Задача обучения распознаванию образов в нестационарной генеральной совокупности

Шавловский М. Б., Красоткина О. В., Моттль В. В.
shavlovsky@yandex.ru, ko180177@yandex.ru, vmottl@yandex.ru
Москва, МФТИ; Тула, ТулГУ; Москва, ВЦ РАН

Целью данного исследования является создание основного математического аппарата и простейших алгоритмов для решения типичных для практики задач обучения распознаванию образов в генеральных совокупностях, свойства которых изменяются во времени. Широко известная классическая постановка задачи распознавания основана на молчаливом предположении, что свойства генеральной совокупности в момент «экзамена» остаются теми же, что и при формировании обучающей выборки. Принятое в данной работе более реалистичное предположение о нестационарности генеральной совокупности неизбежно приводит к необходимости анализа последовательности выборок в некоторые моменты времени и поиска для них разных решающих правил распознавания.

Пусть каждый объект генеральной совокупности $\omega \in \Omega$ представлен точкой в линейном пространстве признаков $\mathbf{x}(\omega) = \big(x^1(\omega), \ldots, x^n(\omega)\big) \in \in \mathbb{R}^n$, а его скрытая фактическая принадлежность к одному из двух классов определяется значением индекса класса $y(\omega) \in \{1, -1\}$. Классический подход к обучению распознаванию двух классов объектов, развитый В. Н. Вапником [1], основан на понимании модели генеральной совокупности в виде дискриминантной функции $f\big(\mathbf{x}(\omega)\big) = \mathbf{a}^T\mathbf{x} + b$, определяемой гиперплоскостью с априори неизвестными наблюдателю направляющим вектором $\mathbf{a} \in \mathbb{R}^n$ и параметром положения $b \in \mathbb{R}$:

$$f\big(\mathbf{x}(\omega)\big) = \mathbf{a}^T\mathbf{x} + b \begin{cases} \text{преимущественно } > 0, & \text{если } y(\omega) = 1, \\ \text{преимущественно } < 0, & \text{если } y(\omega) = -1. \end{cases}$$

Неизвестные параметры разделяющей гиперплоскости подлежат оцениванию на основе анализа обучающей совокупности объектов $\{\omega_j,\ j = \\ = 1, \ldots, N\}$, представленных векторами их признаков и индексами принадлежности к классам, так что выборка в целом является конечным множеством пар $\big\{(\mathbf{x}_j \in \mathbb{R}^n,\ y_j = \pm 1),\ j = 1, \ldots, N\big\}$. Широко известен принцип оптимальной разделяющей гиперплоскости, выбираемой по критерию максимизации числа точек обучающей совокупности, правильно классифицируемых с гарантированным «запасом», условно равным единице:

$$\begin{cases} J(\mathbf{a}, b, \delta_j) = \mathbf{a}^T\mathbf{a} + C\sum\limits_{j=1}^{N}\delta_j \to \min, \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geqslant 1 - \delta_j, \quad \delta_j \geqslant 0, \quad j = 1, \ldots, N. \end{cases} \tag{1}$$

Понятие времени здесь полностью отсутствует.

Принципиальное отличие предлагаемой в данной работе концепции нестационарной генеральной совокупности заключается во введении в рассмотрение фактора времени t. Предполагается, что основное свойство нестационарной генеральной совокупности выражается изменяющейся во времени разделяющей гиперплоскостью, характеризующей преимущественное различие векторов признаков объектов двух классов и, в свою очередь, полностью определяемой своим направляющим вектором и параметром положения, которые должны рассматриваться как функции времени $\mathbf{a}_t$ и b_t:

$$f_t\big(\mathbf{x}(\omega)\big) = \mathbf{a}_t^T\mathbf{x}+b_t \begin{cases} \text{преимущественно} > 0 \text{ в момент } t, & \text{если } y(\omega) = 1, \\ \text{преимущественно} < 0 \text{ в момент } t, & \text{если } y(\omega) = -1. \end{cases}$$

