ПОСТРОЕНИЕ НЕЛИНЕЙНЫХ КЛАССИФИКАТОРОВ В СЛУЧАЕ МНОГИХ КЛАССОВ

Ю.И. Журавлев, Ю. П. Лаптин, А.П. Виноградов

Abstract: *Рассматриваются задачи построения нелинейных классификаторов для случая многих классов. Для произвольной обучающей выборки формулируется NP-трудная задача построения классификатора заданной сложности, обеспечивающего минимизацию эмпирического риска. Рассмотрена непрерывная релаксация задачи минимизации эмпирического риска, предложены некоторые новые подходы к решению таких задач*

Keywords: *кластер, решающее правило, дискриминантная функция, линейное и нелинейное программирование, негладкая оптимизация*

ACM Classification Keywords: *G.1.6 Оптимизация – градиентные методы, I.5 Распознавание образов; I.5.2 Методология проектирования – построение и оценивание классификаторов*

Acknowledgement: *Работа поддержана совместным грантом НАН Украины и РФФИ № 10-01-90419*

Введение

Проблемы построения классификаторов, линейных и нелинейных, обычно рассматриваются для двух непересекающихся множеств. Для линейно разделимых множеств естественным образом формулируются и эффективно решаются задачи построения оптимальных классификаторов [Мест, 2008], [В, 2009], [ШГ, 2004], [Маз, 2010], [DN, 2009]. Для линейно неразделимых множеств, как формулировка, так и решение задач построения классификаторов наталкиваются на определенные трудности. При построении линейных классификаторов в этом случае часто используются эвристические алгоритмы [В, 2009], [ВМ, 1992], [СБЖК, 2009]. Широко используется метод опорных векторов, требующий задания значений некоторых параметров. Выбор этих значений далеко не очевиден. Альтернативой для линейно неразделимых множеств является переход к нелинейным классификаторам – квадратичным, полиномиальным. По-видимому, увеличивая сложность классификатора, всегда можно добиться правильного разделения исходных множеств. Такая цель не всегда оправданна, поскольку исходные множества (обучающие выборки) могут формироваться в условиях возможного возникновения тех, или других ошибок. Вопросы постановки и решения задач построения классификаторов в условиях неопределенности рассматриваются в ([BTGN, 2009], Chapter 12). Естественным в этой ситуации является построение классификатора заданного типа (заданной сложности), минимизирующего число точек обучающей выборки, разделяемых неправильно (минимизация эмпирического риска).

Замечательной особенностью рассматриваемых задач является то, что для достаточно широкого класса нелинейных классификаторов условие правильного разделения множеств классификатором эквивалентно тому, что параметры этого классификатора удовлетворяют некоторой системе линейных неравенств [Маз, 2010]. Эта особенность позволяет создавать эффективные алгоритмы формирования таких нелинейных классификаторов.

В настоящей работе задачи построения нелинейных классификаторов рассматриваются для случая многих классов. Эта работа является развитием результатов, полученных ранее для линейных классификаторов [LLV, 2010], [ZhLV, 2010].

1. Постановка задачи

Для заданной совокупности конечных непересекающихся множеств $\Omega_i \subset R^n$, $i=1,..., m$ (обучающей выборки из m непересекающихся классов) рассматривается задача построения алгоритма классификации следующего вида

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

где W^i – совокупность параметров функции $f_i(x,W^i)$, $i=1,...,m$, $W=(W^1,..., W^m)$. Функции $f_i(x,W^i)$ называются (приоритетными) весовыми функциями.

Говорят, что классификатор $a(x,W)$ правильно разделяет точки из Ω_i, $i=1,..., m$, если $a(x,W)=i$, для всех $x \in \Omega_i$, $i=1,..., m$. Если функции $f_i(x,W^i)$ линейны – $f_i(x,W^i)=(w^i,x)+w_0^i$, $i=1,..., m$, то классификатор $a(x,W)$ называется линейным. Задачи построения линейных классификаторов рассматривались в [LLV, 2010], [ZhLV, 2010].

