text{s.t.} \quad & y_i\left(\mathbf{w}'\mathbf{x}_i - b\right) \geq 1 - \xi_i \quad i \in \mathcal{M} && (4) \\ & \xi_i \geq 0 \ \forall i; \quad \mathbf{w}, b \text{ free.} \end{aligned}$$

Here the L_2-norm $\|\mathbf{w}\|_2$ is a regularization term, aimed at maximizing the margin of separation, whereas the second term in the objective function is a loss function expressing the distance of the misclassified examples from the canonical hyperplane delimiting the correct halfspace. The parameter C is available for adjusting the trade-off between the two terms in the objective function of problem L_2-SVM.

The quadratic formulation L_2-SVM has some advantages, which contributed to its popularity. Among others, it admits fast solution algorithms and, through its dual problem, it allows to implicitly apply kernel transformations for deriving nonlinear separations in the original input space from linear separations obtained in a high-dimensional Hilbert space.

Yet, other norms $\|\mathbf{w}\|_p$ have been considered in the literature as alternative ways for expressing the margin maximization. In particular, linear formulations have attracted much attention [17–19] since they can benefit from the high efficiency of the solution algorithms for linear optimization problems. The linear counterpart of problem L_2-SVM is given by the optimization model

$$\begin{aligned} \min \quad & \|\mathbf{w}\|_1 + C\sum_{i=1}^{m}\xi_i && (L_1\text{-SVM}) \\ \text{s.t.} \quad & y_i\left(\mathbf{w}'\mathbf{x}_i - b\right) \geq 1 - \xi_i \quad i \in \mathcal{M} && (5) \\ & \xi_i \geq 0 \ \forall i; \quad \mathbf{w}, b \text{ free.} \end{aligned}$$

Although not suited to host the kernel transformations applicable to L_2-SVM, the linear problem L_1-SVM has proven even more effective to achieve an accurate separation directly into the input space, particularly when the number of attributes is high and there are noisy unnecessary features.

We propose an extension of problem L_1-SVM by defining a new term aimed at improving the discrimination capability when dealing with time series classification problems. This additional term is given by the sum of the warping distances between all pairs of time series assigned to the same class. By including this term into the objective function we aim at deriving a separating hyperplane which maximizes the overall similarity among time series lying in the same halfspace.

Let d_{ik} denote the warping distance between the pair of time series $(\mathbf{A}_i, \mathbf{A}_k)$. We have to introduce binary variables expressing the number of misclassified examples as

$$p_i = \begin{cases} 0 & \text{if } \mathbf{w}'\mathbf{x}_i - b \geq 1 \\ 1 & \text{otherwise} \end{cases} . \tag{6}$$

In order to determine the best separating hyperplane for time series classification, the following mixed-integer optimization problem L_1-TSVM, termed *L_1-norm temporal support vector machines*, can be formulated

$$\min \quad \sum_{j=1}^{n} u_j + C \sum_{i=1}^{m} \xi_i + \delta \sum_{i=1}^{m} \sum_{k=i+1}^{m} d_{ik} r_{ik} \qquad (L_1\text{-TSVM})$$

$$\text{s.t.} \quad y_i \left(\mathbf{w}'\mathbf{x}_i - b\right) \geq 1 - \xi_i \quad i \in \mathcal{M} \tag{7}$$

$$-u_j \leq w_j \leq u_j \quad j \in \mathcal{N} \tag{8}$$

$$\frac{1}{S}\xi_i \leq p_i \leq S\xi_i \quad i \in \mathcal{M} \tag{9}$$

$$-r_{ik} \leq y_i \left(2p_i - 1\right) + y_k \left(2p_k - 1\right) \leq r_{ik} \quad i, k \in \mathcal{M}, i < k \tag{10}$$

$$u_j, \xi_i, r_{ik} \geq 0 \; \forall i, j, k; \quad p_i \in \{0, 1\} \; \forall i; \quad \mathbf{w}, b \text{ free.}$$

Here S is a sufficiently large constant; C and δ the parameters to control the trade-off among the objective function terms. The family of continuous bounding variables $u_j, j \in \mathcal{N}$, and the constraints (8) are introduced in order to linearize the first term $\|\mathbf{w}\|_1$ in the objective function of problem L_1-SVM. Constraints (9) are required to enforce the relationship between the slack variables ξ_i and the binary misclassification variables p_i. Finally, the family of continuous bounding variables $r_{ik}, i, k \in \mathcal{M}$, together with the constraints (10), are needed to express in linear form via the third term the inclusion of the sum of the warping distances between the time series, as shown in [14].

