Statistical Signal Processing for Novelty Detection.

Radu Balan, Justinian Rosca

Siemens Corporate Research
Princeton, SUA

Paul Bogdan

Politehnica University of Bucharest
Bucharest, Romania

Abstract

The goal of this article is to investigate and suggest techniques for health condition monitoring and diagnosis using machine learning from sensor data. In particular, this article overview and discusses support vector machines methods such as hard margin and soft margin problems. In order to investigate the abnormalities and classify a large set of data an iterative Support Vector Machine algorithm was constructed. However, similar techniques could be applied to analyze or monitor for abnormality various other complex devices or even computer methods.

Key words

Support Vector Machines, Health condition monitoring, Novelty detection and Machine learning methods.

1. Introduction

Support vector machines methods represent a powerful paradigm for classification and regression problems. For these reasons the support vector machines were successfully applied in domains such as data mining, fault and novelty detection problems, health condition monitoring for engines, bioinformatics (protein homology detection, functional interpretation of gene expression data), detection of anomalous windows registry (Krysta Svore, Katherine Heller, Angelos Keromytis, Salvatore Stolfo). The central problem in bioinformatics is predicting the functional and structural features of a protein based on its amino-acidic sequence. The task is to identify the protein homologies so that proteins can be clustered in families. From the statistical perspective, Jaakkola and Haussler construct a generative probabilistic model in order to analyze the protein sequence. For each protein sequence, a probability is assigned. In order to assess the similarity between protein sequences, they used the kernel functions. For extracting features from protein sequences, the model maps all protein sequences to points in a Euclidean feature space of fixed dimension. In order to classify the new points representing the protein sequences Jaakkola and Haussler used a discriminative statistical method(kernel methods). The SVM approach has been successfully applied also to the categorization of gene expression data from DNA microarrays.

2 Problem description

In order to assess a machine learning approach for faulty detection and health condition monitoring of a gas turbine, we consider a short description of the problem. For a gas turbine, we consider two types of measurements: the measurement of the external parameters $Z(t) = \{z_i(t) = z_i, i = 1, 2....n\}$ done by using a gas turbine sentry system, and the measurement of the internal parameters $X(t) = \{x_j(t) = x_j, j = 1, 2....m\}$ where $t \in \{0, T_0, 2T_0, 3T_0, ...\}$ by using a digital audio recorder system. With all these measurements, we want to build the following model state:

$$X(t) = F(Z(t), \sigma, \nu(t)) \text{ where} \tag{1}$$

$\nu(t)$ represents the statistical fluctuations or combustion dynamics, $X(t)$ the input space of the internal parameters, $Z(t)$ the input space of the external parameters and F the model state function. The measurement model that we have assumed is given by:

$$y(t) = x(t) + n(t) = F(z(t), \sigma, \nu(t) + n(t)) \tag{2}$$

where $y(t)$ and $n(t)$ represent the measurements affected by noise and respectively the noise that influences the measurement process. The goal of the model state function is to infer the state σ , which can be '1' - good state or '-1' -bad state, based on data measurements y and z.

3 Temporal Feature Extraction

In the present section, we analyze the methods that we use for processing the data set and how we construct the feature vectors for the training and testing steps of the classification problem. At the initial stage, we apply a filter bank

of k order to the internal measurements. In other words, for each signal vector y_i we apply a window function and transforms it into frequency domain by FFT. The first step of the method takes a part of the vector signal of dimension M (internal measurements), applies a window function (Hanning, Chebyshev, Hamming, etc) and transforms it into frequency domain by Fast Fourier Transform so that the transformed data is given by:

$$Y_{k,l}(\omega) = \sum_{j=1}^{M} e^{2\pi i \frac{\omega}{M}(j-1)} \cdot f(j) y_k(l-1) \cdot b + j)$$

$$\text{where } k = 1, 2, ...32,\ l = 1, 2, ..\text{NTF},\ \omega = 0, 1, ...\frac{M}{2}. \quad (3)$$

For the external measurements, we transform the data into frequency domain and consider the DC component only, because these measurements are slowly time varying signals:

$$z_k \mapsto Z_{k,l}(\omega)|_{\omega=0}$$
$$\text{where } k = 1, 2, ...25, \text{ and } l = 1, 2, ...\text{NTF} \quad (4)$$

