Krassimir Markov, Vladimir Ryazanov,
Vitalii Velychko, Levon Aslanyan
(editors)

New Trends
in
Classification and Data Mining

I T H E A
SOFIA
2010

Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan (ed.)

New Trends in Classification and Data Mining

ITHEA®

Sofia, Bulgaria, 2010

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

ISBN 978-954-16-0042-9

C\o Jusautor, Sofia, 2010

MINIMIZATION OF EMPIRICAL RISK IN LINEAR CLASSIFIER PROBLEM

Yurii I. Zhuravlev, Yury Laptin, Alexander Vinogradov

Abstract: *Mixed-integer formulation of the problem of minimization of empirical risk is considered. Some possibilities of decision of the continuous relaxation of this problem are analyzed. Comparison of the proposed continuous relaxation with a similar SVM problem is performed too.*

Keywords: *cluster, decision rule, discriminant function, linear and non-linear programming, non-smooth optimization*

ACM Classification Keywords: *G.1.6 Optimization - Gradient methods, I.5 Pattern Recognition; I.5.2 Design Methodology - Classifier design and evaluation*

Acknowledgement: *This work was done in the framework of Joint project of the National Academy of Sciences of Ukraine and the Russian Foundation for Fundamental Research No 08-01-90427 'Methods of automatic intellectual data analysis in tasks of recognition objects with complex relations'.*

Introduction

Recently considerable number of researches are devoted to problems of construction of linear algorithms of classification (classifiers). In many cases such problems are considered for classification of two sets. Usually linear classifier problems are formulated for the case of linearly separable sets. In separable case the mentioned problems can be efficiently solved [1–4]. The concept of optimality for two linearly separable sets has a simple geometrical sense – the optimum classifier defines the strip of maximal width separating these sets.

For linear separability of two finit sets it is necessary and sufficient for convex envelops of these sets don't intersect each other. But this condition is not sufficient in the case of more than two sets. In [5–7] some sufficient conditions of linear separability of any number of final sets are formulated.

Minimization of the empirical risk is the natural criterion of choice of the classifier in case of linearly inseparable sets. In this paper, a mixed-integer formulation of the problem of minimization of empirical risk is considered, and some possibilities of decision of the continuous relaxation of this problem are analyzed. Comparison of the proposed continuous relaxation with a similar SVM problem is performed too.

1. Problem formulation

Let a set of linear functions is defined $f_i(x,W^i) = (w^i, x) + w_0^i$, $i = 1,\dots, m$, where $x \in R^n$ is attribute vector, and $W^i = (w_0^i, w^i) \in R^{n+1}$, $i = 1,\dots,m$, are vectors of parameters. We denote $W = (W^1,\dots, W^m)$, $W \in R^L$, $L = m(n+1)$. Let's consider linear algorithms of classification (linear classifiers) of the following kind

$$a(x,W) = \arg\max_i \left\{ f_i(x,W^i) :\ i = 1,\dots, m \right\};\ x \in R^n; W \in R^L \tag{1}$$

In [6] also classifiers, in which f_i are convex piece-wise linear functions, were investigated.

Here it is considered a family of finite not intersected sets Ω_i, $i=1,..., m$. We will say that the classifier $a(x,W)$ separates correctly points from Ω_i, $i=1,..., m$, if $a(x,W)=i$ for all $x\in\Omega_i$, $i=1,..., m$.

Sets Ω_i, $i=1,..., m$ are called *linearly separable* if there is a linear classifier correctly separating points from these sets.

Each set Ω_i, $i=1,..., m$ is a training sample of points from some class $\overline{\Omega}_i$ known only on these sample units. The training process for classifier $a(x,W)$ consists in selection of parameters W at which classes $\overline{\Omega}_i$, $i=1,..., m$ are separated in the best way (in some sense). For definition of the quality of separation various approaches are used.