Будем предполагать, что для каждого класса i задана совокупность базисных функций $\varphi_0^i(x) \equiv 1$ $\varphi_j^i(x),\ j=1,...,L_i$, на основе которых формируются весовые функции $f_i(x,W^i)$, $W^i=(w_0^i,\ w_1^i,..., w_{L_i}^i)$, $i=1,...,m$

$$f_i(x,W^i) = \sum_{j=1}^{L_i} w_j^i \varphi_j^i(x) + w_0^i \tag{2}$$

Совокупности базисных функций могут совпадать для разных классов, например, $\varphi_j^i(x)=x_j$, $j=1,...,n$, $i=1,...,m$ для линейных классификаторов. В общем случае совокупность базисных функций для каждого класса может быть своя.

В зависимости от выбора совокупности базисных функций рассматриваются задачи построения квадратичных, полиномиальных или других алгоритмов классификации.

Множества Ω_i, $i=1,..., m$ называются разделимыми в заданной совокупности базисных функций, если существует классификатор, сформированный на основе этой совокупности базисных функций, правильно разделяющий точки из этих множеств.

Обозначим $\Omega = \bigcup_{i=1}^{m} \Omega_i$. Пусть точки множества Ω перенумерованы, T – совокупность индексов, $\Omega = \left\{ x^t : t \in T \right\}$, T_i – совокупность индексов множества Ω_i, $\Omega_i = \left\{ x^t : t \in T_i \right\}$, $T = \bigcup_{i=1}^{m} T_i$.

Положим $i(t)$ – номер множества Ω_i, которому принадлежит точка x^t, $t \in T$. Величина

$$g^t(W) = \min\left\{ f_i(x^t,W^i) - f_j(x^t,W^j) : j \in \left\{1,..., m\right\} \setminus i,\ i=i(t) \right\} \tag{3}$$

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

Обозначим $z_j^{it} = \varphi_j^i(x^t)$, тогда зазор $g^t(W)$ можно представить в виде

$$g^t(W) = \min\left\{\sum_{k=1}^{L_i} w_k^i z_k^{it} - \sum_{k=1}^{L_j} w_k^j z_k^{jt} + w_0^i - w_0^j \;:\; j \in \{1,\ldots,m\} \setminus i,\; i = i(t)\right\} \tag{4}$$

Заметим, что $g^t(W)$ – вогнутая кусочно-линейная функция. Классификатор $a(x,W)$ допускает ошибку на точке x^t тогда и только тогда, когда зазор $g^t(W)$ отрицателен.

Величина $g(W) = \min\left\{g^t(W) : t \in T\right\}$ 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 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 надёжнее разделяются точки из Ω_i, $i = 1,\ldots, m$), однако, при этом должна учитываться некоторая нормировка совокупности векторов W, которую обозначим $\eta(W)$ и будем называть нормой классификатора $a(x,W)$. Вопрос выбора нормировки имеет самостоятельное значение и должен рассматриваться отдельно.

В дальнейшем будем использовать функцию:

$$\eta(W) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{L_i} (w_j^i)^2} \tag{5}$$

В качестве нормы $\eta(W)$ могут использоваться также другие функции [LLV, 2010]. Норма вида (5) выбрана из тех соображений, чтобы для случая линейных классификаторов и двух множеств (классов) Ω_1 и Ω_2 сформулированные задачи были эквивалентны традиционно рассматриваемым, в частности, задаче, которая решается в методе опорных векторов.

С учетом введенных обозначений задачу построения оптимального классификатора (определения значений параметров W) запишем в следующем виде: найти

$$g^* = \max_W \left\{g(W) : \eta(W) \le 1,\ W \in R^L\right\} \tag{6}$$

где L – размерность вектора W. Поскольку вектор $W = 0$ является допустимым, то задача (6) имеет решение всегда и $g^* \ge g(0) = 0$. Заметим, что $g^* > 0$, если множества Ω_i, $i = 1,\ldots, m$ разделимы в заданной совокупности базисных функций, т.е. существует классификатор, правильно разделяющий эти множества. Рассмотрим также задачу: найти

$$\eta^* = \min_V \left\{\eta(V) : g(V) \ge 1,\ V \in R^L\right\} \tag{7}$$

Близкие по смыслу задачи рассматривались разными авторами (см., напр., [4, 9]).

