УДК 519.7

НАХОЖДЕНИЕ ОПТИМАЛЬНОЙ РАЗДЕЛЯЮЩЕЙ ГИПЕРПЛОСКОСТИ НА ОСНОВЕ ЛОКАЛЬНОЙ МИНИМИЗАЦИИ РИСКА

А.А. Галкин

Киевский национальный университет им. Т. Шевченко, Украина

oleksandr.galkin@mail.ru

Поступила в редакцию 27.11.2012

Исследована методология применения опорно-векторных машин на основе локальной минимизации риска. Рассмотрена проблема нахождения оптимальной разделяющей гиперплоскости в случае линейно неразделимых данных. Представлены алгоритм слабого поля опорно-векторных машин в виде выпуклой аппроксимации линейной минимизации риска со сферической оценкой Гаусса и методология построения гиперплоскости в многомерном пространстве.

Ключевые слова: локальная минимизация риска, разделяющая гиперплоскость, опорно-векторная машина, пространство характеристик.

Введение

Рассмотрим случай локальной минимизации риска для задачи классификации со сферическими окнами Парцена со стандартным отклонением σ и с семейством F линейных функций $f_{w,b}(x) = wx + b$.

Предположим, что учебные данные являются линейно раздельными, а это значит, что существует такое (w,b), что $\forall i,\ y_i(wx_i + b) > 0$.

Отметим, что локальный риск может быть представлен в следующем виде:

$$R_{лок}(w,b) = \frac{1}{n}\sum_{i=1}^{n}\int \mathrm{I}_{sgn}(wx_i + b + \varepsilon_w) \neq$$
$$\neq y_i d\mathrm{N}_\sigma(\varepsilon_w) = \frac{1}{n}\sum_{i=1}^{n}\varphi - \left(\frac{y_i(wx_i + b)}{\sigma}\right). \quad (1)$$

В случае, когда $\sigma \to 0$, наибольшее влияние на локальный риск (1) имеют термины, отвечающие примерам, расстояние которых до предела решения минимально. Действительно,

$$1 - \left|\mathrm{erf}(x)\right| \underset{x\to\pm\infty}{\sim} \frac{\exp(-x^2)}{\sqrt{\pi}x} \text{ и}$$
$$\varphi(-x) = \frac{1 + \mathrm{erf}(-x/\sqrt{2})}{2} \underset{x\to+\infty}{\sim} \frac{\exp(-x^2/2)}{\sqrt{2\pi}x}.$$

Пусть $v_i = y_i(wx_i + b) > 0$, где (w,b) является гиперплоскостью, разделяющей учебные примеры и $v_{\min} = \min v_i$, что является расстоянием от ближайшей точки до гиперплоскости. Оказывается, что

$$R_{лок}(w,b) = \frac{1}{n}\sum_{i=1}^{n}\varphi(-v_i/\sigma) \underset{\sigma\to 0}{\sim} \#$$
$$\#\{i, v_i = v_{\min}\}\frac{\exp(-v_{\min}^2/2\sigma^2)}{\sqrt{2\pi}v_{\min}/\sigma}.$$

Из предыдущего уравнения видно, что σ стремится к нулю, а локальный риск является минимальным, когда $v_{\min}$ является максимальным.

Итак, мотивацией нахождения оптимальной разделяющей гиперплоскости (рис. 1) является:

Найти такую разделяющую гиперплоскость, чтобы расстояние от ближайшей точки до данной гиперплоскости (*поля*) *было максимальным.*

Оптимальная разделяющая гиперплоскость

Как уже отмечалось выше, главной целью является нахождение оптимальной разделяющей гиперплоскости (ОРГ), которая определяется как

$$(w_0, b_0) = \arg\max_{(w,b)}\{\min_i y_i(wx_i + b), \|w\| = 1\}.$$

Если провести масштабирование (w,b) таким образом, что $\min_i y_i(wx_i + b) = 1$, то ОРГ будет также решением следующей задачи оптимизации:

$$\min w^2 \quad (2)$$

при следующем ограничении:

$$y_i(wx_i + b) \geq 1. \quad (3)$$

На рис. 1 учебные примеры разделяются гиперплоскостью правильно. Однако оптимальная разделяющая гиперплоскость на правом рисунке имеет большее поле, а следовательно, и меньший локальный риск. Интуитивно понятно, что этот случай является менее чувствительным к шумам в обучающем множестве.