Let $\Omega=\bigcup_{i=1}^{m}\Omega_i$, points of the set Ω are enumerated, T is the set of indices, $\Omega=\left\{x^t:\ t\in T\right\}$, T_i is a subset of indices correponding to points from Ω_i, $\Omega_i=\left\{x^t:\ t\in T_i\right\}$, $T=\bigcup_{i=1}^{m}T_i$. Let function $i(t)$ returns the index of the set Ω_i, to which the point x^t belongs, $t\in T$. The value

$$g^t(W)=\min\left\{f_i(x^t,W^i)-f_j(x^t,W^j):\ j\in\{1,..., m\}\setminus i,\ i=i(t)\right\}=$$
$$=\min\left\{(w^i-w^j,x^t)+w_0^i-w_0^j\ :\ j\in\{1,..., m\}\setminus i,\ i=i(t)\right\} \quad (2)$$

is called as a *margin* or a *gap* of the classifier $a(x,W)$ on the point x^t, $t\in T$.

The classifier $a(x,W)$ makes a mistake in a point x^t iff the gap $g^t(W)$ is negative.

The value $g(W)=\min\left\{g^t(W): t\in T\right\}$ is called as a gap of the classifier $a(x,W)$ on the family of sets Ω_i, $i=1,..., m$.

The classifier $a(x,W)$ correctly separates points from Ω_i, $i=1,..., m$, if $g(W)>0$.

Remark 1. The classifier $a(x,W)$ is invariant with respect to multiplication of all functions f_i (vectors W^i) by positive number, and the gap $g(W)$ is linear with respect to such multiplication. The classifier $a(x,W)$ and the gap $g(W)$ are invariant concerning to addition of any real number to all f_i.

The value $g(W)$ can be used as a criterion of quality of the classifier $a(x,W)$ (the more value $g(W)$, the more reliably points from Ω_i, $i=1,..., m$ are separated). However, it is necessary to take into account a norm for the family of vectors W which we denote $\eta(W)$ and name *norm* of the classifier $a(x,W)$.

Let's use the following function:

$$\eta(W)=\sqrt{\sum_{i=1}^{m}\sum_{k=1}^{n}(w_k^i)^2} \quad (3)$$

Other functions also can be used as a norm $\eta(W)$ [6].

Let the family of sets Ω_i, $i=1,\ldots,m$ is given. Taking into account the introduced notations the optimal classifier problem we write as following: find

$$g^* = \max_W \left\{ g(W) : \eta(W) \le 1,\ W \in R^L \right\} \tag{4}$$

Since the vector $W=0$ is feasible, the problem (4) always has a solution, and $g^* \ge g(0) = 0$. Let's notice that $g^* > 0$ if sets Ω_i, $i=1,\ldots,m$ are linearly separable, i.e. there is the linear classifier correctly separating these sets. We will consider also a problem: find

$$\eta^* = \min_V \left\{ \eta(V) : g(V) \ge 1,\ V \in R^L \right\} \tag{5}$$

Similar problems were considered by different authors (see, e.g., [4, 8]).

Lemma 1. Let W^* be an optimal solution to the problem (4). Then

1) if $g^* > 0$, the problem (5) also has the optimum solution V^*, and $V^* = W^* / g^*$, $\eta^* = 1 / g^*$;

2) if $g^* = 0$, the problem (5) has no feasible solutions.

The proof is simple (see [6]).

Let's consider in more details problems of construction of linear classifiers for the family of sets $\Omega_i = \left\{ x^t,\ t \in T_i \right\}$, $i=1,\ldots,m$. It is easy to see that the problem (4) can be represented as a LP- problem with additional quadratic constraint: find

$$g^* = \max_{w,\delta} \ \delta \tag{6}$$

subject to

$$(w^i - w^j, x^t) + w_0^i - w_0^j \ge \delta,\ \ j \in \{1,\ldots,m\} \setminus i,\ t \in T_i,\ i=1,\ldots,m \tag{7}$$

$$\sum_{i=1}^{m} \sum_{k=1}^{n} (w_k^i)^2 \le 1 \tag{8}$$

The problem (5) is a quadratic programming problem: find

$$\eta^* = \min_v \sum_{i=1}^{m} \sum_{k=1}^{n} (v_k^i)^2 \tag{9}$$

subject to

$$(v^i - v^j, x^t) + v_0^i - v_0^j \ge 1,\ \ j \in \{1,\ldots,m\} \setminus i,\ t \in T_i,\ i=1,\ldots,m \tag{10}$$

It is possible to show that in case $m=2$ the problem (9) – (10) is equivalent to the problem which is used for construction of the strip of the maximum width separating some linearly separable sets Ω_1, Ω_2 .