For determining a feasible suboptimal solution to model L_1-TSVM, we propose the following approximation procedure based on a sequence of linear optimization (LO) problems. In what follows R-TSVM denotes the LO relaxation of model L_1-TSVM, and t is the iteration counter.

Procedure L_1-TSVM$_{\text{SLO}}$

1. Set $t = 0$ and consider the relaxation R-TSVM$_0$ of L_1-TSVM.
2. Solve problem R-TSVM$_t$.
3. Suppose first that problem R-TSVM$_t$ is feasible. If its optimal solution is integer, the procedure is stopped and the solution generated at iteration t is retained as an approximation to the optimal solution of L_1-TSVM; otherwise, proceed to step 5.
4. Otherwise, if problem R-TSVM$_t$ is unfeasible, modify previous problem R-TSVM$_{t-1}$ by fixing to 1 all of its fractional variables. Problem R-TSVM$_t$ redefined in this way is necessarily feasible and any of its optimal solutions is integer. Thus, the procedure is stopped and the solution found is retained as an approximation to the optimal solution of L_1-TSVM.

5. Next problem R-TSVM$_{t+1}$ in the sequence is obtained by fixing to zero the relaxed binary variable with the smallest fractional value in the optimal solution of the predecessor R-TSVM$_t$; then proceed to step 2.

4 Computational Experiments

Computational experiments were performed on two datasets both composed by microarray time series gene expression data. As stated in the introduction our aim was twofold; from one side, we intended to evaluate the effectiveness of L_1-TSVM and to compare it to its continuous counterpart in terms of accuracy. From the other side, we were interested in investigating whether DTW distance may be conveniently used in conjunction with alternative supervised learning methods for gene expression time series classification.

The first dataset considered in our tests, denoted as *Yeast*[1] and originally described in [24], contains the genome characterization of the mRNA transcript levels during the cell cycle of the yeast *Saccharomyces cerevisiae*. Gene expression levels were gathered at regular intervals during the cell cycle. In particular, measurements were performed at 17 time points with an interval of ten minutes between each pair of recorded values. The gene expression time series of this dataset are known to be associated to five different phases, namely Early G1, Late G1, S, G2 and M, which represent the class values in our setting. The second dataset, indicated as *MS-rIFNβ* and first analyzed in [6], contains gene expression profiles of patients suffering from relapsing-remitting multiple sclerosis (MS), who are classified as either good or poor responders to recombinant human interferon beta (rIFNβ). The dataset is composed by the expression profiles of 70 genes isolated from each patient at 7 time points: before the administration of the first dose of the drug ($t = 0$), every three months ($t = 1, 2, 3, 4$) and every six months ($t = 5, 6$) in the first and second year of the therapy, respectively. For a few patients entire profile measurements are missing at one or two time points. From the complete *MS-rIFNβ* dataset we retained only twelve genes whose expression profiles at $t = 0$ have shown to accurately predict the response to rIFNβ, as described in [6]. Furthermore, for each possible number of time points from 2 to 7 we extracted the corresponding gene expression time series, in order to obtain six different datasets. The distinctive features of *Yeast* and *MS-rIFNβ* in terms of number of available examples, classes and time series length are summarized in Table 1.

Five alternative methods were considered for comparison with L_1-TSVM: L_1-SVM, SVM with radial basis function (SVM$_{\mathrm{RBF}}$) and dynamic time warping (SVM$_{\mathrm{DTW}}$) kernels, k-nearest neighbor classifier based respectively on Euclidean distance (k-NN$_{\mathrm{Eucl}}$) and dynamic warping distance (k-NN$_{\mathrm{DTW}}$). For solving L_1-TSVM and L_1-SVM models we respectively employed the heuristic procedure described in section 4 and standard LO code, both framed within the CPLEX environment; for nonlinear kernels SVM we used the LIBSVM library [25], extending its standard version with the DTW kernel. Among dynamic time

[1] `http://genomics.stanford.edu/yeast_cell_cycle/cellcycle.html`

Table 1. Summary of gene expression time series datasets

	Dataset	
Summary	*Yeast*	*MS-rIFNβ*
Examples	388	52
Classes	Early G1 (67), Late G1 (136), S (77) G2 (54), M (54)	Good responder (33), Poor responder (19)
Time series length	17	[5,7]

warping kernels we implemented the one proposed in [13], which has been proven to be positive definite under favorable conditions. Finally, in order to perform the classification of *Yeast*, which represents a multicategory dataset, SVM-based methods were framed within the all-against-all scheme.