Поскольку w^2 является выпуклым, минимизация уравнения (2) при линейных ограничениях (3) может быть достигнута за счет использо-

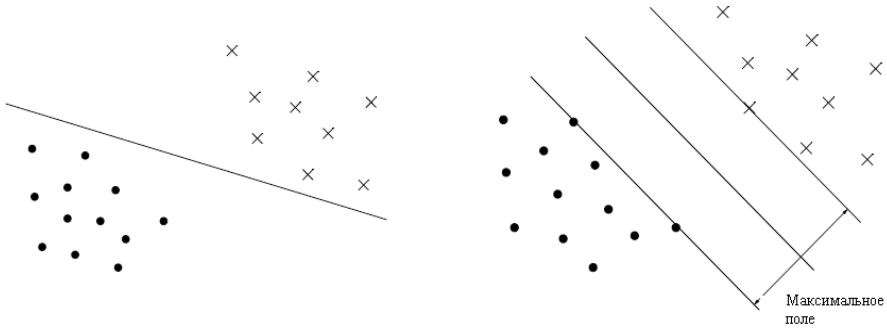

Рис. 1. Разделение учебных примеров с помощью гиперплоскости

вания множителей Лагранжа. Пусть мы имеем n неотрицательных множителей Лагранжа, связанных с ограничениями (3) через $\alpha = (\alpha_1,...,\alpha_n)$. Для минимизации (2) необходимо отыскание седловой точки функции Лагранжа

$$L(w,b,\alpha) = \frac{1}{2}w^2 - \sum_{i=1}^{n}\alpha_i[y_i(wx_i + b) - 1]. \quad (4)$$

Для того чтобы найти седловую точку, необходимо минимизировать функцию (4) по w и b, а также максимизировать ее по множителям Лагранжа $\alpha_i \geq 0$. Седловая точка должна удовлетворять следующим условиям:

$$\frac{\partial L(w,b,\alpha)}{\partial b} = \sum_{i=1}^{n} y_i\alpha_i = 0, \quad (5)$$

$$\frac{\partial L(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0. \quad (6)$$

При подстановке уравнений (5) и (6) в (4) задача оптимизации сводится к максимизации

$$W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j \quad (7)$$

с $\alpha_i \geq 0$ и ограничением (5). Это может быть достигнуто за счет использования стандартных методов квадратичного программирования [1].

Как только вектор решения $\alpha^0 = (\alpha_1^0,...,\alpha_n^0)$ задачи максимизации (7) найден, получаем, учитывая (6), что оптимальная разделительная гиперплоскость (w_0, b_0) будет иметь следующее расширение:

$$w_0 = \sum_{i=1}^{n}\alpha_i^0 y_i x_i, \quad (8)$$

в то время как b_0 может быть определено из условия Куна–Таккера

$$\alpha_i^0[y_i(w_0 x_i + b_0) - 1] = 0. \quad (9)$$

Заметим, что из уравнения (9) следует, что точки, для которых $\alpha_i^0 \geq 0$, удовлетворяют (3). Геометрически это означает, что они являются ближайшими точками к оптимальной гиперплоскости (рис. 1). Эти точки играют важную роль, поскольку только они являются точками, которые необходимы в выражении ОРГ. Они называются опорными векторами, что указывает на то, что они «поддерживают» расширение w_0.

