УДК 519.8

Ю.П. ЛАПТИН, Ю.И. ЖУРАВЛЕВ, А.П. ВИНОГРАДОВ

МИНИМИЗАЦИЯ ЭМПИРИЧЕСКОГО РИСКА И ЗАДАЧИ ПОСТРОЕНИЯ ЛИНЕЙНЫХ КЛАССИФИКАТОРОВ[1]

Ключевые слова: *линейные алгоритмы классификации, линейная разделимость множеств, эмпирический риск, метод опорных векторов, методы оптимизации.*

ВВЕДЕНИЕ

Проблемам построения линейных алгоритмов классификации (классификаторов) посвящено большое количество работ [1–5]. Часто такие проблемы рассматриваются для классификации двух множеств. Задачи построения оптимальных линейных классификаторов обычно формулируются для линейно разделимых множеств. Для этого случая такие задачи решаются достаточно эффективно. Понятие оптимальности для двух линейно разделимых множеств имеет простой геометрический смысл — оптимальный классификатор определяет полосу максимальной ширины, разделяющую эти множества.

Для линейной разделимости двух конечных множеств необходимо и достаточно, чтобы выпуклые оболочки этих множеств не пересекались. В случае многих множеств этих условий недостаточно. В [6–8] формулируются некоторые достаточные условия линейной разделимости произвольного числа множеств. Более углубленный геометрический анализ таких условий приведен в [9].

В случае линейно неразделимых множеств естественным критерием выбора классификатора является минимизация эмпирического риска. В настоящей статье рассматриваются частично целочисленная формулировка задачи минимизации эмпирического риска, возможности решения непрерывной релаксации этой задачи. Проводится сравнение предложенной непрерывной релаксации с задачами, которые решаются при использовании других подходов для построения линейных классификаторов в случае линейно неразделимых множеств. Описываются особенности использования методов негладкой оптимизации для решения сформулированных задач.

1. ПОСТАНОВКА ЗАДАЧИ

Пусть задана совокупность линейных функций $f_i(x, W^i) = (w^i, x) + w_0^i$, $i = 1, \dots, m$, где $x \in R^n$ — вектор признаков, $W^i = (w_0^i, w^i) \in R^{n+1}$ — вектор параметров, $i = 1, \dots, m$.

Обозначим $W = (W^1, \dots, W^m)$, $W \in R^L$, $L = m(n+1)$. Будем рассматривать линейные алгоритмы классификации (линейные классификаторы) следующего вида:

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

В [7] рассматривались также классификаторы, в которых f_i — выпуклые кусочно-линейные функции.

Считается заданной совокупность конечных непересекающихся множеств Ω_i, $i = 1, \dots, m$. Будем говорить, что классификатор $a(x, W)$ правильно разделяет точки из Ω_i, $i = 1, \dots, m$, если $a(x, W) = i$, для всех $x \in \Omega_i$, $i = 1, \dots, m$.

[1]Работа выполнена в рамках совместного проекта НАН Украины и Российского фонда фундаментальных исследований № 10-01-90419 «Оптимизационные подходы в задачах машинного обучения и анализа данных».

Множества Ω_i, $i=1,\ldots,m$, называются линейно разделимыми, если существует линейный классификатор, правильно разделяющий точки из этих множеств.

Известно, что два множества линейно разделимы, если выпуклые оболочки этих множеств не пересекаются (см., например, [2]). В случае $m>2$ этого условия недостаточно.

Рассмотрим условия линейной разделимости множеств для произвольного m.

Лемма 1 [7]. Пусть каждое множество Ω_i состоит из одной точки, $i=1,\ldots,m$. Тогда существует линейный классификатор $a(x,W)$, правильно разделяющий точки из Ω_i, $i=1,\ldots,m$ (множества Ω_i, $i=1,\ldots,m$, линейно разделимы).

Теорема 1 [7, 8]. Пусть вокруг каждого множества Ω_i может быть построена сфера S_i, $i=1,\ldots,m$, так, что $S_i \cap S_j = \varnothing$, $i \neq j$. Тогда множества Ω_i, $i=1,\ldots,m$, линейно разделимы.

