ТЕОРІЯ ОПТИМАЛЬНИХ РІШЕНЬ

УДК 519.8

Н.Г. ЖУРБЕНКО

ПОСТРОЕНИЕ ЛИНЕЙНЫХ КЛАССИФИКАТОРОВ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ ОПОРНЫХ ФУНКЦИЙ

Предложен алгоритм построения линейных бинарных классификаторов. Объекты распознавания представляются выпуклыми компактами евклидового пространства. Алгоритм основан на использовании опорных функций выпуклых компактов и методов негладкой оптимизации.

Введение. Обычно объект распознавания для линейного классификатора представляется точкой в n-мерном евклидовом пространстве R^n. Однако возникают задачи, когда объект распознавания представляется более сложным образом. Это может быть связано с погрешностями измерения параметров объекта или для параметра объекта задается диапазон его значений. Последнее характерно для задач медицинской диагностики. В работе [1] объект распознавания определялся эллипсоидом в R^n. В данной работе предполагается, что объект распознавания определялся выпуклым компактом в R^n. В работе представляется описание алгоритма построения линейных бинарных классификаторов для такого достаточно широкого класса задач. Приводится также краткое описание разработанного программного обеспечения.

Линейный классификатор и опорная функция. Линейный бинарный классификатор $L(b,c)$ задается ненулевым вектором $b \in R^n, b \neq 0$ и константой $c \in R^1$. Классификатор $L(b,c)$ определяет открытые полупространства в R^n:

$$R^+ = \{x \in R^n \mid (x,b) + c > 0\};$$

$$R^- = \{x \in R^n \mid (x,b) + c < 0\}.$$

Объект распознавания представляется выпуклым компактом W в R^n. Объект распознается классификатором, если $W \subset R^-$ («объект класса 1») или $W \subset R^+$ («объект класса 2»). В противном случае говорим, что

объект W не распознается классификатором $L(b,c)$. Объект распознается классификатором, если $W \subset R^-$ («объект класса 1») или $W \subset R^+$ («объект класса 2»). В противном случае говорим, что объект W не распознается классификатором $L(b,c)$.

Пусть $h(u)$ опорная функция [2], выпуклого компакта W:

$$h(u) = \max\{(x,u) \mid x \in W\}. \quad (1)$$

$h(u)$ – выпуклая функция, $h(u)$ – положительно однородна и субаддитивна на R^n: $h(\lambda u) = \lambda h(u)$ при $\lambda \geq 0$; $h(u+v) \leq h(u) + h(v)$.

Пусть $x(u)$ – некоторое (возможно не единственное) решение задачи (1): $h(u) = (x(u),u)$. Из задачи (1) следует, что $x(u)$ – субградиент опорной функции $h(u)$. Таким образом $x(u)$ определяет не только значение опорной функции, но и ее субградиент в точке u.

Приведем примеры опорной функции.

1. W – точка в $R^n : W = \{z \in R^n\}$. $x(u) = z$; $h(u) = (z,u)$.

2. W – эллипсоид в $R^n : W = \{x \in R^n \mid (x-z, D(x-z) \leq r^2\}$, где $z \in R^n$ – центр эллипсоида; D – положительно определенная матрица $n \times n$; $r \geq 0$.

$$x(u) = z + \begin{cases} \dfrac{r}{\sqrt{(u, D^{-1}u}} D^{-1}u, & u \neq 0; \\ 0, & u = 0. \end{cases}$$

$$h(u) = (z,u) + r\sqrt{(u, D^{-1}u}.$$

Нетрудно показать справедливость следующих отношений эквивалентности:

$$W \subset R^+ \leftrightarrow (x(-b), -b) - c < 0; \quad (2)$$

$$W \subset R^- \leftrightarrow (x(b), b) + c < 0. \quad (3)$$

Из $(x(-b), b) = -(x(-b), -b) = -h(-b)$ следует:

$$W \subset R^+ \leftrightarrow h(-b) - c < 0; \quad (4)$$

$$W \subset R^- \leftrightarrow h(b) + c < 0. \quad (5)$$

Отметим также следующее свойство опорных функций ([2], теорема 12.4).

