Bayesian Approach to the Pattern Recognition Problem in Nonstationary Environment

O.V. Krasotkina[1], V.V. Mottl[2], and P.A. Turkov[1]

[1] Tula State University, Tula, Russia
krasotkina@tsu.tula.ru
[2] Computing Centre of RAS

Abstract. The classical learning problem of the pattern recognition in a finite-dimensional linear space of real-valued features is studied under the conditions of a non-stationary universe. The training criterion of non-stationary pattern recognition is formulated as a generalization of the classical Support Vector Machine. The respective numerical algorithm has the computation complexity proportional to the length of the training time series.

The majority of modern methods for pattern recognition works under the assumption that properties of the environment and consequently required decision rule doesn't change in a process of data collection as well as during the exam. However, recently there have appeared many applications, in which samples of training set are entering the system over a long period of time, when the properties of the analyzed phenomenon may undergo considerable changes. The example of such problem is task of filtering irrelevant or publicity hyperlinks as a result of retrieval request. The behavior of internet advertising distributors constantly turns and it causes changes of publicity hyperlinks features that should result in the adequate correction of search mechanism. Often in this kind of tasks both the nature of changes in the environment and the fact of changes itself are hidden from direct observation and this makes learning even more difficult. In literature this problem is called *concept drift* [1].

At present there are three approaches to the construction of algorithms taking into account non-stationary character of decision rule: algorithms based on selecting instances, algorithms based on weighted instances and algorithms based on classifier selection and mergers. The goal of algorithms using selection of instances is the choice of prototypes which are relevant for decision rule at the present moment. As a rule, it is realized with the help of running window technology when decision rule at the present moment is made only on the basis of the instances obtained from previous time points. The examples of such algorithms can be FLORA family of algorithms [1] and TMF [2].

Algorithms based on weighting of instances [3] obtained from different time points use the ability of some learning algorithms such as Support Vector Machines (SVMs) method to assign weights to different instances. As a rule, weights are assigned to the instances according to their "age" (e.g. period of time from their obtaining).

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 24–29, 2011.

By ensemble-based approach [5] for pattern recognition in non-stationary environment, required decision rule is made as voting or weighted voting of classifiers obtained for different conditions. For example, in the paper [4] it is proposed to cluster training samples and to construct its own classifier for each cluster.

In general, we can mark that all existing algorithms are more or less heuristic and a certain set of heuristics is determined by a specificity of the current task. In this paper we propose stochastic concept of the non-stationary environment based on Bayesian approach to the pattern recognition problem. The main instantaneous property of the non-stationary environment is understood as time-dependent separating hyperplane that in the best way describes the differences between feature vectors of samples of two classes. The proposed concept brings about the learning algorithm that is a generalization of the classical SVM for the case when the parameters of the separating hyperplane change with time.

1 Bayesian Definition of the Pattern Recognition Problem in Non-stationary Environment

Let every instance of the environment $\omega \in \Omega$ is presented by a point in the linear feature space $\mathbf{x}(\omega) = \big(x^1(\omega), \ldots, x^n(\omega)\big) \in \mathbb{R}^n$, and its hidden membership in one of two classes is determined by an index value of the class $y(\omega) \in \{1, -1\}$. The paper [6] proposed a stochastic model of the universe. The main model assumption is a priori parametric distribution of the instances

$$\phi\left(\mathbf{x}|\mathbf{a}, b, y; c\right) = \begin{cases} \text{const}, & yz\left(\mathbf{a}, \mathbf{x}\right) \geq 1, \\ e^{-c(1-yz(\mathbf{a},x))}, & yz\left(\mathbf{a}, \mathbf{x}\right) < 1, \end{cases} \tag{1}$$