Доказательство в [7] более простое и получено независимо.

Теорема 2 [7]. Пусть $n \geq C_2^m$, для любых двух множеств: Ω_i, Ω_j, $i \neq j$, существует гиперплоскость, разделяющая эти множества и не пересекающая выпуклые оболочки остальных множеств. Тогда множества Ω_i, $i=1,\ldots,m$, линейно разделимы.

Другие условия линейной разделимости приведены в [9]. Следует отметить, что условия сформулированных теорем достаточности весьма жесткие. Можно привести много примеров, когда эти условия не выполняются, но, тем не менее, существует линейный классификатор, правильно разделяющий точки из Ω_i, $i=1,\ldots,m$.

Каждое множество Ω_i, $i=1,\ldots,m$, есть обучающая выборка точек из некоторого класса $\overline{\Omega}_i$, известного только на элементах выборки. Обучение классификатора $a(x,W)$ заключается в подборе значений параметров W, при котором классы $\overline{\Omega}_i$, $i=1,\ldots,m$, разделяются «наилучшим» образом. Для определения качества разделения существуют различные подходы.

Обозначим $\Omega = \bigcup\limits_{i=1}^{m} \Omega_i$. Пусть точки множества Ω перенумерованы, T — совокупность индексов, $\Omega = \{x^t : t \in T\}$, T_i — совокупность индексов множества Ω_i, $\Omega_i = \{x^t:\ t \in T_i\}$, $T = \bigcup\limits_{i=1}^{m} T_i$. Положим $i(t)$ — номер множества Ω_i, которому принадлежит точка x^t, $t \in T$. Величина

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

называется отступом (margin) или зазором (gap) классификатора $a(x,W)$ на точке x^t, $t \in T$.

Классификатор $a(x,W)$ допускает ошибку на точке x^t тогда и только тогда, когда зазор $g^t(W)$ отрицателен. Величина $g(W) = \min\{g^t(W) : t \in T\}$ называется зазором классификатора $a(x,W)$ на совокупности множеств Ω_i, $i=1,\ldots,m$. Классификатор $a(x,W)$ правильно разделяет точки из Ω_i, $i=1,\ldots,m$, если $g(W)>0$.

Замечание 1. Классификатор $a(x,W)$ инвариантен относительно умножения всех функций f_i (векторов W^i) на положительное число, зазор $g(W)$ линеен относительно такой операции умножения. Классификатор $a(x,W)$ и отступ $g(W)$ инвариантны относительно добавления произвольного числа ко всем f_i.

Величину $g(W)$ можно использовать как критерий качества классификатора $a(x,W)$ (чем больше значение $g(W)$, тем надежнее разделяются точки из Ω_i, $i=1,\ldots,m$), однако при этом должна учитываться некоторая нормировка совокупности векторов W, которую обозначим $\eta(W)$ и назовем нормой классификатора $a(x, W)$.

В дальнейшем будем полагать

$$\eta(W) = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (w_j^i)^2}. \tag{3}$$

В качестве нормы $\eta(W)$ могут использоваться также другие функции [7]. Пусть задана совокупность множеств Ω_i, $i = 1, \ldots, m$. С учетом введенных обозначений задачу построения оптимального классификатора (определения значений параметров W) запишем в следующем виде: найти

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

Поскольку вектор $W = 0$ является допустимым, то задача (4) имеет решение всегда и $g^* \ge g(0) = 0$. Заметим, что $g^* > 0$, если множества Ω_i, $i = 1, \ldots, m$, линейно разделимы, т.е. существует линейный классификатор, правильно разделяющий эти множества. Рассмотрим также задачу: найти

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

Близкие по смыслу задачи рассматривались разными авторами (см., например, [5, 10]).

Лемма 2. Пусть W^* — оптимальное решение задачи (4). Тогда:

1) если $g^* > 0$, то задача (5) также имеет оптимальное решение V^*, $V^* = W^* / g^*$; $\eta^* = 1 / g^*$;

2) если $g^* = 0$, то задача (5) не имеет допустимых решений.

Доказательство простое и приведено в [7].