Лемма 1 [LLV, 2010]**.** Пусть W^* – оптимальное решение задачи (6). Тогда если $g^* > 0$, то задача (7) также имеет оптимальное решение V^* и $V^* = W^* / g^*$, $\eta^* = 1 / g^*$, если $g^* = 0$, то задача (7) не имеет допустимых решений.

Рассмотрим более подробно задачи построения нелинейных классификаторов. Учитывая (4), представим задачу (6) в форме задачи линейного программирования с дополнительным квадратичным ограничением: найти

$$g^* = \max_{w,\delta} \delta \tag{8}$$

при ограничениях

$$\sum_{k=1}^{L_i} w_k^i z_k^{it} - \sum_{k=1}^{L_j} w_k^j z_k^{jt} + w_0^i - w_0^j \geq \delta,\ \ j \in \{1,..., m\} \setminus i,\ t \in T_i,\ i = 1,..., m \tag{9}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{L_i} (w_j^i)^2 \leq 1 \tag{10}$$

Соответственно, задача (7) есть задача квадратичного программирования: найти

$$\eta^* = \min_{v} \sum_{i=1}^{m} \sum_{j=1}^{L_i} (v_j^i)^2 \tag{11}$$

при ограничениях

$$\sum_{k=1}^{L_i} v_k^i z_k^{it} - \sum_{k=1}^{L_j} v_k^j z_k^{jt} + w_0^i - w_0^j \geq 1,\ \ j \in \{1,..., m\} \setminus i,\ t \in T_i,\ i = 1,..., m \tag{12}$$

Можно показать, что в случае линейных классификаторов и $m = 2$ задача (11)-(12) эквивалентна задаче, которая решается с целью построения полосы максимальной ширины, разделяющей линейно разделимые множества Ω_1, Ω_2 .

Для решения рассмотренных задач может использоваться существующее эффективное программное задач оптимизации общего назначения, если число точек в обучающей выборке не очень велико. Результаты такого рода вычислительных экспериментов приведены в [LLV, 2010]. При большом числе точек в обучающей выборке целесообразно использовать методы негладкой оптимизации [СБЖК, 2009], [Sh, 1998].

Задачи (8)-(10) и (11)-(12) позволяют находить оптимальный классификатор только в случае, когда множества Ω_i, $i = 1,..., m$ разделимы в заданной совокупности базисных функций. В противном случае совокупность базисных функций можно расширить, т.е. от линейных классификаторов можно перейти к квадратичным или к полиномиальным более высокого порядка. Такие переходы к полиномиальным классификаторам должны быть обоснованы содержательно в результате анализа рассматриваемой прикладной задачи, поскольку неразделимость множества Ω_i, $i = 1,..., m$ в заданной совокупности базисных функций может быть связана не со сложностью конфигурации классов, а с ошибками определения координат точек множеств Ω_i.

Таким образом, возникает необходимость разработки таких формулировок задач, которые позволяют строить рациональные классификаторы в случае, когда множества Ω_i неразделимы в заданной совокупности базисных функций.

Для случая двух множеств Ω_1, Ω_2 в ряде работ (см., напр. [B, 2009], [BM, 1992], [СБЖК, 2009]) предлагаются эвристические формулировки задач, позволяющие находить линейные классификаторы для линейно неразделимых множеств. Наиболее известным в этой связи является метод опорных векторов (см., напр. [В, 2009]).

2. Минимизация эмпирического риска

В случае, когда множества Ω_i, $i=1,...,m$, неразделимы в заданной совокупности базисных функций, естественным критерием выбора классификатора является минимизация эмпирического риска, т.е. числа точек обучающей выборки, которые классификатор разделяет неправильно.