Здесь всякий объект $\omega \in \Omega$ рассматривается всегда только вместе с указанием момента времени, в который он предъявлен (ω, t). В результате обучающая совокупность приобретает структуру множества троек $\big\{(\mathbf{x}_j \in \mathbb{R}^n,\ y_j = \pm 1,\ t_j),\ j = 1, \ldots, N\big\}$, а не пар. Естественно нумеровать объекты обучающей совокупности в порядке поступления объектов, тогда уместно говорить скорее об обучающей последовательности, нежели об обучающей совокупности, рассматривая ее как временной ряд, вообще говоря, с переменным шагом по времени.

В разные момент времени t_j скрытая от наблюдателя разделяющая гиперплоскость характеризуется разными неизвестными значениями направляющего вектора и параметра положения. Таким образом, объективно существует двухкомпонентный временной ряд со скрытой и наблюдаемой компонентами, соответственно, $(\mathbf{a}_j, b_j)$ и $(\mathbf{x}_j, y_j)$.

В динамической постановке задача обучения превращается в задачу анализа двухкомпонентного временного ряда, в котором требуется, анализируя наблюдаемую компоненту, дать оценку скрытой компоненты. Это стандартная задача анализа сигналов (временных рядов), специфика которой заключается лишь в предполагаемой модели связи между скрытой и наблюдаемой компонентами. Согласно классификации задач оценивания скрытой компоненты сигнала, введенной Н. Винером [2], естественно различать, по крайнем мере, два вида задач обучения.

Задача фильтрации обучающей последовательности. Пусть t_j — момент поступления очередного объекта, к которому уже зарегистрированы векторы признаков и индексы классов объектов $\big\{\ldots, (\mathbf{x}_{j-2}, y_{j-2}), (\mathbf{x}_{j-1}, y_{j-1}), (\mathbf{x}_j, y_j)\big\}$, поступивших в предыдущие моменты времени до текущего момента включительно $(\ldots, t_{j-2}, t_{j-1}, t_j)$. Требуется непосредственно в процессе наблюдения давать оценку параметров разделяющей гиперплоскости $(\hat{\mathbf{a}}_j, \hat{b}_j)$ в каждый текущий момент времени t_j.

Задача интерполяции. Пусть к моменту обработки обучающая последовательность уже зарегистрирована в некотором интервале времени $\big\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\big\}$. Требуется оценить изменяющиеся параметры разделяющей гиперплоскости во всем интервале наблюдения $\big\{(\hat{\mathbf{a}}_1, \hat{b}_1), \ldots, (\hat{\mathbf{a}}_N, \hat{b}_N)\big\}$.

Предполагается, что параметры разделяющей гиперплоскости $\mathbf{a}_t$ и b_t изменяются во времени достаточно медленно в том смысле, что величины

$$\tfrac{1}{t_j - t_{j-1}}(\mathbf{a}_j - \mathbf{a}_{j-1})^{\mathrm{T}}(\mathbf{a}_j - \mathbf{a}_{j-1}) \approxeq \varepsilon_{\mathbf{a}} \quad \text{и} \quad \tfrac{1}{t_j - t_{j-1}}(b_j - b_{j-1})^2 \approxeq \varepsilon_b$$

являются, как правило, достаточно малыми. Это предположение препятствует вырождению задач фильтрации и интерполяции в совокупность независимых некорректных задач обучения распознаванию двух классов объектов по единственному наблюдению.

С формальной точки зрения оценка параметров разделяющей гиперплоскости в последний момент интервала наблюдения $(\hat{\mathbf{a}}_N, \hat{b}_N)$, полученная при решении задачи интерполяции, является решением задачи фильтрации для этого момента. Однако смысл задачи фильтрации заключается в том, чтобы очередные оценки вычислялись непосредственно в процессе поступления новых наблюдений, без решения всякий раз задачи интерполяции для временного ряда возрастающей длины.