Проблема классификации новой точки x решается путем учета знака

$$w_0 x + b_0.$$

С учетом расширения w_0 (8) функция решения гиперплоскости может быть записана как

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n}\alpha_i^0 y_i x_i x + b).$$

Несмотря на наличие мотивации максимизации поля с использованием принципа индукции минимизации локального риска (МЛР), стоит обратить внимание, что стандартным способом обоснования оптимальной разделяющей гиперплоскости является использование аргумента размерности Вапника–Червоненкиса (ВЧ). Действительно, для размерности ВЧ гиперплоскости, которая разделяет учебные точки, поле ограничено R^2/M^2, где R является радиусом наименьшей сферы, которая содержит учебные точки, а M – полем, полученным на учебных точках [2]. Обобщающие ограничения, зависящие от поля, можно найти в [3, 4].

Случай линейной неразделимости

Если данные не являются линейно раздельными, проблема нахождения оптимальной разделяющей гиперплоскости становится бесцельной. Кроме того, проблема ОРГ была мотивирована минимизацией локального риска (1), когда пропускная способность σ стремится к нулю. Когда σ не стремится к нулю, соответствующая функция потерь – сигмоидная функция, которая не является выпуклой. Кусочно-линейная функция потерь на рис. 2 представляет собой выпуклую аппроксимацию локальных потерь: в начале она имеет такой же наклон, как и функция потерь φ, а точкой соединения является $-\sqrt{\pi/2}$.

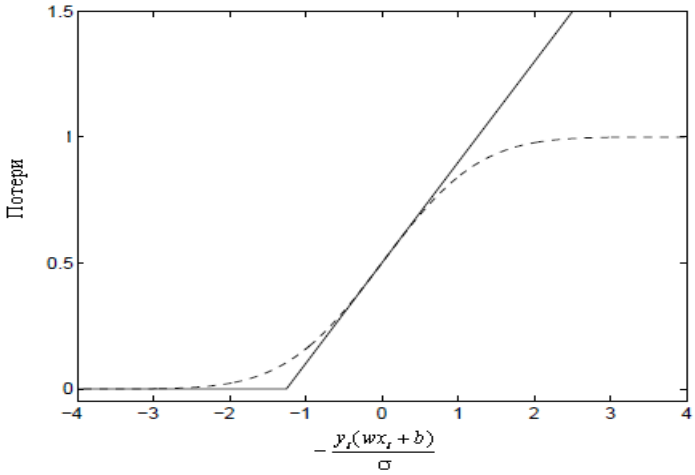

Рис. 2. Функция потерь ЛМР и ее выпуклая аппроксимация

Новой оптимизационной проблемой, соответствующей этой кусочно-линейной функции потерь, является

$$\min \sum_{i=1}^{n} \xi_i \qquad (10)$$

со следующими ограничениями:

$$\xi_i \geq 0,$$

$$\frac{y_i(wx_i+b)}{\sigma} \geq \sqrt{\frac{\pi}{2}} - \xi_i,$$

$$w^2 = 1.$$

С учетом обозначения $A^2 = \frac{2}{\pi\sigma^2}$ эта проблема является эквивалентной минимизации (10) при следующих ограничениях:

$$y_i(wx_i+b) \geq 1-\xi_i, \qquad (11)$$

$$w^2 \leq A^2.$$

Задача может быть решена путем введения множителей Лагранжа [5]:

$$L(w,b,\alpha,\beta,\gamma) = \sum_{i=1}^{n}\xi_i - \frac{1}{2}\gamma(A^2 - w^2) -$$

$$-\sum_{i=1}^{n}\alpha_i[y_i(wx_i+b)-1+\xi_i] - \sum_{i=1}^{n}\beta_i\xi_i,$$

после чего получаем двойственную проблему

$$W(\alpha) = \sum_{i=1}^{n}\alpha_i - A\sqrt{\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j},$$