Рассмотрим более подробно задачи построения линейных классификаторов для заданной совокупности множеств $\Omega_i = \{x^t, t \in T_i\}$, $i = 1, \ldots, m$. Нетрудно видеть, что задача (4) может быть представлена в форме задачи линейного программирования с дополнительным квадратичным ограничением: найти

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

при ограничениях

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

$$\sum_{i=1}^{m} \sum_{j=1}^{n} (w_j^i)^2 \le 1. \tag{8}$$

Соответственно задача (5) есть задача квадратичного программирования: найти

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

при ограничениях

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

Заметим, что в случае $m = 2$ задача (9),(10) эквивалентна задаче, которая используется для построения полосы максимальной ширины, разделяющей линейно разделимые множества Ω_1, Ω_2, и записывается (см., например, [3]) в следующем виде: найти значения параметров $u \in R^n$, $u_0 \in R$ такие, что

$$(u, u) \to \min \tag{11}$$

при ограничениях

$$(u, x^t) - u_0 \geq 1,\ t \in T_1, \tag{12}$$

$$(u, x^t) - u_0 \leq -1,\ t \in T_2. \tag{13}$$

Пусть $\overline{v}^1$, $\overline{v}_0^1$, $\overline{v}^2$, $\overline{v}_0^2$ — оптимальное решение задачи (9), (10), u^*, u_0^* — оптимальное решение задачи (11)–(13). Тогда $\overline{v}^1 = -\overline{v}^2$, $u^* = \overline{v}^1 - \overline{v}^2$, $u_0^* = \overline{v}_0^1 - \overline{v}_0^2$.

Для решения рассмотренных задач может использоваться существующее эффективное программное обеспечение задач оптимизации общего назначения, если число точек в обучающей выборке невелико. Результаты такого рода вычислительных экспериментов приведены в [7]. При большом числе точек в обучающей выборке целесообразно использовать методы негладкой оптимизации [10, 13].

Задачи (6)–(8) и (9),(10) позволяют находить оптимальный линейный классификатор только для линейно разделимых множеств. Для линейно неразделимых множеств задачи должны формулироваться иначе.

Для случая двух множеств: Ω_1, Ω_2, в ряде работ (см., например, [3, 11, 12]) предлагаются некоторые формулировки задач, позволяющие находить линейные классификаторы в случае линейно неразделимых множеств. В других работах (см., например, [5]) для этого предлагаются специальные алгоритмы. Наиболее известным в этой связи является метод опорных векторов (см., например, [3]).

2. МИНИМИЗАЦИЯ ЭМПИРИЧЕСКОГО РИСКА

В случае линейно неразделимой выборки естественным критерием выбора классификатора является минимизация эмпирического риска, т.е. числа точек обучающей выборки, которые классификатор разделяет неправильно.

Будем считать, что задан некоторый параметр $\delta > 0$ надежности разделения точек обучающей выборки Ω_i, $i = 1, \ldots, m$. Будем говорить, что точки x^t, $t \in T$, для которых величина зазора $g^t(W) < \delta$, разделяются классификатором $a(x, W)$ ненадежно.

В дальнейшем значение эмпирического риска будем определять с учетом надежности, определяемой параметром δ — эмпирический риск равен числу точек обучающей выборки, которые классификатор разделяет неправильно или ненадежно.

Лемма 3. Пусть $x^\alpha \in \Omega_i$, $x^\beta \in \Omega_j$, классификатор $a(x, W)$ правильно разделяет эти точки, для нормы классификатора выполняется ограничение (8). Тогда

$$-R \leq w_0^i - w_0^j \leq R, \tag{14}$$

где $R = \sqrt{2} \max \{||x|| : x \in \Omega_i,\ i = 1, \ldots, m\}$.

Доказательство. Для точек x^α, x^β запишем неравенства $-(w^i - w^j, x^\alpha) \leq$ $\leq w_0^i - w_0^j \leq -(w^i - w^j, x^\beta)$. Учитывая (8), можно показать, что $-\sqrt{2}||x^\alpha|| \leq$ $\leq w_0^i - w_0^j \leq \sqrt{2}||x^\beta||$. Отсюда следует утверждение леммы.