A preprocessing step was applied on both datasets before classification. In particular, each expression profile of *Yeast* was normalized as described in [24]. The expression levels of *MS-rIFNβ* were instead standardized, by subtracting from each value in a gene profile the mean of the values of the same gene in temporal-homologous sequences, and dividing the result by the corresponding standard deviation. Since all methods apart from SVM_{DTW} and $k\text{-NN}_{\text{DTW}}$ are not able to cope with sequences of variable length, we replaced missing profiles with series of an out-of-range value, and then sequenced genes and time periods for every patient in order to obtain a rectangular representation for each of the six *MS-rIFNβ* datasets.

The accuracy of the competing methods was evaluated by applying five times 4-fold cross-validation, each time randomly dividing the dataset into four folds for training and testing. To achieve a fair comparison we used the same folds for all methods. Furthermore, on each training set we applied 3-fold cross-validation in order to figure out the optimal parameters setting for all classifiers, represented by the regularization constant C and the kernel parameter σ for L_1-norm and L_2-norm SVM methods, and by the number k of neighbors for k-NN classifiers. For L_1-TSVM a further parameter to be optimized was represented by the weight δ in the objective function, regulating the trade-off between misclassification and the sum of time series warping distances. The values tested for each parameter are reported in Table 2.

The results of each method are shown in Table 3 which indicates the average accuracy values obtained by applying five times 4-fold cross-validation. These results allow us to draw some empirical conclusions concerning the effectiveness of the proposed method and the usefulness of DTW distance. The temporal variant L_1-TSVM was capable of outperforming its counterpart L_1-SVM on all datasets, achieving an increase in accuracy ranging between 0.8% and 5.4%. The novel technique appeared rather accurate also with respect to the other classifiers, being able to provide the highest rate of correct predictions on *Yeast* and on most *MS-rIFNβ* datasets. Especially on these datasets, in which

Table 2. Parameters values tested for each family of methods

Method	Parameters values
k-NN$_{\text{Eucl}}$	$\text{k} = 2, 4, 6, 8, 10$
k-NN$_{\text{DTW}}$	
SVM$_{\text{RBF}}$	$\text{C} = 10^j, j \in [-1, 3]$
SVM$_{\text{DTW}}$	$\sigma = 10^j, j \in [-4, 2]$
L_1-SVM	$\text{C} = 10^j, j \in [-1, 3]$
L_1-TSVM	$\delta = 10^j, j \in [-1, 1]$

Table 3. Classification accuracy (%) on the gene expression time series datasets. Intervals in brackets indicate the time points considered for each *MS-rIFNβ* dataset.

	Method					
Dataset	k-NN$_{\text{Eucl}}$	k-NN$_{\text{DTW}}$	SVM$_{\text{RBF}}$	SVM$_{\text{DTW}}$	L_1-SVM	L_1-TSVM
Yeast	68.5	51.8	73.3	73.7	72.4	**73.9**
MS-rIFNβ						
t∈[0,1]	83.8	76.9	82.7	**84.2**	76.9	80.8
t∈[0,2]	81.9	78.9	82.7	84.6	80.0	**85.4**
t∈[0,3]	82.7	75.0	81.9	75.4	78.5	**83.8**
t∈[0,4]	76.9	73.1	76.9	71.2	79.2	**80.0**
t∈[0,5]	75.8	69.2	71.5	78.5	79.6	**80.8**
t∈[0,6]	71.2	66.9	68.5	70.8	76.5	**78.8**

examples are composed by sequences of variable length, also the use of DTW as the kernel function appeared promising. Notice that the average accuracy provided by most classifiers on *MS-rIFNβ* datasets decreased when more than four expression time series were considered for each example. This phenomenon is possibly related to the increase of missing profiles in the last measurements which may have slightly compromised the classification results. Nevertheless, L_1-TSVM showed the mildest degradation of its classification performances with respect to most of the other methods. On the $Yeast$ dataset the competing techniques L_1-TSVM and SVM$_{\text{DTW}}$ provided comparable results. Even in this case, however, L_1-TSVM was able to obtain the best rate of correct predictions. By investigating the confusion matrices of all methods we observed that the higher accuracy of L_1-TSVM mainly derived from the correct classification of a greater number of examples belonging to the classes S and M.

5 Conclusions and Future Extensions

In this paper we have proposed a new supervised learning method for time series gene expression classification based on a temporal extension of L_1-norm SVM.