Пусть W_i – непустые выпуклые компакты в R^n с опорными функциями h_i, $i = 1,\ldots m$. $W = conv\left(\bigcup_{i=1}^{m} W_i\right)$ (W – выпуклая оболочка множеств W_i). Тогда $h(u) = \max\limits_{1 \leq i \leq m} h_i(u)$ – опорная функция компакта W.

Заметим, что отсюда следует достаточно очевидный факт: если линейный классификатор разделяет два множества компактов I_1, I_2, то этот классификатор разделяет и множества $conv(I_1)$, $conv(I_2)$.

Задача определения классификатора $L(b,c)$ состоит в следующем. Задана «обучающая выборка»: множество I_1 ($m_1 = |I_1|$) объектов класса 1; множество I_2 ($m_2 = |I_2|$) объектов класса 2. Необходимо определить классификатор $L(b,c)$ «правильно» определяющий класс этих объектов (в этом случае плоскость $\{x \in R^n \mid (x,b)+c=0\}$ «разделяет» множества I_1, I_2). То есть необходимо определить классификатор $L(b,c)$, для которого справедливы соотношения:

$$h_i(-b)-c<0,\ i\in I_1; \qquad h_i(b)+c<0,\ i\in I_2. \tag{6}$$

Система (6) эквивалентна системе

$$\max\{h_i(-b)-c \mid i\in I_1\}<0; \quad \max\{h_i(b)+c \mid i\in I_2\}<0.$$

Таким образом, формально определение классификатора $L(b,c)$ сводится к решению системы (6). Однако решение этой системы может быть не единственно или система несовместна. Поэтому возникает достаточно сложная проблема определения «хорошего» классификатора как для случая совместности, так и несовместности (6). Этой проблеме посвящено множество публикаций [3]. Разумеется, предлагаемый далее алгоритм построения классификатора не претендует на решение этой проблемы.

Алгоритм определения классификатора. В данном пункте приводится краткое описание предлагаемого алгоритма определения классификатора. Фактически этот алгоритм является обобщением алгоритма работы [1] на случай представления объектов распознавания выпуклыми компактами R^n.

В работе [4] был предложен робастный алгоритм решения задачи построения разделяющей гиперплоскости для двух множеств точек. В работе [5] для ее решения использовался алгоритм негладкой оптимизации – *r*-алгоритм [6]. По аналогии работы [1] построим двухэтапный алгоритм для рассматриваемой задачи, в которой объектами выборок являются компакты.

На первом этапе определяются параметры b^* и c^* путем решения задачи минимизации выпуклой негладкой функции $F(b,c)$:

$$\min_{b,c}\left[F(b,c)\equiv\frac{1}{m_1}\sum_{i\in I_1}\max\{0;h_i(-b)-c+1\}+\frac{1}{m_2}\sum_{i\in I_2}\max\{0;h_i(b)+c+1\}\right]. \tag{7}$$

На втором этапе проводится процедура улучшения разделяющей гиперплоскости по переменной c. Эта процедура состоит в решении следующей задачи минимизации выпуклой одномерной кусочно-линейной выпуклой функции:

$$\min_{c}\left[\frac{1}{m_1}\sum_{i\in I_1}\max\{0;h_i(-b^*)-c\}+\frac{1}{m_2}\sum_{i\in I_2}\max\{0;h_i(b^*)+c\}\right].$$

Основная численная трудоемкость описанной модели состоит в решении задачи первого этапа (6).

Модель (7) в основном отражает интуитивные представления разделения двух классов объектов. Так, например, если классы разделимы, то так определяемый классификатор также их разделяет. При этом значение целевой функции в задаче (7) равно нулю.

Недостатком модели (7) является то, что она не обеспечивает построение разделяющей полосы максимальной ширины. Поэтому решение задачи проводится по следующей схеме. Если в результате решения задачи (7) определено, что классы разделимы, то решается следующая задача (задача определения классификатора с максимальным «зазором» [1]):

$$F(b^*,c^*)=\min_{b,c}\left\{\max\left\{h_i(-b)-c, i\in I_1; h_i(b)+c, i\in I_2;\right\} \mid \sum_{i=1}^{n} b_i^2 \le 1\right\}. \qquad (8)$$

Алгоритмы решения задач (7), (8) основаны на использовании методов негладкой оптимизации.