Предлагаемая постановка задачи обучения в режиме интерполяции отличается от совокупности классических задач обучения по методу опорных векторов (1) для каждого момента времени только наличием дополнительных членов, штрафующих различие между смежными значениями параметров гиперплоскости $(\mathbf{a}_{j-1}, b_{j-1})$ и $(\mathbf{a}_j, b_j)$:

$$\begin{cases} J(\mathbf{a}_j, b_j, \delta_j, j = 1, \ldots, N) = \sum\limits_{j=1}^{N} \big(\tfrac{1}{N}\mathbf{a}_j^T \mathbf{a}_j + \delta_j\big) + \\ + \sum\limits_{j=2}^{N} \tfrac{1}{t_j - t_{j-1}}\big[D^{\mathbf{a}}(\mathbf{a}_j - \mathbf{a}_{j-1})^{\mathrm{T}}(\mathbf{a}_j - \mathbf{a}_{j-1}) + D^b(b_j - b_{j-1})^2\big] \to \min, \\ y_j(\mathbf{a}_j^T \mathbf{x}_j + b_j) \geqslant 1 - \delta_j, \quad \delta_j \geqslant 0, \quad j = 1, \ldots, N. \end{cases} \tag{2}$$

Здесь коэффициенты $D^{\mathbf{a}} > 0$ и $D^b > 0$ являются параметрами критерия, задающими желаемую степень сглаживания оцениваемой последовательности мгновенных значений параметров разделяющей гиперплоскости.

Критерий (2) реализует концепцию оптимальной достаточно гладкой последовательности разделяющих гиперплоскостей в отличие от концепции единственной оптимальной гиперплоскости в (1). Искомые гиперплоскости должны обеспечивать правильную классификацию векторов признаков объектов для как можно большего числа моментов времени с гарантированным «запасом», принятым равным единице, как и в (1).

Как и классическая задача обучения, динамическая задача (2) является задачей квадратичного программирования, но содержит $N(n+1)+N$ переменных, в отличие от $(n+1)+N$ переменных в (1). Известно, что вычислительная сложность задачи квадратичного программирования общего вида пропорциональна кубу числа переменных, т. е. динамическая задача, на первый взгляд, существенно сложнее классической.

Однако целевая функция $J(\mathbf{a}_j, b_j, \delta_j,\, j = 1, \ldots, N)$ в динамической задаче является парно-сепарабельной, т.е. представляя собой сумму частных функций, каждая из которых зависит от переменных, связанных только с одним либо двумя моментами времени в порядке их возрастания. Это обстоятельство позволяет построить алгоритм численного решения задачи, вычислительная сложность которого линейна относительно длины обучающей последовательности N.

Применение теоремы Куна-Таккера к динамической задаче (2) переводит ее в двойственную форму относительно множителей Лагранжа $\lambda_j \geqslant 0$ при ограничениях-неравенствах $y_j(\mathbf{a}_j^T \mathbf{x}_j + b_j) \geqslant 1 - \delta_j$:

$$\begin{cases} W(\lambda_1, \ldots, \lambda_N) = \\ \quad = \sum\limits_{j=1}^{N} \lambda_j - \frac{1}{2} \sum\limits_{j=1}^{N} \sum\limits_{l=1}^{N} y_j y_l \big(\mathbf{a}_j^{\mathrm{T}} \mathbf{Q}_{jl} \mathbf{a}_l + f_{jl}\big) \lambda_j \lambda_l \to \max, \\ \sum\limits_{j=1}^{N} y_j \lambda_j = 0, \quad 0 \leqslant \lambda_j \leqslant C/2, \quad j = 1, \ldots, N. \end{cases} \tag{3}$$

Здесь матрицы $\mathbf{Q}_{jl}(n \times n)$ и $\mathbf{F} = (f_{jl})\,(N \times N)$ не зависят от обучающей последовательности и определяются только коэффициентами штрафа, соответственно, $D^{\mathbf{a}}$ на негладкость последовательности направляющих векторов искомых гиперплоскостей и D^b на негладкость последовательности их параметров положения в (2).

