О согласованных оценках сложности задач и алгоритмов классификации

Романов Л. Ю.
lromanov@gmail.com
Москва, Вычислительный центр РАН

В докладе рассматриваются два независимо определяемых понятия сложности: геометрическая сложность конфигурации объектов обучающей выборки и функциональная сложность выборки как сложность разделяющей поверхности. Вычислительные эксперименты подтверждают гипотезу о корреляции указанных сложностей. Строится экспериментальная зависимость, позволяющая по одной из величин произвести оценку другой.

Постановка задачи

Рассматривается стандартная задача классификации с двумя непересекающимися классами в евклидовом пространстве размерности n.

Будем изучать возможную зависимость сложностей на примере алгоритма SVM [1, 2], строящего в заданном пространстве линейную разделяющую поверхность. Для построения нелинейного решающего правила используется следующая схема: исходное пространство расширяется дополнительными осями (признаками) таким образом, чтобы SVM разделял обучающую выборку без ошибок. Значения координат по каждой из дополнительных осей являются нелинейными функциями от исходных координат: $x_{n+k} = g_k(x_1, \ldots, x_n),\ k = 1, 2, \ldots$ Построенное таким образом пространство называют *спрямляющим*. В исходном пространстве разделяющая поверхность оказывается нелинейной и разделяет обучающую выборку без ошибок.

Основной проблемой при построении спрямляющего пространства является нахождение оптимального набора дополнительных осей. Поэтому важной задачей представляется оценивание сложности спрямляющего пространства до его построения, на основе лишь геометрической сложности обучающей выборки.

Геометрическая сложность обучающей выборки

Геометрическую сложность можно определять различными способами. Наибольший интерес представляет степень взаимного проникновения классов друг в друга. Для оценки этой величины можно разными способами измерять расстояние между классами.

Определение 1. *Геометрической сложностью обучающей выборки в n-мерном пространстве будем называть отношение стороны описан-*

ного около выборки n-мерного куба к минимальному расстоянию между объектами двух классов.

Функциональная сложность обучающей выборки

Основная идея нахождения сложности разделяющей поверхности состоит в том, что в качестве сложности поверхности принимается определенная некоторым образом сложность спрямляющего пространства, достаточного для разделения обучающей выборки без ошибок.

Будем говорить, что дополнительная ось задается функционалом $g(x_1, \ldots, x_n)$, если вводится новая координата, значения которой для рассматриваемого множества точек задаются указанным функционалом.

Функционалы будем задавать в виде суперпозиции функций некоторого базиса $\mathscr{F} = \left\{ f_i \colon \mathbb{R}^{k_i} \to \mathbb{R} \;\middle|\; i = 1, \ldots, m \right\}$.

Будем считать, что каждая функция f_i из базиса $\mathscr{F}$ обладает некоторой сложностью c_i, которая задается исследователем априори. Сложность суперпозиции функций определим как суммарную сложность входящих в нее функций.

Рассмотрим всевозможные наборы дополнительных осей, дающих линейное разделение выборки. Сложность набора осей определим как сумму сложностей функционалов, задающих оси из набора. Функциональную сложность выборки определим как минимальную сложность среди таких наборов осей.

Алгоритм нахождения функциональной сложности

Пусть мы имеем некоторую обучающую выборку. Следующий алгоритм описывает процедуру нахождения ее функциональной сложности.

Перебираются всевозможные наборы функций, образованные суперпозициями функций из базиса, в порядке увеличения сложности набора. Для каждого набора, задающего дополнительные оси спрямляющего пространства, строится линейное разделение в расширенном ими пространстве с помощью алгоритма SVM. Если расширенное пространство является спрямляющим, то функциональная сложность найдена, иначе перебор продолжается.

Эмпирическая зависимость функциональной сложности от геометрической

Для эмпирического исследования зависимости функциональной сложности от геометрической строятся обучающие выборки с заданной геометрической сложностью, для каждой из них с помощью приведенного алгоритма вычисляется функциональная сложность.

В качестве реализации метода SVM в данной работе используется алгоритм SMO [3].

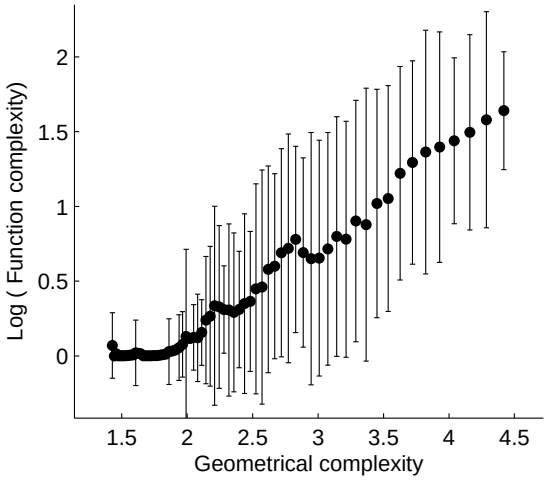

Рис. 1. Зависимость функциональной сложности от геометрической (усреднение по 50 ближайшим значениям геометрической сложности).

В настоящей работе описанный выше метод анализа проводится при некоторых упрощениях. Во-первых, рассматриваются двумерные выборки. Во-вторых, в качестве функций, задающих координаты дополнительных осей, рассматриваются мономы $f_i(x_1, x_2) = x_1^{\alpha_1^i} x_2^{\alpha_2^i}$, сложность которых определяется как $C(\alpha_1) + C(\alpha_2)$, где

$$C(\alpha) = \begin{cases} \alpha, & \text{если } \alpha = 1, 2, 3, \ldots; \\ \beta + 1, & \text{если } \alpha = -\beta, \quad \beta = 1, 2, 3, \ldots; \\ \beta, & \text{если } \alpha = \frac{1}{\beta}, \quad \beta = 2, 3, \ldots; \\ \beta + 1, & \text{если } \alpha = -\frac{1}{\beta}, \quad \beta = 2, 3, \ldots. \end{cases}$$

График экспериментальной зависимости функциональной сложности от геометрической приведен на Рис. 1. Для отображения функциональной сложности выбрана логарифмическая шкала. Для каждого значения также отложена дисперсия.

Работа выполнена при поддержке РФФИ, проект №06-07-89315-а.

Литература

[1] *Vapnik V.* Estimation of Dependences Based on Empirical Data — Springer–Verlag, 1982.

[2] *Burges C. J. C.* A tutorial on support vector machines for pattern recognition. — Data Mining and Knowledge Discovery. — Vol. 2, No. 2. — 1998.

[3] *Platt J. C.* Sequential minimal optimization: A fast algorithm fo training support vector machines. — Technical Report MSR–TR–98–14, Microsoft Research, 1998.