Existing efficient software packages for optimization problems of general purpose can be used for considered problems, if the number of points in training sample is not too large [6]. For a large number of points in training sample, it is appropriate to use non-smooth optimization methods [8, 9].

Problems (6) – (8) and (9) – (10) allow to find the optimum linear classifier only for linearly separable sets. For linearly inseparable sets the problem should be formulated in other way.

2. Empirical risk minimization

In the case of linearly inseparable training sample a natural criterion for choosing classifier is that of minimizing empirical risk, i.e. the number of training sample points which are separated by the classifier incorrectly.

Suppose that a reliability parameter $\overline{\delta} > 0$ is fixed for separation of points of the training sample Ω_i, $i = 1, ..., m$. We say that the points x^t, $t \in T$ are separated by the classifier unreliably, if $g^t(W) < \overline{\delta}$.

Below the value of empirical risk will be determined by reliability, characterized by parameter $\overline{\delta}$, i.e. the empirical risk is equal to the number of points of the training sample, which are separated by the classifier incorrectly or unreliably.

Lemma 3 [6]. Let $x^{\alpha} \in \Omega_i$, $x^{\beta} \in \Omega_j$, classifier $a(x, W)$ separates these points correctly, and for the norm of the classifier the constraint (8) is valid. Then

$$-R \le w_0^i - w_0^j \le R \tag{11}$$

where $R = \max\left\{\|x\| : x \in \Omega_i,\ i = 1, ..., m\right\}$.

Let $\Omega_i = \left\{x^t,\ t \in T_i\right\}$, $i = 1, ..., m$, $T = \bigcup_{i=1}^{m} T_i$. To each point x^t, $t \in T$ we associate a variable $y_t = 0 \vee 1$ so that $y_t = 0$, if the point x^t is considered in the problem (6)–(8), and $y_t = 1$ – otherwise.

Let a large positive number B be given. Empirical risk minimization problem based on reliability parameter $\overline{\delta}$ has the following form: find

$$Q^* = \min_{w,y}\left\{\sum_{t \in T} y_t\right\} \tag{12}$$

subject to

$$(w^i - w^j, x^t) + w_0^i - w_0^j \ge \overline{\delta} - B \cdot y_t,\ \ j \in \{1, ..., m\} \setminus i,\ \ t \in T_i,\ \ i = 1, ..., m \tag{13}$$

$$\eta(W) \le 1 \tag{14}$$

$$\sum_{t \in T_i} y_t \le |T_i| - 1,\quad i = 1, ..., m \tag{15}$$

$$0 \le y_t \le 1,\ t \in T \tag{16}$$

$$y_t = 0 \vee 1,\ t \in T \tag{17}$$

From (14), (11) follows that if $y_t = 1$, then for sufficiently large values B the corresponding inequalities of the form (13) are always valid, i.e. point x^t is excluded from the problem. Constraints (15) mean that at least one point from each set Ω_i must be included in the problem.

The optimal value Q^* is equal to the minimum empirical risk based on reliability $\overline{\delta}$. Problem (12)-(17) is NP-hard; the branch and bound method can be used to solve it. To calculate the lower bounds for Q^* (minimum empirical risk), let's consider the continuous relaxation of the mentioned above problem – the problem (12)–(16). The optimum value of the relaxed problem is denoted q^*. To solve this problem we use decomposition on the variables W. Let variables W are fixed. Given (2), the problem of minimizing on the variables y takes the following form: find

$$q(W) = \min_y \left\{ \sum_{t \in T} y_t \right\} \tag{18}$$

subject to

$$y_t \geq \frac{1}{B}\left(\overline{\delta} - g^t(W)\right),\ t \in T \tag{19}$$