With all this procesing we prepare two vector $Y_{k,l}$ and $Z_{k,l}$:

$$Y_l = \begin{pmatrix} Y_{k,l}(0) \\ Y_{k,l}(1) \\ .... \\ Y_{k,l}(\frac{M}{2}) \end{pmatrix} z_l = \begin{pmatrix} Z_{1,l}(0) \\ Z_{2,l}(1) \\ .... \\ z_{25,l}(\frac{M}{2}) \end{pmatrix}. \quad (5)$$

From the FFT instance, we transform the Y and Z parameters according a predefined set of frequecies $\Omega = \Omega_1, \Omega_2, ...\Omega_K$ and using filterbank principle obtaining the following vector Y_l:

$$Y_l = \begin{pmatrix} \sum_{\omega\in\Omega_1} |Y_{k,l}(\omega)| \\ \sum_{\omega\in\Omega_2} |Y_{k,l}(1)| \\ .... \\ \sum_{\omega\in\Omega_{NF=10}} |Y_{k,l}(\frac{M}{2})| \end{pmatrix}. \quad (6)$$

Due to the sumation $\sum_{\omega\in\Omega_i} |Y_{k,l}(\omega)|$, this step represents the nonlinear filtering of the measurements. At the end of this method, we compute a least square filter Φ which is the solution of the minimization problem:

$$\begin{aligned} \Phi &:= argmin_{\Phi\in Class} \sum_t \|u(t) - \Phi(z(t))\|^2 \\ &= argmin_{\Phi\in Class} \sum_{segment} \sum_{t\in segment} \|u(t) - \Phi(z(t))\|^2 \\ &= argmin_{\Phi_k\in Class} \sum_t \|U_{k,l}(t) - \Phi_k(z)\|^2 \\ &= argmin_{\Phi_k} \sum_{l=1}^{NTF} \|U_{k,l}(n) - \Phi_k^T Z_l\|^2. \end{aligned} \quad (7)$$

where $n = 1, 2, ...NF = 10$ and $k = 1, 2, ...32$. The feature vector at frame l after applying the linear predictor has the following form:

$$F_l = \begin{pmatrix} \sum_{\omega\in\Omega_1} |Y_{1,l}(\omega)| - \phi_1^T Z_l \\ \sum_{\omega\in\Omega_2} |Y_{1,l}(1)| - \phi_2^T Z_l \\ ... \\ \sum_{\omega\in\Omega_{NF}} |Y_{1,l}(1)| - \phi_{NF}^T Z_l \\ \sum_{\omega\in\Omega_1} |Y_{2,l}(1)| - \phi_{NF+1}^T Z_l \\ \sum_{\omega\in\Omega_2} |Y_{2,l}(1)| - \phi_{NF+2}^T Z_l \\ ... \\ \sum_{\omega\in\Omega_{NF}} |Y_{2,l}(1)| - \phi_{2NF}^T Z_l \\ \sum_{\omega\in\Omega_1} |Y_{32,l}(1)| - \phi_{31NF+1}^T Z_l \\ ... \\ \sum_{\omega\in\Omega_{NF=10}} |Y_{32,l}(\frac{M}{2})| - \phi_{32}^T Z_l \end{pmatrix}. \quad (8)$$

In the above mentioned derivations, NTF designates the total number of frames, NTS designates the number of total samples and NF designates the number of frequencies.

4 Support Vector Machines

In many classification problems, the task is reduced to the mere identification of object classes. Therefore, the article will approach the problem of One Class Classification and will propose an algorithm focused on large size data classification. For the beginning, we will define the context and the requirements of the One Class Classification problem. Hence, we will consider a set of measurements $X = \{x_j \in R^d, j = 1, 2....m\}$ entitled input space, to which we want to assign a label from output space $Y = \{y_i = \pm 1, i = 1, 2...m\}$. In other words, we want to design a decision or discriminant function which performs the classification: $f(x) = y$. If we suppose the data set is separable, the above mentioned requirement is equivalent with the construction of a hyperplane given by the equation:

$$\langle w, x\rangle + b = 0 \quad (9)$$

The vector w has the role of indicating the orientation of the hyperplane and the parameter b represents the offset respective to the origin that was considered. As there is no additional constraint concerning the data set there is an infinity of possibilities for defining the separation hyperplane:

$$H = \{\langle w, x\rangle + b = 0 |\ x \in X\} \quad (10)$$

The construction of the hyperplane gives us the opportunity of taking a Bayesian decision invariant to the positive scaling of the decision function argument :

$$f : X \to Y,\ f(x) = sgn\{\langle w, x\rangle + b\} \quad (11)$$

As the decision is invariant at the level of positive scaling we can define the canonical hyperplane which leads to the following conclusions:

$$\begin{aligned} \langle \overline{w}, x_1\rangle + b = 1, \text{ then} x_1 \in \text{ TargetClass} \\ \langle \overline{w}, x_2\rangle + b = -1, \text{ then} x_2 \in \text{ OutlierClass} \end{aligned} \quad (12)$$

The $\overline{w}$ vector of the canonical hyperplane is the result of the relation $w/\left\|w\right\|_2^2$ and the margin is given by $1/\left\|w\right\|_2^2$. We have observed judging by the above mentioned data that there is an infinity of possibilities for the construction of the separating hyperplane. Taking into account that we want the misclassification error to be minimized as much as possible, the hyperplane needs to be as far away as possible from both convex data sets. This is equivalent to the following optimization problem(Hard Margin Problem):

$$\begin{aligned} &minimize && \frac{1}{2}\langle w,w\rangle \\ &\text{subject to} && y_i(\langle w,x_i\rangle+b)\geq 1, i=1,2..m. \end{aligned} \quad (13)$$

The maximization problem of the margin is translated through the minimization of the quantity and leads to the support vector machines learning rule. As it is an optimization problem of a convex criterion we can resort to a Lagrangian treatment which leads to a dual approach of the primal formulation. The dual problem is obtained by constructing the Lagrangian function:

$$L(w,b,\alpha)=\frac{1}{2}\langle w,w\rangle-\sum_{i=1}^{m}\alpha_i[y_i((\langle w,x_i\rangle)+b)-1] \quad (14)$$

where α are called Lagrangian multipliers. Setting the Karush-Kuhn-Tucker conditions

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{m}y_i\alpha_i x_i=0 \quad (15)$$

and

$$\frac{\partial L}{\partial b}=\sum_{i=1}^{m}y_i\alpha_i \quad (16)$$

and re-substituting in the Lagrangian function the relation becomes :

$$L(w,b,\alpha)=\sum_{i=1}^{m}\alpha_i-(1/2)\sum_{i,j=1}^{m}y_iy_j\alpha_i\alpha_j\langle x_i,x_j\rangle. \quad (17)$$

Taking into consideration the result of the MinMax theorem we reach the following optimization problem:

$$\begin{aligned} &minimize && -\frac{1}{2}\sum_i\sum_j y_iy_j\alpha_i\alpha_j\langle x_i,x_j\rangle+\sum_i\alpha_i \\ &\text{subject to} && \sum_i y_i\alpha_i=0\ \alpha_i\geq 0. \end{aligned} \quad (18)$$

As it can be noticed in the expression of the optimization problem, the objects x_i and x_j data appear in the context of a scalar product. This aspect corroborated with the case in which the initial data set is not linearly separable leads us to the conclusion that the idea of projecting the data in a new space that allows for a better representation. Such a space is called feature space and the function that performs the projection of the data is called feature map. The Reproducing Kernel Hilbert space concept permits the definition of the kernel function on the basis of the relation between objects $\Phi(x_i)$ and $\Phi(x_j)$. Thus, given the adequate choice of the kernel function, the data can be classified linearly in feature space and this aspect provides benefits as the map function does not need to be explicitly known. With the help of the kernel function, the optimization problem can be re-written in the following way:

$$\begin{aligned} &maximize && \frac{1}{2}\sum_i\sum_j y_iy_j\alpha_i\alpha_jK(x_i,x_j)+\sum_i\alpha_i \\ &\text{subject to} && \sum_i y_i\alpha_i=0\ \alpha_i\geq 0. \end{aligned} \quad (19)$$