Пусть $\Omega_i = \{x^t, t \in T_i\}$, $i = 1, \ldots, m$, $T = \bigcup_{i=1}^m T_i$. Каждой точке x^t, $t \in T$, поставим в соответствие переменную $y_t = 0 \vee 1$ так, что $y_t = 0$, если точка x^t учитывается при формировании задачи (6)–(8), $y_t = 1$ в противном случае.

Пусть задано большое положительное число B. Задача минимизации эмпирического риска с учетом надежности, определяемой параметром $\overline{\delta}$, имеет вид: найти

$$Q^* = \min_{w, y} \{\sum_{t \in T} y_t\} \tag{15}$$

при ограничениях

$$(w^i - w^j, x^t) + w_0^i - w_0^j \geq \bar{\delta} - B \cdot y_t, \;\; j \in \{1, \ldots, m\} \setminus i, \;\; t \in T_i, \;\; i = 1, \ldots, m, \tag{16}$$

$$\eta(W) \leq 1, \tag{17}$$

$$\sum_{t \in T_i} y_t \leq |T_i| - 1, \;\; i = 1, \ldots, m, \tag{18}$$

$$0 \leq y_t \leq 1, \; t \in T, \tag{19}$$

$$y_t = 0 \vee 1, \; t \in T. \tag{20}$$

Из ограничений (14), (17) следует, что если $y_t = 1$, то при достаточно большом значении числа B соответствующие неравенства вида (16) выполняются всегда, т.е. точка x^t исключается из задачи. Ограничения (18) определяют условие того, что, по крайней мере, одна точка из каждого множества Ω_i должна быть включена в задачу.

Оптимальное значение Q^* равно минимальному эмпирическому риску с учетом надежности $\bar{\delta}$. При формулировке задачи минимизации эмпирического риска могут учитываться дополнительные условия и информация:

- цена ошибки на разных точках может быть разной, т.е. целевой функцией может быть взвешенная сумма ошибок,
- на допустимую область значений векторов W могут накладываться дополнительные ограничения.

Задача (15)–(20) является *NP*-трудной, ее приближенные решения могут быть найдены в рамках общей схемы метода ветвей и границ. Для вычисления оценки снизу для величины Q^* (минимального эмпирического риска) рассмотрим непрерывную релаксацию сформулированной задачи — задачу (15)–(19) Оптимальное значение релаксированной задачи обозначим q^*. Для ее решения используем схему декомпозиции по переменным W. Пусть переменные W зафиксированы. Учитывая (2), задачу минимизации по переменным y представим в следующем виде: найти

$$q(W) = \min_y \left\{ \sum_{t \in T} y_t \right\} \tag{21}$$

при ограничениях

$$y_t \geq \frac{1}{B} (\bar{\delta} - g^t(W)), \; t \in T, \tag{22}$$

$$\eta(W) \leq 1, \tag{23}$$

$$\sum_{t \in T_i} y_t \leq |T_i| - 1, \;\; i = 1, \ldots, m, \tag{24}$$

$$0 \leq y_t \leq 1, \; t \in T. \tag{25}$$

Обозначим $d^t(W) = \max\left(0, \frac{1}{B}(\bar{\delta} - g^t(W))\right)$. Очевидно, что если задача (21)–(25) имеет решение y^*, то $y_t^* = d^t(W)$. Отсюда получаем задачу минимизации по переменным W: найти

$$q^* = \min \sum_{t \in T} d^t(W) \tag{26}$$

при ограничениях

$$\eta(W) \leq 1, \tag{27}$$

$$\sum_{t \in T_i} d^t(W) \leq |T_i| - 1, \; i = 1, \ldots, m, \tag{28}$$

$$d^t(W) \leq 1, \; t \in T. \tag{29}$$

Функции $d^t(W)$ выпуклые кусочно-линейные, $\eta(W)$ — квадратичная положительно определенная. Для решения задачи (26)–(29) целесообразно применять эффективные методы негладкой оптимизации [13]. Особенности такого применения рассматриваются в разд. 4.