The novel technique relies on a mixed-integer optimization problem which incorporates in the objective function an additional term aiming at improving the discrimination capability when dealing with the classification of time series datasets. This term is represented by the sum of the warping distances of time series assigned to the same class, where the warping distance is used as a similarity measure among temporal sequences. The inclusion of this term in the objective function is aimed at deriving separating hyperplanes which are also optimal with respect to time series similarity. In this paper we have also investigated from a computational perspective the convenience of combining the warping distance with alternative classification methods for time series gene profiles labeling. Experiments performed on two datasets showed the effectiveness of the proposed method and the usefulness of the warping distance when used as the kernel function in L_2-norm SVM. Future research development will be pursued along three main directions, by testing the novel technique on a wider range of time series gene expression classification problems, by investigating other similarity measures to be included in the model and by studying alternative heuristic procedures for solving the resulting mixed-integer formulations.

References

1. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
2. Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16, 906–914 (2000)
3. Lai, C., Reinders, M.J.T., van't Veer, L.J., Wessels, L.F.A.: A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics 7, 235 (2006)
4. Cho, S.B., Won, H.H.: Cancer classification using ensemble of neural networks with multiple significant gene subsets. Applied Intelligence 26, 243–250 (2007)
5. Peddada, S., Lobenhofer, E., Li, L., Afshari, C., Weinberg, C., Umbach, D.: Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics 19, 834–841 (2003)
6. Baranzini, S., Mousavi, P., Rio, J., Caillier, S., Stillman, A., Villoslada, P., Wyatt, M., Comabella, M., Greller, L., Somogyi, R., Montalban, X., Oksenberg, J.: Transcription-based prediction of response to IFNβ using supervised computational methods. PLoS Biology 3, 166–176 (2005)
7. Lin, T., Kaminski, N., Bar-Joseph, Z.: Alignment and classification of time series gene expression in clinical studies. In: ISMB (Supplement of Bioinformatics), pp. 147–155 (2008)
8. Kadous, M.W., Sammut, C.: Classification of multivariate time series and structured data using constructive induction. Machine Learning 58, 179–216 (2005)
9. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowledge and Information Systems 7, 358–386 (2004)
10. Xi, X., Keogh, E., Shelton, C., Wei, L.: Fast time series classification using numerosity reduction. In: Proc. of the 23rd International Conference on Machine Learning, pp. 1033–1040 (2006)

11. Shimodaira, H., Noma, K.I., Nakai, M., Sagayama, S.: Dynamic time-alignment kernel in support vector machine. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) NIPS, pp. 921–928. MIT Press, Cambridge (2001)
12. Bahlmann, C., Haasdonk, B., Burkhardt, H.: On-line handwriting recognition with support vector machines: A kernel approach. In: IWFHR '02: Proc. of the Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 49–54. IEEE Computer Society, Washington (2002)
13. Cuturi, M., Vert, J.P., Birkenes, O., Matsui, T.: A kernel for time series based on global alignments. In: Proc. of ICASSP, pp. 413–416 (2007)
14. Orsenigo, C., Vercellis, C.: Combining discrete SVM and fixed cardinality warping distances for multivariate time series classification. Pattern Recognition 43, 3787–3794 (2010)
15. Orsenigo, C., Vercellis, C.: Discrete support vector decision trees via tabu-search. Journal of Computational Statistics and Data Analysis 47, 311–322 (2004)
16. Orsenigo, C., Vercellis, C.: Multicategory classification via discrete support vector machines. Computational Management Science 6, 101–114 (2009)
17. Bradley, P.S., Mangasarian, O.L.: Massive data discrimination via linear support vector machines. Optimization Methods and Software 13, 1–10 (2000)
18. Zhu, J., Rosset, S., Hastie, T., Tibshirani, R.: 1-norm support vector machines. Neural Information Processing Systems 16 (2003)
19. Mangasarian, O.L.: Exact 1-norm support vector machines via unconstrained convex differentiable minimization. Journal of Machine Learning Research 7, 1517–1530 (2006)
20. Allwein, E., Schapire, R., Singer, Y.: Reducing multiclass to binary: a unifying approach for margin classifiers. Journal of Machine Learning Research 1, 113–141 (2000)
21. Vapnik, V.: The nature of statistical learning theory. Springer, New York (1995)
22. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
23. Schölkopf, B., Smola, A.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge (2002)
24. Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W.: A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 2, 65–73 (1998)
25. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`