By analogy the decision function becomes:

$$f(x_j)=sgn\left\{\sum_{i=1}^{nSV}\alpha_iy_iK(x_i,x_j)+b\right\}. \quad (20)$$

The maximization problem of the margin hyperplane is equivalent to finding the closest points to the separating surface having the Lagrange multipliers $\alpha_i>0$ and being called support vectors. These points are interpretated as the most informative pattern vectors in the data set. If the data set is not linearly separable or equivalently the two convex hulls are intersecting we need to penalize the criterion by a quantity that represents these points of intersection. Vapnik introduces the slack variables for these trouble objects. These slack variables stand for an error resulting from the classification process. Adding this error to the objective criterion the problem becomes(Soft Margin Problem):

$$\begin{aligned} &minimize && \frac{1}{2}\langle w,w\rangle+C\sum_i\xi_i \\ &\text{subject to} && y_i(\langle w,x_i\rangle+b)\geq 1-\xi_i. \end{aligned} \quad (21)$$

or in terms of the kernel functions:

$$\begin{aligned} &minimize && \frac{1}{2}\sum_{i,j}^{m}\alpha_i\alpha_jy_iy_jK(x_i,x_j)-\sum_i^m\alpha_i \\ &\text{subject to} && 0\leq\alpha_i\leq C,\ \sum_i^m\alpha_iy_i=0. \end{aligned} \quad (22)$$

The difference between the initial form and the above mentioned expression is that the Lagrange multipliers are upper bounded by a parameter C ranging from 0 to 1.

4.1 Iterative SVM approach

The IterativeSVM algorithm is dedicated to processing large data sets and finding a fixed number of support vectors according to limited computational conditions. The

proposed SVM method requires two stages: one stage is dedicated to training in which we try to solve a quadratic optimization problem in order to find the desired support vectors, and the second stage that implies computing the decision function for each testing pattern. The algorithm can summarized as follows:

Algorithm IterativeSVM(Given a large training data set the algorithm finds the desired support vectors, the Lagrange multipliers corresponding to each support vector, the bias parameter rho used for classification of the testing data set)
1.Training Step
1.1.Initialization
1.1.1.m:number of maximum support vectors from each step
1.1.2.label=1 all training vectors are considered true positive points
1.2.For each bunch of data
1.2.1.Prepare the new problem
1.2.1.1.Old part of the size equal to the number of its support vectors
1.2.1.2.Load new bunch of data from the data set
1.2.2.Call the SVM code for finding the support vectors
1.2.3.Select m support vectors at most, based on the *alpha-rule*
2.Testing Step
2.1.Compute for each pattern the decision function and classify it

As it can be noticed, we have introduced in the above algorithm an *alpha rule* which permits us to select a desired number of support vectors. In other words we will select the most informative support vector based on the largeness of the Lagrange multipliers. For solving the quadratic optimization problem, we can use the conjugated gradient method, the quasi-Newton method or the interior points method.

5. Summary and Conclusions

The support vector machines methods present important features like the use of kernels in order to obtain a better representation of the data, the sparseness of the solution and the capacity control obtained by acting on the margin. Since the criterion that should be minimized is convex, the optimization solution is unique and we will not experience problems concerning the local minimum. We can easily observe that the dimensionality of feature space does not interfere with the solution of the minimization problem. The only problem that arises at this stage is represented by the kernel choice which should offer more information about the data. For the optimization task, there exist well established algorithms like gradient ascent or sequential minimal optimization. Also the support vector machines method turned out to be robust to the noise measurements. Even if the support vector machines approach is a powerful technique for classification, regression and novelty detection problems, many issues are still debateble. One of these issues is represented by the desired number of support vectors which should perform similar results on different datasets. Another issue is represented by the adequate choice of kernel function and well parametriztion of this function.

References

[1] Bernhard Scholkopf - *Kernel Learning Methods* Cambridge Press, 1999.

[2] Paul Hayton, Bernhard Scholkopf, Lionel Tarassenko, Paul Anuzis - *Support Vector Novelty Detection Applied to Jet Engine Vibration Spectra,*, Neural Information Processing Systems, 2000.

[3] Colin Cambell, Kristin P. Bennet - *A Linear Programming Approach to Novelty Detection*, Neural Information Processing Systems, 2000.

[4] Elzbieta Pekalska, David Tax, Robert Duin - *One class LP classifier for Dissimilarity representation*, Neural Information Processing Systems, 2002.

[5] David Tax, Robert Duin - *Uniform object generation for optimizing one class classifiers*, Journal of Machine Learing Research, 2001

[6] David Crisp, Christopher Burges -*A geometric interpretation of nu - SVM classifiers*, 2000