3. СРАВНИТЕЛЬНЫЙ АНАЛИЗ

Предложенный выше подход сравнивается с методом опорных векторов (см., например, [3]) и методом робастного разделения [10–12] для случая двух линейно неразделимых множеств. Пусть, как и ранее, $\Omega_i = \{x^t, t \in T_i\}$, $i = 1, 2$, $T = T_1 \cup T_2$.

В методе опорных векторов решается задача, которая может быть представлена в следующем виде: найти

$$\min_{u, u_0, \xi} \left[\frac{1}{2}(u, u) + C \sum_{t \in T} \xi^t \right] \tag{30}$$

при ограничениях

$$(u, x^t) + u_0 \geq 1 - \xi^t, \ t \in T_1, \tag{31}$$

$$(-u, x^t) - u_0 \geq 1 - \xi^t, \ t \in T_2, \tag{32}$$

$$\xi^t \geq 0, \ t \in T, \tag{33}$$

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

При робастном разделении двух множеств [10–12] построение классификатора выполняется в два этапа. На первом решается задача: найти

$$\min_{u, u_0, \xi} \left\{ \frac{1}{|T_1|} \sum_{t \in T_1} \xi^t + \frac{1}{|T_2|} \sum_{t \in T_2} \xi^t \right\} \tag{34}$$

при ограничениях (31)–(33). На втором этапе при фиксированных значениях вектора u уточняется значение параметра u_0.

Для сравнения различных подходов представим аналог задачи (15)–(20) для случая двух множеств в следующем виде:

$$Q^* = \min_{w, w_0, y} \left\{ \sum_{t \in T} y_t \right\} \tag{35}$$

при ограничениях

$$(w, x^t) + w_0 \geq \bar{\delta} - B \cdot y_t, \ \ t \in T_1, \tag{36}$$

$$(-w, x^t) - w_0 \geq \bar{\delta} - B \cdot y_t, \ \ t \in T_2, \tag{37}$$

$$(w, w) \leq 1, \tag{38}$$

$$\sum_{t \in T_i} y_t \leq |T_i| - 1, \ \ i = 1, 2, \tag{39}$$

$$0 \leq y_t \leq 1, \ t \in T, \tag{40}$$

$$y_t = 0 \vee 1, \ t \in T. \tag{41}$$

Сделаем замену переменных в задаче (35)–(41): $w=\overline{\delta}u,\ w_0=\overline{\delta}u_0$, $\xi^t=\dfrac{By_t}{\overline{\delta}},\ t\in T_1\cup T_2$. Тогда непрерывная релаксация этой задачи примет вид

$$q^*=\frac{\overline{\delta}}{B}\cdot\min_{u,u_0,\xi}\left\{\sum_{t\in T}\xi^t\right\} \tag{42}$$

при ограничениях

$$(u,x^t)+u_0\ge 1-\xi^t,\ \ t\in T_1, \tag{43}$$

$$(-u,x^t)-u_0\ge 1-\xi^t,\ \ t\in T_2, \tag{44}$$

$$(u,u)\le 1/\overline{\delta}^2, \tag{45}$$

$$\xi^t\ge 0,\ t\in T, \tag{46}$$

$$\xi^t\le\frac{B}{\overline{\delta}},\ t\in T, \tag{47}$$

$$\sum_{t\in T_i}\xi^t\le\frac{B}{\overline{\delta}}(|T_i|-1),\ \ i=1,2. \tag{48}$$

Обозначим $\chi,\gamma_i,\ i=1,2$, двойственные переменные для ограничений (45), (48) и рассмотрим функцию Лагранжа

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

Положим

$$\varphi(\chi,\gamma)=\min_{u,u_0,\xi}L(\chi,\gamma,\xi,u) \tag{49}$$

при ограничениях (43), (44), (46), (47).

Пусть задан штрафной коэффициент C в задаче (30)–(33). Нетрудно видеть, что, полагая $\gamma=0$ и выбирая χ из условия $\dfrac{\overline{\delta}}{2\chi B}=C$, получаем

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

т.е. задача (49), (43), (44), (46) эквивалентна задаче (30)–(33) при указанном выборе значений двойственных переменных.