которая должна быть максимизирована при следующих ограничениях:

$$\sum_{i=1}^{n} y_i\alpha_i = 0,$$

$$0 \leq \alpha_i \leq 1.$$

Однако эта последняя оптимизационная проблема не является квадратичной и более трудной для решения. По этой причине была введена следующая формулировка задачи обобщенной ОРГ [6]:

минимизировать

$$\frac{1}{2}w^2 + C\sum_{i=1}^{n}\xi_i$$

при ограничениях (11) и $\xi_i \geq 0$. Первый член сводится к минимуму для того, чтобы контролировать поле, как и в случае раздельности; целью второго члена является контроль количества неправильно классифицированных точек. Параметр C, увеличение значения которого приводит к пенализации ошибок, выбирается пользователем.

По аналогии с тем, что было сделано для случая раздельности, использование множителей Лагранжа приводит к следующей оптимизационной проблеме:
максимизировать

$$W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j$$

при следующих ограничениях:

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \ \text{ и } \ 0 \leq \alpha_i \leq C.$$

Единственным отличием от случая раздельности является то, что теперь α_i имеет верхнюю границу C.

Заметим, что эти два подхода эквивалентны. Действительно, ограничения (11) являются идентичными в обоих случаях, т.е. как в первом случае выполняется минимизация $\sum\xi_i$ при ограничениях $w^2 \leq A^2$, так и во втором случае выполняется минимизация $\sum\xi_i + \lambda w^2$ (здесь $\lambda = 1/2C$). Первый подход заключается в минимизации эмпирических потерь на ограниченном множестве функций, а второй – в минимизации потерь регуляризации: для каждого значения C существует такое значение A (следовательно, σ), что обе проблемы имеют одинаковые решения, и наоборот.

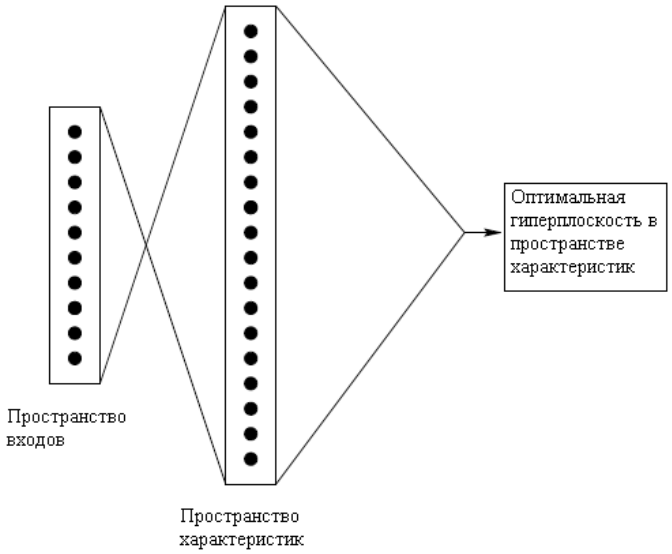

Рис. 3. ОВМ отображает пространство входных данных в многомерное пространство характеристик, а затем строит оптимальную гиперплоскость в пространстве характеристик

Мы показали, что алгоритм слабого поля опорно-векторных машин (ОВМ) может рассматриваться как выпуклая аппроксимация ЛМР со сферической оценкой Гаусса. Соответствующая ширина σ непосредственно связана со слабым параметром поля C. Например, в случае жесткого поля $\sigma \to 0$ и $C \to \infty$.

Нелинейные опорно-векторные машины

Идея опорно-векторных машин заключается в отображении входных данных в многомерное *пространство характеристик* с помощью определенного нелинейного отображения, которое выбирается априорно. Заметим, что в этом пространстве строится оптимальная разделяющая гиперплоскость (рис. 3).