Теорема 1. *Решение задачи обучения (2) полностью определяется значениями множителей Лагранжа $(\lambda_1, \ldots, \lambda_N)$, полученными как решение двойственной задачи (3), и обучающей последовательностью:*

$$\hat{\mathbf{a}}_j = \sum_{l:\ \lambda_l > 0} y_l \lambda_l \mathbf{Q}_{jl} \mathbf{x}_l; \qquad \hat{b}_j = b + \sum_{l:\ \lambda_l > 0} y_l \lambda_l f_{jl}; \tag{4}$$

$$b = \frac{\sum\limits_{j:\ 0<\lambda_j<C/2} \lambda_j \sum\limits_{l:\ \lambda_j>0} y_l \lambda_l \big(\mathbf{x}_l^{\mathrm{T}} \mathbf{Q}_{jl} \mathbf{x}_j + f_{jl}\big) + (C/2) \sum\limits_{j:\ \lambda_j = C/2} y_j}{\sum\limits_{j:0<\lambda_j<C/2} \lambda_j}. \tag{5}$$

Из этих формул видно, что решение задачи динамического обучения определяется только теми элементами обучающей последовательности

$(\mathbf{x}_j, y_j)$, множители Лагранжа при которых получили положительные значения $\lambda_j > 0$. Уместно назвать векторы признаков соответствующих объектов опорными векторами, так что мы пришли к некоторому обобщению метода опорных векторов [1], вытекающего из концепции оптимальной разделяющей гиперплоскости (1).

Классическая задача обучения (1) является частным случаем задачи (2) при бесконечно больших значениях штрафов на изменение параметров гиперплоскости $D^{\mathbf{a}} \to \infty$ и $D^b \to \infty$. В этом случае $\mathbf{Q}_{jl} \to \mathbf{I}$, $f_{jl} \to 0$, и двойственная задача (3) превращается в классическую двойственную задачу [1], соответствующую исходной задаче (1), а формулы (4) и (5) определяют результат обучения по классическому методу опорных векторов $\hat{\mathbf{a}} = \hat{\mathbf{a}}_1 = \ldots = \hat{\mathbf{a}}_N$, $\hat{b} = \hat{b}_1 = \ldots = \hat{b}_N$.

Хотя двойственная задача (3) и не является парно-сепарабельной, в силу парной сепарабельности исходной задачи (2) вычисление градиента целевой функции $\nabla_{\boldsymbol{\lambda}} W(\lambda_1, \ldots, \lambda_N)$ в любой точке $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ и, далее, определение оптимального допустимого направления поиска с учетом ограничений обеспечиваются алгоритмом, имеющим линейную вычислительную сложность относительно длины обучающей последовательности. В результате оказывается, что использование любого градиентного метода для решения двойственной задачи приводит к алгоритму с линейной вычислительной сложностью относительно N. В частности, стандартный метод наискорейшего спуска для решения задач квадратичного программирования [3], примененный к функции $-W(\lambda_1, \ldots, \lambda_N)$, дает алгоритм, являющийся, по сути, обобщением известного алгоритма SMO (Sequential Minimum Optimization) [4], обычно используемого при решении двойственных задач.

Работа выполнена при поддержке РФФИ, проекты №05-01-00679, №06-01-00412.

Литература

[1] *Vapnik V.* Statistical Learning Theory. — New York: John-Wiley & Sons, Inc., 1998. — 732 p.

[2] *Wiener N.* Extrapolation, Interpolation, and Smoothing of Stationary Random Time Series with Engineering Applications. — Technology Press of MIT, John Wiley & Sons, 1949. — 163 p.

[3] *Базара М., Шетти К.* Нелинейное программирование. Теория и алгоритмы. — М.: Мир, 1982.

[4] *Platt J.C.* Fast training of support vector machines using sequential minimal optimization / Advances in Kernel Methods: Support Vector Learning. — MIT Press, Cambridge, MA, 1999.