Полагая $\chi=0$, получаем $L(\chi,\gamma,\xi,u)=\sum_{i=1}^{2}\left\{\left(\dfrac{\overline{\delta}}{B}+\gamma_i\right)\sum_{t\in T_i}\xi^t\right\}-\sum_{i=1}^{2}\gamma_i\left(\dfrac{B}{\overline{\delta}}(|T_i|-1)\right)$.

Откуда следует, что, подбирая соответствующим образом значения γ, получаем задачу (49), (43), (44), (46), эквивалентную задаче, которая решается при робастном разделении двух множеств.

Таким образом, задачи, которые решаются в методе опорных векторов и при робастном разделении двух множеств, являются частным случаем задачи (49), (43), (44), (46).

4. АЛГОРИТМЫ РЕШЕНИЯ РЕЛАКСИРОВАННОЙ ЗАДАЧИ

Задача (26)–(29) вычисления нижней оценки минимального эмпирического риска является частным случаем задачи выпуклого программирования

$$f^*=\min f(x) \tag{50}$$

при ограничениях

$$f_i(x)\le 0,\ \ i=1,\ldots,m, \tag{51}$$

где $x\in R^n$.

Для решения таких задач в настоящее время известны различные алгоритмы [13]. Одним из подходов является метод негладких штрафных функций, в котором задача (50), (51) сводится к эквивалентной задаче безусловной выпуклой оптимизации с целевой функцией, имеющей вид $f(x)+\lambda h^{+}(x)$, где $\lambda \geq 0$, $h^{+}(x)=\max(0, h(x))$, $h(x)=\max\{f_i(x), i=1,\ldots,m\}$.

Известно (см., например, [13]), что если для задачи (50), (51) выполняется условие Слейтера, то при достаточно большом значении коэффициента λ решение задачи

$$\bar{f}^{*}=\min_{x}\{f(x)+\lambda h^{+}(x)\} \tag{52}$$

совпадает с решением задачи (50), (51). Для поиска решения задачи (52) могут применяться эффективные методы негладкой оптимизации [13]. При таком подходе требуется определенным образом подбирать коэффициент λ.

Рассмотрим другой подход [14, 15] к сведению задач (50), (51) с ограничениями к задачам безусловной оптимизации. Обозначим $S=\{x\in R^{n}: h(x)\leq 0\}$. Предположим, что S — замкнутое выпуклое множество, задана допустимая точка $x^{0}\in S$ такая, что $h(x^{0})<0$. Для $x\notin S$ обозначим $\pi_S(x)$ точку пересечения отрезка $[x^{0}, x]$ с границей множества S. Одномерный поиск для определения $\pi_S(x)$ может быть реализован достаточно эффективно.

Пусть задано некоторое число E, $E<f(x^{0})$, которое будем называть параметром продолжения целевой функции. Положим

$$\chi^{E}(x)=E+(f(\pi_S(x))-E)\frac{||x-x^{0}||}{||\pi_S(x)-x^{0}||}, \tag{53}$$

$$\psi^{E}(x)=\begin{cases} f(x), & \text{если } x\in S,\\ \chi^{E}(x), & \text{если } x\notin S.\end{cases} \tag{54}$$

Рассмотрим задачу

$$\psi^{*E}=\inf\{\psi^{E}(x): x\in R^{n}\}. \tag{55}$$

Имеют место следующие утверждения.

Лемма 4 [15]. Пусть $E\leq f^{*}$, тогда $\psi^{*E}=f^{*}$ и решения задач (55) и (50), (51) совпадают.

Теорема 3 [15]. Пусть f, h — выпуклые функции, принимающие конечные значения при любых $x\in R^{n}$. Тогда существует конечное E^{*} и для всех $E<E^{*}$ функция $\psi^{E}(x)$ выпуклая.

В работе [15] также приводятся соотношения, позволяющие вычислять субградиент функции $\chi^{E}(x)$ в произвольной точке $x\notin S$.