Пример. Для того чтобы построить поверхность решения, отвечающую многочлену степени 2, можно определить следующее пространство характеристик размерности $\frac{d(d+3)}{2}$:

$$z_i = x_i,\ 1 \le i \le d,$$
$$z_{d+i} = (x_i)^2,\ 1 \le i \le d,$$
$$z_{2d+1} = x_1 x_2, ...,\ z_N = x_d x_{d-1},$$

где $x = (x_1, ..., x_d)$ является вектором входных значений, а $z = (z_1, ..., z_N) = \Phi(x)$ – образом x с использованием отображения Φ. Разделяющая гиперплоскость, построенная в пространстве характеристик, является полиномом второй степени в пространстве входных данных.

Возникает одна вычислительная проблема: поскольку размерность пространства характеристик может быть очень большой, встает вопрос, как построить разделяющую гиперплоскость в этом многомерном пространстве.

Ответ на этот вопрос может быть получен из того, что для построения оптимальной разделяющей гиперплоскости в пространстве характеристик отображение $z = \Phi(x)$ может явно не выполняться. Действительно, если заменить x на $\Phi(x)$, уравнение (7) будет иметь вид

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \Phi(x_i)\Phi(x_j).$$

Итак, учебный алгоритм будет зависеть только от данных, используемых в скалярных произведениях в пространстве характеристик, т.е. от функций вида $\Phi(x_i)\Phi(x_j)$. Теперь предположим, что у нас есть такая симметричная функция K, что $K(x_i, x_j) = \Phi(x_i)\Phi(x_j)$. В учебном алгоритме необходимо наличие только K, а отображение Φ явно никогда не используется.

Для отображения Φ ядром K является, очевидно, $K(x, y) = \Phi(x)\Phi(y)$. Однако, при учете ядра Φ, какие существуют условия для неявного отображения? Ответ можно получить из условий Мерсера [2].

Теорема 1. *Пусть $K(x, y)$ является непрерывной симметричной функцией в $L_2(\aleph^2)$. Можно утверждать, что существуют отображение Φ и расширение*

$$K(x, y) = \sum_{i=1}^{\infty} \Phi(x)_i \Phi(y)_i$$

тогда и только тогда, когда для любого компакта C и $g \in L_2(C)$

$$\int_{C\times C} K(x,y)g(x)g(y)dxdy \geq 0. \qquad (12)$$

Заметим, что в определенных случаях довольно сложно проверить, удовлетворяется ли условие Мерсера, поскольку уравнение (12) должно выполняться для любого $g \in L_2(C)$. Тем не менее, легко доказать, что условие удовлетворяется для полиномиального ядра $K(x,y) = (xy + c)^p$, $c \geq 0$ [7].

Рассмотрим пример. Пусть наши входные данные находятся в R^2, а ядро выбирается в виде $K(x,y) = (xy)^2$. В данном случае верным является следующее отображение:

$$\Phi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}.$$

В такой ситуации пространством характеристик является R^3.

После того как ядро K, удовлетворяющее условию Мерсера, является выбранным, учебный алгоритм заключается в максимизации

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j K(x_i, x_j),$$

где функцией решения является

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b).$$

Идея замены скалярного произведения на положительно определенное ядро K называется «трюком ядра». Данная идея была впервые предложена в [8].

Первыми ядрами, которые были применены к исследованию проблемы распознавания образов, были следующие:

полиномиальное ядро $K(x,y) = (xy+1)^p$;

ядро радиальной базисной функции $K(x,y) = \exp(-\|x-y\|^2 / 2\sigma^2)$;

ядро нейронной сети $K(x,y) = \tanh(axy - b)$.

В первом случае следствием является классификатор, который имеет полиномиальную функцию решения. Во втором случае следствием является классификатор гауссовой радиальной базисной функции (РБФ). Наконец, в последнем случае имеет место особый вид двуслойной сигмоидальной сети. В случае РБФ число центров (число опорных векторов), сами центры (опорные векторы), веса (α_i) и порог (b) создаются автоматически посредством обучения ОВМ и дают отличные результаты по сравнению с классической РБФ. Таким же образом для случая нейронных сетей архитектура (число скрытых элементов) определяется обучением ОВМ. Однако ядро гиперболического тангенса удовлетворяет условию Мерсера лишь при некоторых значениях параметров a и b.