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

В дальнейшем значение эмпирического риска будем определять с учетом надежности, определяемой параметром $\overline{\delta}$ – 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 (10). Тогда существует конечное число $R>0$ такое, что

$$-R\le w_0^i-w_0^j\le R \tag{13}$$

Доказательство. Для точек x^α, x^β можно записать неравенства $\sum_{k=1}^{L_j} w_k^j z_k^{j\alpha}-\sum_{k=1}^{L_i} w_k^i z_k^{i\alpha}\le w_0^i-w_0^j\le\sum_{k=1}^{L_j} w_k^j z_k^{j\beta}-\sum_{k=1}^{L_i} w_k^i z_k^{i\beta}$. Учитывая (10), можно показать, что $-\sqrt{\sum_{k=1}^{L_j}\left(z_k^{j\alpha}\right)^2}-\sqrt{\sum_{k=1}^{L_i}\left(z_k^{i\alpha}\right)^2}\le w_0^i-w_0^j\le\sqrt{\sum_{k=1}^{L_j}\left(z_k^{j\beta}\right)^2}+\sqrt{\sum_{k=1}^{L_i}\left(z_k^{i\beta}\right)^2}$. Откуда следует утверждение леммы. !

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 (8)-(10), $y_t=1$ – в противном случае. Пусть задано большое положительное число B. Задача минимизации эмпирического риска имеет вид: найти

$$Q^*=\min_{w,y}\left\{\sum_{t\in T} y_t\right\} \tag{14}$$

при ограничениях

$$\sum_{k=1}^{L_i} w_k^i z_k^{it}-\sum_{k=1}^{L_j} w_k^j z_k^{jt}+w_0^i-w_0^j\ge\overline{\delta}-B\cdot y_t,\ \ j\in\{1,...,m\}\setminus i,\ \ t\in T_i,\ \ i=1,...,m \tag{15}$$

$$\eta(W)\le 1 \tag{16}$$

$$\sum_{t\in T_i} y_t\le|T_i|-1,\quad i=1,...,m \tag{17}$$

$$0\le y_t\le 1,\ t\in T \tag{18}$$

$$y_t=0\vee 1,\ t\in T \tag{19}$$

Из ограничений (16), (13) следует, что если $y_t=1$, то при достаточно большом значении числа B соответствующие неравенства вида (15) выполняются всегда, т.е. точка x^t исключается из задачи.

Ограничения (17) 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 $\bar{\delta}$. Задача (14)-(19) является NP-трудной, но приближенные решения могут быть найдены сравнительно просто по решению непрерывной релаксации этой задачи и могут быть уточнены в рамках общей схемы метода ветвей и границ. При формулировке задачи минимизации эмпирического риска могут учитываться дополнительные условия и информация.

Для вычисления оценки снизу для величины Q^* (минимального эмпирического риска) рассмотрим непрерывную релаксацию сформулированной задачи – задачу (14)-(18) Оптимальное значение релаксированной задачи обозначим q^*. Для ее решения будем использовать схему декомпозиции по переменным W. Пусть переменные W зафиксированы. Учитывая (4), задачу минимизации по переменным y представим в следующем виде: найти

$$q(W) = \min_y \left\{ \sum_{t \in T} y_t \right\} \tag{20}$$

при ограничениях

$$y_t \geq \frac{1}{B}\left(\bar{\delta} - g^t(W)\right),\ t \in T \tag{21}$$

$$\eta(W) \leq 1 \tag{22}$$

$$\sum_{t \in T_i} y_t \leq |T_i| - 1, \quad i = 1, \ldots, m \tag{23}$$

$$0 \leq y_t \leq 1,\ t \in T \tag{24}$$

Обозначим $d^t(W) = \max\left(0,\ \frac{1}{B}\left(\bar{\delta} - g^t(W)\right)\right)$. Очевидно, что если задача (20)-(24) имеет решение, то $y^t = d^t(W)$. Откуда получаем задачу минимизации по переменным W: найти

$$q^* = \min \sum_{t \in T} d^t(W) \tag{25}$$

при ограничениях

$$\eta(W) \leq 1 \tag{26}$$

$$\sum_{t \in T_i} d^t(W) \leq |T_i| - 1,\ i = 1, \ldots, m \tag{27}$$

$$d^t(W) \leq 1,\ t \in T \tag{28}$$

Функции $d^t(W)$ – выпуклые кусочно-линейные, $\eta(W)$ – квадратичная положительно определенная.

Задача (25)-(28) для случая линейных классификаторов рассматривалась в [ZhLV, 2010]. Было показано, что лагранжева релаксация этой задачи при специальном подборе значений множителей Лагранжа эквивалентна задаче, которая решается в методе опорных векторов.