Если величина E удовлетворяет условиям $E\leq f^{*}$ и $E<E^{*}$, то для решения задачи (55) может применяться любой сходящийся алгоритм минимизации выпуклых функций.

При неизвестных f^{*} и E^{*} значение E должно уточняться итеративно по ходу работы алгоритма минимизации. В [15] приведены простые условия, которые должны проверяться на каждой итерации. Уточнение значения E проводится, если на текущей итерации эти условия не выполняются. При этом если уточнение (уменьшение) значения E проводится на величину $\Delta\geq\bar{\Delta}>0$, где $\bar{\Delta}$ — заданный параметр, то такие уточнения в процессе работы алгоритма будут проводиться конечное число раз. После чего алгоритм сходится к решению задачи (50)–(51).

Результаты вычислительных экспериментов [15] показали эффективность метода выпуклых продолжений. При этом, однако, каждое обращение к функции

$\psi^E(x)$ требует выполнения достаточно трудоемкой (по сравнению с вычислением значений функций f и h) процедуры одномерного поиска.

Для задачи (26)–(29) с учетом ее специфики процедура одномерного поиска точки на границе допустимого множества может быть реализована существенно эффективнее обычной процедуры дихотомии.

Представим вектор в виде $W = W^0 + p\tau$, где W^0 — заданная допустимая точка, p — направление поиска, и подставим выражение для W в (27)–(29). Значение τ, при котором (27) превращается в равенство, находится как решение квадратного уравнения. Рассмотрим процедуру определения значения τ, при котором точка $W = W^0 + p\tau$ лежит на границе множества, описываемого неравенствами (28), (29). Представим последние в виде

$$f^m(\tau) \equiv \sum_{k \in K_m} f_{km}(\tau) \le 0,\ m \in M, \tag{56}$$

где $\tau \in R$, $f_{km}(\tau)$ — выпуклые кусочно-линейные функции,

$$f_{km}(\tau) = \max\{a^l \tau + b^l : l \in L_{km}\}. \tag{57}$$

Заметим, что скалярные произведения для определения коэффициентов каждой линейной функции $a^l\tau + b^l$ вычисляются только один раз.

Будем считать, что $L_{km} \cap L_{k'm'} = \varnothing$, если $km \ne k'm'$. Для каждой функции $f_{km}(\tau)$ разобьем интервал $(-\infty, \infty)$ на отрезки $(\underline{\tau}_l, \overline{\tau}_l)$ так, что при $\tau \in (\underline{\tau}_l, \overline{\tau}_l)$ максимум в (57) достигается на линейной функции $a^l\tau + b^l$, $l \in L_{km}$. Для этого достаточно упорядочить элементы множества L_{km} по возрастанию коэффициентов a^l (трудоемкость $O(|L_{km}| \cdot \log_2 |L_{km}|)$) и вычислить значения $\underline{\tau}_l, \overline{\tau}_l$ (трудоемкость $O(|L_{km}|)$).

Обозначим $L = \cup\{L_{km} : k \in K_m, m \in M\}$, $\underline{\mathrm{T}} = \{\underline{\tau}_l, l \in L\}$, $\overline{\mathrm{T}} = \{\overline{\tau}_l, l \in L\}$. Упорядочим множества $\underline{\mathrm{T}}$, $\overline{\mathrm{T}}$ по возрастанию элементов (трудоемкость $O(|L| \cdot \log_2 |L|)$).

Для поиска граничной точки π используем итеративную процедуру, на каждом шаге n которой

1) выбирается (правило выбора объясняется ниже) точка τ_n;

2) уточняются множества M и L_{km}:

- если $\max\{f^m(\tau_n) : m \in M\} < 0$, т.е. $\pi > \tau_n$, то из множества L_{km} исключаются все элементы l такие, что $\overline{\tau} < \tau_n$;
- если $\max\{f^m(\tau_n) : m \in M\} > 0$, т.е. $\pi < \tau_n$, то из множества M исключаются все элементы m такие, что $f^m(\tau_n) < 0$, из множества L_{km} исключаются все элементы l такие, что $\underline{\tau}_l > \tau_n$;

3) если каждое множество L_{km} состоит из одного элемента, то задача сведена к системе линейных неравенств, точка π находится просто, процедура завершается; в противном случае выполняется переход на п. 1).