Результаты экспериментов

В данном разделе представлены экспериментальные результаты использования опорно-векторных машин на некоторых тестовых базах данных.

Распознавание рукописных цифр часто используется в качестве стандарта для сравнения классификаторов. С учетом этого факта опорно-векторные машины были применены на базе данных USPS [9] и MNIST [10]. Преимущество последней базы данных в том, что имея 60000 учебных примеров и 10000 тестовых примеров, она обеспечивает точное сравнение между классификаторами. С другой стороны, база данных USPS содержит 9298 рукописных цифр (7291 для обучения и 2007 для тестирования) и используется для быстрого сравнения между алгоритмами. Мы использовали эту базу данных несколько раз в ходе проведения экспериментов. Таблица 1 содержит тестовые ошибки различных алгоритмов обучения на базе данных MNIST [11].

Результаты с применением ОВМ были получены в [6]. Было использовано мягкое поле ОВМ с полиномиальным ядром степени 4. Включение предварительных знаний заметно улучшает эффективность, но стандартный алгоритм ОВМ дает лучшие результаты среди классификаторов, которые не принимают предварительные знания в расчет.

Исходя из этого сравнения было выведено, что классификатор оптимального поля обладает отличной точностью потому, что в отличие от других классификаторов высокой производительности, он не включает априорных знаний о проблеме [11].

Таблица 1

Тестовые ошибки на базе данных MNIST

Классификатор	**Тестовая ошибка**
Линейный классификатор	8.4%
Сеть РБФ	3.6%
Нейронная сеть	1.6%
ОВМ	1.1%

Список литературы

1. Bazaraa M., Shetty C.M. Nonlinear programming. New York: John Wiley, 1979.

2. Vapnik V. The Nature of Statistical Learning. Theory. New York: Springer, 1995.

3. Shawe-Taylor J., Bartlett P.L., Wiliamson R.C., Anthony M. Structural risk minimization over data-dependent hierarchies // IEEE Transactions on Information Theory. 1998. V. 44(5). P. 1925–1940.

4. Bartlett P., Shawe-Taylor J. Generalization performance of support vector machines and other pattern classifiers // In: Scholkopf B., Burges C., Smola A., editors. Advances in Kernel Methods – Support Vector Learning. Cambridge, MA: MIT Press, 1999.

5. Vapnik V. Statistical Learning Theory. John Wiley & Sons, 1998.

6. Cortes C., Vapnik V. Support vector network // Machine learning. 1995. V. 20. P. 1–25.

7. Burges C. A tutorial on support vector machines for pattern recognition // Data Mining and Knowledge Discovery. 1998. V. 2(2). P. 121–167.

8. Aizerman M., Braverman E., Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning // Automation and Remote Control. 1964. V. 25. P. 821–837.

9. LeCun Y., Boser B., Denker J.S., et al. Backpropagation applied to handwritten zip code recognition // Neural Computation. 1989. V. 1. P. 541–551.

10. LeCun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition // Proceedings of the IEEE. 1998. V. 86. P. 2278–2324.

11. LeCun Y., Jackel L.J., Bottou L., et al. Comparison of learning algorithm for handwritten digit recognition // In: International Conference on Artificial Neural Networks. 1995. P. 50–53.

FINDING THE OPTIMAL SEPARATING HYPERPLANE BASED ON VICINAL RISK MINIMIZATION

O.A. Galkin

We investigate the methodology of support-vector machines (SVM) based on the vicinal risk minimization principle. The problem of finding the optimal separating hyperplane in the case of linearly inseparable data is considered. An algorithm of the SVM weak field is presented in the form of a convex approximation of linear risk minimization with a spherical Gaussian assessment. The methodology of constructing hyperplanes in a multidimensional space is presented.

Keywords: vicinal risk minimization, separating hyperplane, support vector machine, feature space.