$$\eta(W) \leq 1 \tag{20}$$

$$\sum_{t \in T_i} y_t \leq |T_i| - 1, \quad i = 1, \dots, m \tag{21}$$

$$0 \leq y_t \leq 1,\ t \in T \tag{22}$$

Denote $d^t(W) = \max\left(0, \frac{1}{B}\left(\overline{\delta} - g^t(W)\right)\right)$. Obviously, if the problem (18)-(22) has a solution, then $y^t = d^t(W)$. So, we get the minimization problem on variables W: find

$$q^* = \min \sum_{t \in T} d^t(W) \tag{23}$$

subject to

$$\eta(W) \leq 1 \tag{24}$$

$$\sum_{t \in T_i} d^t(W) \leq |T_i| - 1,\ i = 1, \dots, m \tag{25}$$

$$d^t(W) \leq 1,\ t \in T \tag{26}$$

Functions $d^t(W)$ are convex piecewise-linear, $\eta(W)$ is quadratic and positively defined. To solve the problem (23)-(26) it is appropriate to apply efficient methods of nonsmooth optimization [9].

3. Comparison with support vector machine

Let us consider the case of two classes. Suppose, as previously, $\Omega_i = \left\{x^t, \ t \in T_i\right\}, \ i = 1, \ 2, \ T = T_1 \bigcup T_2$. In the method of support vectors (see eg [1]) to build a classifier which separates the two linearly inseparable sets, one has to solve the following problem: find

$$\min_{u, \, u_0, \xi} \left\{ \frac{1}{2}(u, \, u) + C \cdot \sum_{t \in T} \xi^t \right\} \tag{27}$$

subject to

$$(u, \, x^t) + u_0 \geq 1 - \xi^t, \ t \in T_1 \tag{28}$$

$$(-u, \, x^t) - u_0 \geq 1 - \xi^t, \ t \in T_2 \tag{29}$$

$$\xi^t \geq 0, \ t \in T \tag{30}$$

where $u \in R^n$, $u_0 \in R$, $\xi^t \in R, \ t \in T$.

To compare these approaches we consider an analogue of (12)–(16) for the case of two sets (in the case of two sets $\Omega_i = \left\{x^t, \ t \in T_i\right\}, \ i = 1, 2$ to build a linear classifier we need only two functions $f_i(x, W^i) = (w^i, x) + w_0^i, \ i = 1, 2$, where $f_1(x) = -f_2(x)$): find

$$q^* = \min_{w, w_0, y} \left\{ \sum_{t \in T} y_t \right\} \tag{31}$$

subject to

$$(w, x^t) + w_0 \geq \overline{\delta} - B \cdot y_t, \ \ t \in T_1 \tag{32}$$

$$(-w, x^t) - w_0 \geq \overline{\delta} - B \cdot y_t, \ \ t \in T_2 \tag{33}$$

$$(w, w) \leq 1 \tag{34}$$

$$\sum_{t \in T_i} y_t \leq \left|T_i\right| - 1, \quad i = 1, 2 \tag{35}$$

$$0 \leq y_t \leq 1, \ t \in T \tag{36}$$

Change of variables in the problem (31)–(36): $w = \overline{\delta} u, \ \ w_0 = \overline{\delta} u_0$, $\xi^t = \dfrac{B y_t}{\overline{\delta}}, \ t \in T_1 \bigcup T_2$, gives

$$q^* = \frac{\overline{\delta}}{B} \cdot \min_{u, u_0, \xi} \left\{ \sum_{t \in T} \xi^t \right\} \tag{37}$$

subject to

$$(u, x^t) + u_0 \geq 1 - \xi^t, \ \ t \in T_1 \tag{38}$$

$$(-u, x^t) - u_0 \geq 1 - \xi^t, \ \ t \in T_2 \tag{39}$$

$$(u, u) \leq 1 \big/ \overline{\delta}^2 \tag{40}$$

$$\xi^t \geq 0,\ t \in T \tag{41}$$

$$\xi^t \leq \frac{B}{\overline{\delta}},\ t \in T \tag{42}$$

$$\sum_{t \in T_i} \xi^t \leq \frac{B}{\overline{\delta}}\left(|T_i| - 1\right), \quad i = 1, 2 \tag{43}$$