Выбор точки τ_n выполняется из условия $|\{\underline{\tau}_l : \underline{\tau}_l > \tau_n, l \in L\}| = |\{\overline{\tau}_l : \overline{\tau}_l < \tau_n, l \in L\}|$. Трудоемкость предложенной процедуры имеет порядок $O(|L| \cdot \log_2 |L|)$.

ЗАКЛЮЧЕНИЕ

В работе рассмотрены задачи построения линейных классификаторов для случая многих классов. Для линейно неразделимых множеств сформулирована частично целочисленная задача минимизации эмпирического риска. Показано, что непрерывная релаксация сформулированной задачи обобщает известные подходы к построению линейных классификаторов для линейно неразделимых множеств. Рассмотрены вопросы использования методов негладкой оптимизации для решения задачи вычисления нижней оценки минимального эмпирического риска. Показано, что специфика задачи позволяет повысить эффективность методов, используемых для решения выпуклых задач оптимизации с ограничениями.

СПИСОК ЛИТЕРАТУРЫ

1. Zhuravlev Yu.I. An algebraic approach to recognition or classification problems // Pattern Recognition and Image Analysis. — 1998. — **8**, N 1. — P. 59–100.
2. Местецкий Л.М. Математические методы распознавания образов. — http://www.intuit.ru/department/graphics/imageproc/.
3. Воронцов К.В. Машинное обучение. — http://www.machinelearning.ru/wiki/images/6/68/voron-ML-Lin.pdf.
4. Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания. — Киев: Наук. думка, 2008. — 232 с.
5. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наук. думка, 2004. — 545 с.
6. Laptin Yu., Vinogradov A. Exact discriminant function design using some optimization techniques // Classification, Forecasting, Data Mining International. Book Series "INFORMATION SCIENCE & COMPUTING", N 8, Sofia (Bulgaria), 2009. — P. 14–19.
7. Laptin Yu.P., Likhovid A.P., and Vinogradov A.P. Approaches to construction of linear classifiers in the case of many classes // Pattern Recognition and Image Analysis. — 2010. — **20**, N 2. — P. 137–145.
8. Петунин Ю.И, Шульдешов Г.А. Проблемы распознавания образов с помощью линейных дискриминантных функций Фишера // Кибернетика. — 1979. — № 6. — С. 134–137.
9. Рублев Б.В., Петунин Ю.И., Литвинко П.Г. Структура гомотетичных линейно разделимых множеств в *n*-мерном евклидовом пространстве. Ч. 1, 2 // Кибернетика и системный анализ. Ч. 1. — 1992. — № 1. — С. 3–15; Ч. 2. — 1992. — № 2. — С. 23–33.
10. Методи негладкої оптимізації у спеціальних задачах класифікації / П.І. Стецюк, О.А. Березовський, М.Г. Журбенко, Д.О. Кропотов. — Київ, 2009. — 28 с. — (Препр. / НАН України. Ін-т кібернетики ім. В.М. Глушкова; 2009–1).
11. Bennett K.P., Mangasarian O.L. Robust linear programming discrimination of two linearly inseparable sets // Optimiz. Methods and Software. — 1992. — N 5. — P. 23–34.
12. Журбенко Н.Г., Саимбетов Д.Х. К численному решению одного класса задач робастного разделения двух множеств // Методы исследования экстремальных задач. — Киев: Ин-т кибернетики им. В.М. Глушкова НАН Украины, 1994. — С. 52–55.
13. Shor N.Z. Nondifferentiable optimization and polynomial problems. — Dordrecht: Kluwer, 1998. — 394 p.
14. Лаптин Ю.П. Один подход к решению нелинейных задач оптимизации с ограничениями // Кибернетика и системный анализ. — 2009. — № 3. — С. 182–187.
15. Лаптин Ю.П., Лиховид А.П. Использование выпуклых продолжений функций для решения нелинейных задач оптимизации // Управляющие системы и машины. — 2010. — № 6. — С. 25–31.

Поступила 14.07.2010