Для решения задачи (25)-(28) вычисления нижней оценки минимального эмпирического риска может использоваться метод негладких штрафных функций и эффективные алгоритмы безусловной выпуклой оптимизации [Sh, 1998]. При использовании такого подхода требуется каким-либо образом подбирать значения штрафных коэффициентов.

Другими подходами являются использование выпуклых конических продолжений [ЛЛ, 2010] или конических аппроксимаций целевой функции [ЛБ, 2011].

Заключение

В работе для случая многих классов рассмотрены задачи построения нелинейных классификаторов, весовые функции которых представимы в виде линейной комбинации заданной совокупности базисных (в общем случае нелинейных) функций. При условии того, что обучающая выборка разделима в заданной совокупности классификаторов, естественным образом формулируются и эффективно решаются задачи построения оптимальных классификаторов. В случае неразделимости обучающей выборки формулируется задача построения классификатора, обеспечивающего минимизацию эмпирического риска. Такая задача является частично целочисленной и NP-трудной. Для практического использования рассмотрена непрерывная релаксация задачи минимизации эмпирического риска, которая сводится к задаче выпуклого программирования с кусочно-линейными функциями, и предложены некоторые новые подходы к решению таких задач.

References

[BM, 1992] Bennett K.P., Mangasarian O.L. Robust Linear Programming Discrimination of Two Linearly Inseparable Sets // Optimization Methods and Software. – 1992. –№5. – P. 23-34.

[BTGN, 2009] Ben-Tal A., Ghaoui L.E., Nemirovsri A. Robust Optimization. – Princeton University Press, 2009. – 542 p.

[DD, 2009] 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.

[DN, 2009] Koel Das, Zoran Nenadic. An efficient discriminant-based solution for small sample size problem // Pattern Recognition – Volume 42, Issue 5, 2009, Pages 857-866.

[LLV, 2010] 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, Vol. 20, No. 2, 2010, p. 137-145.

[Sh, 1998] Shor N.Z. Nondifferentiable Optimization and Polynomial Problems. – Dordrecht, Kluwer, 1998. – 394 p.

[ZhLV, 2010] Zhuravlev Yu., Laptin Yu, Vinogradov A. Minimization of empirical risk in linear classifier problem // New Trends in Classification and Data Mining, ITHEA, Sofia, Bulgaria, 2010. – Pages 9-15

[В, 2009] 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

[ГС, 2008] Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания. - Киев: Наук. думка, 2008. - 232 с.

[ЛБ, 2011] Лаптин Ю.П., Бардадым Т.А. Некоторые подходы к регуляризации нелинейных задач оптимизации // Проблемы управления и информатики. 2011, № 3. – С. 57-68.

[ЛЛ, 2010] Лаптин Ю.П. Лиховид А.П. Использование выпуклых продолжений функций для решения нелинейных задач оптимизации // Управляющие системы и машины. 2010, № 6. – С. 25–31.

[Маз, 2010] Мазуров Вл.Д. Математические методы распознавания образов.– Екатеринбург, 2010. – 101 с. – Уральский государственный университет им. А.М.Горького.

[Мест, 2008] Местецкий Л.М. Математические методы распознавания образов. – http://www.intuit.ru/department/graphics/imageproc/ – Опубликовано 30.04.2008

[СБЖК, 2009] 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)

[ШГ, 2004] Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. – К.: Наукова думка, 2004. – 545 с.

Информация об авторах

Ю.И. Журавлев *– академик РАН, заместитель директора по научной работе Вычислительного центра им. А. А. Дородницына РАН, ул. Вавилова 40, 119333 Москва, Российская федерация; e-mail: zhuravlev@ccas.ru*

Ю.П. Лаптин *– старший научный сотрудник, Институт кибернетики им. В.М. Глушкова НАН Украины, пр. академика Глушкова, 40, 03650 Киев, Украина; e-mail: laptin_yu_p@mail.ru*

А.П. Виноградов *– старший научный сотрудник, Вычислительный центр им. А. А. Дородницына РАН, ул. Вавилова 40, 119333 Москва, Российская федерация; e-mail: vngrccas@mail.ru*