Denote $\chi,\ \gamma_i,\ i = 1,2$ the dual variables for constraints (40), (43) and consider the Lagrangian

$$L(\chi,\gamma,\xi,u) = \frac{\overline{\delta}}{B}\sum_{t \in T}\xi^t + \chi \cdot ((u,u) - {}^{1}\!/\!_{\overline{\delta}^2}) + \sum_{i=1}^{2}\gamma_i\left(\sum_{t \in T_i}\xi^t - \frac{B}{\overline{\delta}}\left(|T_i| - 1\right)\right)$$

Let:

$$\varphi(\chi,\gamma) = \min_{u,u_0,\xi} L(\chi,\gamma,\xi,u) \tag{44}$$

subject to (38), (39), (41), (42).

Suppose a penalty factor C in the problem (27)-(30) is given. It is easy to see that, if we take $\gamma = 0$ and choose χ from the condition $\dfrac{\overline{\delta}}{2\chi B} = C$, we obtain

$$L(\chi,\gamma,\xi,u) = 2\chi\left\{\frac{1}{2}(u,u) + C \cdot \sum_{t \in T}\xi^t\right\} - \frac{\chi}{\overline{\delta}^2}.$$

So, the problem (44), (38), (39), (41) is equivalent to (27)–(30) for the dual variables chosen above. Constraints (42) can be neglected at small $\overline{\delta}$ and large B.

Thus, the SVM problem is a special case of (44), (38), (39), (41).

References

1. Воронцов К.В. Машинное обучение. – http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций%2C_К.В.Воронцов) – Последнее изменение: 30 мая 2009
2. Местецкий Л.М. Математические методы распознавания образов. – http://www.intuit.ru/department/graphics/imageproc/ – Опубликовано 30.04.2008
3. Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания. - Киев: Наук.думка, 2008. - 232 с.
4. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. – К.: Наукова думка, 2004. – 545 с.
5. Laptin Yu., Vinogradov A. Exact discriminant function design using some optimization techniques // "Classification, Forecasting, Data Mining" International Book Series "INFORMATION SCIENCE & COMPUTING", Numder 8, Sofia, Bulgaria, 2009. – Pages 14-19.
6. Лаптин Ю.П., Виноградов А.П., Лиховид А.П. О некоторых подходах к проблеме построения линейных классификаторов в случае многих классов // Pattern Recognition and Image Analysis, 2010 (Сдана в печать)
7. Петунин Ю.И, Шульдешов Г.А. Проблемы распознавания образов с помощью линейных дискриминантных функций Фишера – Кибернетика, 1979, № 6, с. 134-137.
8. Методи негладкої оптимізації у спеціальних задачах класифікації, Стецюк П.І., Березовський О.А., Журбенко М.Г., Кропотов Д.О. – Київ, 2009. – 28 с. – (Препр./НАН України. Ін-т кібернетики ім.. В.М.Глушкова; 2009–1)

9. Shor N.Z. Nondifferentiable Optimization and Polynomial Problems. – Dordrecht, Kluwer, 1998. – 394 p.
10. Koel Das, Zoran Nenadic. An efficient discriminant-based solution for small sample size problem // Pattern Recognition – Volume 42, Issue 5, 2009, Pages 857-866.
11. Juliang Zhang, Yong Shi, Peng Zhang. Several multi-criteria programming methods for classification // Computers & Operations Research – Volume 36, Issue 3, 2009, Pages 823-836.
12. E. Dogantekin, A. Dogantekin, D. Avci Automatic Hepatitis Diagnosis System based on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System // Expert Systems with Applications, In Press, 2009.

Information about authors

Yurii I. Zhuravlev *– Academician, Deputy Director, Dorodnicyn Computing Centre of the RAS, Vavilova 40, 119333 Moscow, Russian Federation*

Yury Laptin *– Senior Researcher, V.M.Glushkov Institute of Cybernetics of the NASU, Prospekt Academika Glushkova, 40, 03650 Kyiv, Ukraine; e-mail: laptin_yu_p@mail.ru*

Alexander Vinogradov *– Senior Researcher, Dorodnicyn Computing Centre of the RAS, Vavilova 40, 119333 Moscow, Russian Federation; e-mail: vngrccas@mail.ru*