Замечание. Если классификатор правильно распознает не все объекты обучающей выборки (ситуация неразделимости множеств I_1, I_2), то это не означает его непригодность. Это может быть связано с ошибками самой обучающей выборки. Однако более реалистично, что поставленная задача классификации не может быть точно формализована в рамках принятой модели. Качество линейного классификатора может определяться лишь на основе статистического анализа результатов для представительного множества обучающих выборок.

Программное обеспечение. Программная реализация описанных алгоритмов представлена классами в объектно-ориентированном стиле на языке C++. Разработанные классы отображают концепции объектов рассматриваемых задач (линейный классификатор, объект распознавания, обучающая выборка). Программное обеспечение может использоваться для построения линейного классификатора любых объектов, определяемых выпуклыми компактами в R^n. Для этого пользователю достаточно определить опорную функцию распознавательных объектов его задачи. Опорная функция объекта определяется функцией-членом определяемого пользователем класса объекта. Эта функция объявлена как чисто виртуальная функция абстрактного класса **CObjectRecognition** (класс **CObjectRecognition** реализует концепцию «объект распознавания»).

Кроме абстрактных классов программное обеспечение содержит конкретные классы (порожденные от соответствующих абстрактных классов). В частности, имеются классы представления таких объектов распознавания как точка, шар, эллипс.

Программная реализация решения задач (7), (8) основана на модификациях *r*-алгоритма [7].

М.Г. Журбенко

ПОБУДОВА ЛІНІЙНИХ КЛАСИФІКАТОРІВ НА ОСНОВІ ВИКОРИСТАННЯ ОПОРНИХ ФУНКЦІЙ

Запропоновано алгоритм побудови лінійних бінарних класифікаторів. Об'єкти розпізнавання представляються опуклими компактами евклідового простору. Алгоритм заснований на використанні опорних функцій опуклих компактів і методів негладкої оптимізації.

N.G. Zhurbenko

THE CONSTRUCTION OF LINEAR CLASSIFIERS ON THE BASIS OF SUPPORT FUNCTIONS

An algorithm for construction of linear binary classifiers is proposed. The objects of recognition are presented by convex compacts of Euclidean space. The algorithm is based on the use of support functions of convex compacts and nonsmooth optimization methods.

1. *Березовский О.А., Журбенко Н.Г., Стецюк П.И.* Алгоритмы построения линейных бинарных классификаторов при неточных измерениях // Компьютерная математика. – Киев: Ин-т кибернетики имени В.М. Глушкова НАН Украины, 2014. – № 2. – С. 133 – 138.
2. *Лейхтвейс К.* Выпуклые множества. – М.: Наука, 1985. – 336 с.
3. *Шлезингер М., Главач В.* Десять лекций по статистическому и структурному распознаванию. – Киев: Наук. думка, 2004. – 545 с.
4. *Bennett K.P., Mangasarian O.L.* Robust Linear Programming Discrimination of Two Linearly Inseparable Sets // Optimization Methods and Software. – 1992. – **5** (1). – P. 23 – 34.
5. *Журбенко Н.Г., Саимбетов Д.Х.* К численному решению одного класса задач робастного разделения двух множеств // Методы исследования экстремальных задач. – Киев: Ин-т кибернетики имени В.М. Глушкова НАН Украины, 1994. – С. 52 – 55.
6. *Шор Н.З., Журбенко Н.Г.* Метод минимизации, использующий операцию растяжения пространства в направлении разности двух последовательных градиентов // Кибернетика. – 1971. – № 3. – С. 51 – 59.
7. *Шор Н.З., Журбенко Н.Г., Лиховид А.П. и др.* Развитие алгоритмов недифференцируемой оптимизации и их приложения // Кибернетика и системный анализ. – 2003. – № 4. – С. 80 – 94.

Получено 19.04.2016