And this distribution is defined by the objectively existing hyperplane $z(\mathbf{x}, \mathbf{a}) = \mathbf{a}^T\mathbf{x} + b = 0$ with an unknown directing vector of features $\mathbf{a} = (a_1, a_2, ..., a_n)$, having a priori distribution $\Psi(\mathbf{a}, b|\sigma^2) \propto \exp\big(-\frac{1}{2\sigma^2}\mathbf{a}^T\mathbf{a}\big)$. Using the maximum a posteriori probability principle for the directing vector parameters estimation $(\mathbf{a}, b)$ leads us to the widely-known support vector criterion

$$\begin{cases} J(\mathbf{a}, b, \delta_1, ..., \delta_N|c) = \mathbf{a}^T\mathbf{a} + c\sum\limits_{j=1}^{N}\delta_j \rightarrow \min, \\ y_j\left(\mathbf{a}^T\mathbf{a} + b\right) \geq 1 - \delta_j,\ \delta_j \geq 0, j = 1, ..., N, \end{cases} \tag{2}$$

that was proposed by V. N. Vapnik from a strictly deterministic point of view. Of course, the concept of time is completely absent here.

The principal distinction of the concept of non-stationary environment given in this paper consists in taking into consideration time factor t. We suppose that the main property of non-stationary environment is expressed by time-dependent separating hyperplane that characterizes primary difference of the feature vectors of instances of two classes. This separating hyperplane, in turn, is completely defined by its own direction vector and location parameter, which should be considered as time functions $\mathbf{a}_t$ b_t: $f_t\big(\mathbf{x}(\omega)\big) = \mathbf{a}_t^T\mathbf{x} + b_t$. In the terms of the improper probability distributions of two classes in the feature space such idea is formulated in the form of hypothesis that parameters of this pair of distributions are time varying:

$$\phi\left(\mathbf{x}|\mathbf{a}_t, b_t, y; c\right) = \begin{cases} \text{const}, & yz(\mathbf{a}_t, \mathbf{x}) \geq 1, \\ e^{-c(1-yz(\mathbf{a}_t, x))}, & yz(\mathbf{a}_t, \mathbf{x}) < 1, \end{cases} \tag{3}$$

where $z(\mathbf{x}, \mathbf{a}_t) = \mathbf{a}_t^T \mathbf{x} + b = 0$.

We will suppose that at the zero time moment a priori distribution of the separating hyperplane is improper and has the appearance: $\psi_0(\mathbf{a}_0, b_0) \propto \psi_0(\mathbf{a}_0) = N(\mathbf{a}_0|\mathbf{0}, \mathbf{I})$. In turn we will regard directing vector $\mathbf{a}_j$ as a stochastic stationary process: $\mathbf{a}_t = q\mathbf{a}_{t-1} + \boldsymbol{\xi}_t, M(\boldsymbol{\xi}_t) = \mathbf{0}, M(\boldsymbol{\xi}_t \boldsymbol{\xi}_t^T) = d\mathbf{I}, 0 \leq q < 1$.

2 Dynamic Support Vector Machine Criterion

Now every instance $\omega \in \Omega$ is considered only together with the indication of time point when this instance was presented (ω, t). As a result the training set represents as the set of triplet $\left\{(\mathbf{X}_t \in \mathbb{R}^n,\ \mathbf{Y}_t,\ t)\right\}_{t=1}^T$, $(\mathbf{X}_t, \mathbf{Y}_t) = \{(\mathbf{x}_{k,t}, y_{k,t})\}_{k=1}^{N_t}$ - a subset of instances, entered in time point t.

We obtain the sought-for sequence $(\mathbf{a}_t, b_t)_{t=1}^T$ as maximum point of joint a priori distribution of separating hyperplane's parameters and training sample. The maximum a posteriori probability principle lets to the criterion

$$\begin{gathered} J(\mathbf{a}_t, b_t, \delta_{t,j}, t = 0, ..., T) = \mathbf{a}_0^T \mathbf{a}_0 + \tfrac{1}{d} \sum_{t=1}^{T} (\mathbf{a}_t - q\mathbf{a}_{t-1})^T (\mathbf{a}_t - q\mathbf{a}_{t-1}) + \\ + \tfrac{1}{d'} \sum_{t=1}^{T} (b_t - b_{t-1})^2 + \sum_{t=1}^{T} \sum_{j=1}^{N_t} \delta_{j,t} \to \min_{[\mathbf{a}_t, b_t]_{t=1}^T} \\ y_{j,t}(\mathbf{a}_t^T \mathbf{x}_{j,t} + b_t) \geq 1 - \delta_{j,t}, \delta_{j,t} \geq 0, j = 1, \ldots, N_t, t = 1, ..., T \end{gathered} \tag{4}$$

The criterion (4) realizes the conception of the rather smooth sequence of optimal separating hyperplanes as opposed to the conception of the single optimal hyperplane in (2).

As the classic learning problem, the dynamical problem (4) is a quadratic programming problem but contains $T(n + 1) + N$ variables in contrast to $(n + 1) + N$ variables in (2). It is known that a computational complexity of the general quadratic programming problem is proportional to the cube of the number of variables, i.e. a dynamic problem ex facte is more complex than the classical problem.

But the objective function in a dynamic problem (4) is pair-wise separable, i.e. representing a sum of private functions every of which depends on the variables connected with one or two time points in their increasing order. The algorithm of the pair-wise separable criterion optimization suggested in this paper consists in an approximate implementation of the dynamic programming procedure and permits to solve the problem mentioned above within the number of iterations proportional to the length of the training sequence.

3 Quickly Optimization Procedure for a Dynamic SVM Criterion

The algorithm for the optimization of the obtained pair-wise separable criterion proposed in this paper is based on the using a general principle of the dynamic programming. Let us to introduce the following notation $\mathbf{z}_t = \left[\mathbf{z'}_t^T \; z''_t \right]^T, \mathbf{z'}_t = \left[\mathbf{a}_t^T \; b_t \right]^T, \mathbf{z''}_t = [\delta_j]_{j=1}^{N_t}$:

$$\zeta_t(\mathbf{z'}_t) = (\mathbf{z'}_t - \mathbf{z}_t^0)^T \mathbf{Q}_t^0 (\mathbf{z'}_t - \mathbf{z}_t^0), \chi_t(\mathbf{z''}_t) = C \mathbf{e}_t^T \mathbf{z''}_t, \; \mathbf{e}_t = [1]_1^{N_t}, t = 1, ..., T,$$
$$\gamma_t(\mathbf{z'}_{t-1}, \mathbf{z'}_t) = (\mathbf{z'}_t - \mathbf{A}_j \mathbf{z'}_{t-1})^T \mathbf{U}_j (\mathbf{z'}_t - \mathbf{A}_j \mathbf{z'}_{\mathbf{t-1}})$$

and rewrite the criterion (4) in a more convenient form

$$J(\mathbf{z}_0, ..., \mathbf{z}_T) = \sum_{t=0}^{T} \zeta_t(\mathbf{z}'_t) + \sum_{t=0}^{T} \chi(\mathbf{z}''_t) + \sum_{t=1}^{T} \gamma_t(\mathbf{z}'_{t-1}, \mathbf{z}'_t) \rightarrow \min, \; \mathbf{z}_t \in Z_t$$

where the areas of acceptable values for variables are determined by the conditions

$$\left\{ \mathbf{z} \in R^{n+2} : \mathbf{g}_j^T \cdot \mathbf{z}'_t + z''_j - 1 \geq 0 \;, j = (N_{t-1} + 1), ..., N_t, \; t = 0, ..., T, \mathbf{z}''_t \geq 0 \right\}$$

The central idea of the dynamical programming method uses the concept of a sequence of Bellman functions $\tilde{J}_t(\mathbf{z}_t) = \min_{\mathbf{z}_0, \ldots, \mathbf{z}_{t-1}} J_t([\mathbf{z}_s]_{s=1}^t)$, $[\mathbf{z}_s \in Z_s]_{s=0}^{t-1}$, connected with partial criteria $J_t(\mathbf{z}_0, ..., \mathbf{z}_t) = \sum_{s=0}^{t} \zeta_s(\mathbf{z}'_s) + \sum_{s=0}^{t} \chi(\mathbf{z}_s) + \sum_{s=1}^{t} \gamma_s(\mathbf{z'}_{s-1}, \mathbf{z}_s{}')$ having the same structure as the full objective function but defined on the set of variables $Z_t = (\mathbf{z}_s, s = 0, ..., t)$. In order to obtain the filtering estimations of the separating hyperplane's parameters we will use the fundamental property of Bellman function

$$\tilde{J}_t(\mathbf{z_t}) = \zeta_t(\mathbf{z'}_t) + \chi(\mathbf{z''}_t) + \min_{\mathbf{z'}_{t-1}, \mathbf{z''}_{t-1}} \left[\gamma_t(\mathbf{z'}_{t-1}, \mathbf{z'}_t) + \tilde{J}_{t-1}(\mathbf{z}_{t-1}) \right], \qquad (5)$$

which is called the direct recurrence relation. The procedure begins with the first Bellman function $\tilde{J}_0(\mathbf{z}_0) = \zeta_0(\mathbf{z}_0{}') + \chi(\mathbf{z''_0})$. Then Bellman functions are recurrently evaluated for the following observation. And the minimum of Bellman function on every step determines the filtering value of the parameters of the optimal separating hyperplane

$$\hat{\mathbf{z}}_t = \arg\min_{\mathbf{z}_t} \hat{J}_t(\mathbf{z}_t), \mathbf{z}_t \in Z_t. \qquad (6)$$

The optimization procedure is based on the hypothesis that there exists an appropriate compact form for Bellman functions representation, allowed to store this functions in the memory. But in the case then inequality constraints are imposed upon the sought-for variables and the Bellman functions are piecewise quadratic and consequently the dynamic programming procedure cannot be applied immediately. In order to save the computation advantages of the dynamic programming procedure we use here the following trick.

We heuristically replace the non quadratic functions $F_t(\mathbf{z}'{}_t) = \min\limits_{\mathbf{z_{t-1}} \in \mathbf{Z_{t-1}}} \left[\gamma_t(\mathbf{z}'{}_{t-1}, \mathbf{z}'{}_t) + \tilde{J}_{t-1}(\mathbf{z}_{t-1})\right]$ by the some appropriate quadratic approximation $\hat{F}_t(\mathbf{z}'_t) = \hat{c}_t + (\mathbf{z}'_t - \hat{\mathbf{z}}_t)^T \hat{\mathbf{Q}}_t (\mathbf{z}'_t - \hat{\mathbf{z}}_t)$. Then the following approximations of Bellman functions will be quadratic too and a numerical implementation of the dynamic programming procedure will be possible. Thus, the quadratic approximation of the Bellman function comes to the selection of appropriate values of the parameters $(\hat{c}_t, \hat{\mathbf{z}}_t, \hat{\mathbf{Q}}_t)$ of the quadratic function $\hat{F}_t(\mathbf{z}'_t)$, which would ensure invariance conservation of the main features of, generally speaking, non-quadratic function and consequently the initial Bellman function. Such features are the minimum point position $\hat{\mathbf{z}}_t = \arg\min F_t(\mathbf{z}'_\mathbf{t})$, minimum point value $\hat{c}_t = \min F_t(\mathbf{z}'_t)$, and also the matrix of the second derivatives at the minimum point $\hat{\mathbf{Q}}_t = \nabla^2 F_t(\mathbf{z}'_t)\big|_{\arg\min F_t(\mathbf{z}'_t)}$.

4 Case Study: Spam-Filtering Problem

The object of the experimental study in this paper is the problem of filtering spam-addresses in a result of retrieval request. The behavior of distributors of network advertising constantly improves and as a result a classifier of advertising links used by a search engine should adapt to the behavior of spammers. Consequently we come to problem of construction the time-dependent decision rule.

As the training data we took an anonymous set of hyperlinks URL Reputation Data Set from the repository UCI, first described in the paper [7]. This set consists of the addresses of Web resources for 121 days grouped according to the observation days and contains both advertising and relevant hyperlinks. Every instance in the data base is characterized by 3.2 million features which may be divided into two groups: lexical (hostname, primary domain, TLD and etc.) and host-based (WHOIS info, IP prefix, geographic and etc.). The values of features are normalized from 0 to 1, and the features themselves are anonymous. For carrying out the experiments from this data base in different ways there have been randomly selected 10 instances for each of 11 days taken from the 1st to the 110th day with the step of 10 days. The testing set was comprised by accidentally selected 4000 instances of the 120th day. We deleted the features with the values that are equal on all instances of the training set out of feature descriptions of the instances participating in the experiment, the number of the rest was 326. The experiment was repeated for different sets of training instances. Two algorithms were compared in the course of the experiment: normal support vector machine with the hyperplane, constructed on all instances of the training set and the algorithm of successive refinement of the decision rule, suggested in the paper. In the proposed algorithm of dynamic pattern recognition the parameters of the required decision rule d and d' were chosen by cross-validation procedure, on each iteration of which non-stationary decision rule was constructed without taking into account the instances entered over one of the days. The general per cent of erroneously classified links ε was calculated for each algorithm: the per cent of malicious URLs erroneously classified as benign ones, ε^- and the per cent of benign URLs, erroneously classified as malicious ones ε^+. As the results given in the table show, the incremental learning algorithm of the pattern recognition allows to improve significantly the quality of recognition that indirectly confirms non-stationary character of the data used for the experiments.

Algorithm	ε^-, %	ε^+, %	ε, %
SVM	32.32	5.94	16.0
Incremental Algorithm	14.96	4.49	8.49

5 Conclusions

In the paper the adaptation of the general definition of the pattern recognition learning problem with two classes of instances in the finite-dimensional space of real features is made under the conditions of non-stationary environment. The non-stationary property of environment is considered as time-changing decision rule. Under this assumption the learning criterion appeared a dynamic modification of SVM criterion. Also we have suggested optimization algorithm for Dynamic SVM having the linear computational complexity relative to the length of the training time series. Our algorithm was applied to the problem of filtering irrelevant or publicity hyperlinks as a result of retrieval request and its application achieved a good enough result as compared with classical SVM.

References

1. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Machine Learning 23(1), 69–101 (1996)
2. Salganicoff, M.: Tolerating concept and sampling shift in lazy learning using prediction error context switching. AI Review, Special Issue on Lazzy Learning 11(1-5), 133–155 (1997)
3. Klinkenberg, R.: Learning drifting concepts example selection vs. example weighting. Intelligent data analysis, Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift 8(3) (2004)
4. Harries, M., Sammut, C., Horn, K.: Extracting hidden context. Machine Learning 32(2), 101–126 (1998)
5. Muhlbaier, M.D., Polikar, R.: An Ensemble Approach for Incremental Learning in Nonstationary Environments. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 490–500. Springer, Heidelberg (2007)
6. Tatarchuk, A.I., Sulimova, V.V., Mottl, V.V., Windridge, D.: Method of relevant potential functions for selective combination of diverse information in the pattern recognition learning based on Bayesian approach. In: MMRO-14: Conf. Proc., Suzdal, pp. 188–191 (2009)
7. Ma, J., Saul, K.L., Savage, S., Voelker, G.: Identifying Suspicious URLs: An Application of Large-Scale Online Learning. In: Proceedings of the International Conference on Machine Learning (ICML), Montreal, Quebec, June 2009, pp. 681–688 (2009)