Адаптивное планирование эксперимента в распознавании образов и в регрессионном анализе с использованием класса логических решающих функций

Лбов Г. С., Бериков В. Б.
lbov@math.nsc.ru, berikov@math.nsc.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН

В достаточно широком круге задач интеллектуального анализа данных, возникающих в различных трудноформализуемых областях исследований, имеется возможность активного влияния на выборку. Так, например, при изучении нового способа лечения можно целенаправленно проводить отбор пациентов с интересующей формой патологии; при исследовании мутационных спектров ДНК имеется возможность выбора генетической последовательности, подвергающейся действию мутагена, и т. д. Сбор экспериментальной информации и анализ полученных данных связаны с большими затратами, которые определяются числом изучаемых объектов. Выбор объектов может быть организован так, чтобы добиться наилучшего качества при заданном числе наблюдений.

Одним из возможных методов решения такого рода задач является предлагаемый метод, основанный на адаптивном планировании случайного эксперимента с использованием класса логических решающих функций. При адаптивном планировании проводится последовательный случайный отбор объектов с учетом уже выявленных закономерностей в структуре данных. Логические решающие функции, наиболее удобная форма представления которых — деревья решений, позволяют строить легко интерпретируемые модели, одновременно проводя отбор наиболее информативных показателей. Основная задача состоит в том, чтобы построить дерево решений, наилучшим образом приближающее оптимальную байесовскую решающую функцию (либо соответствующую регрессионную функцию) f_0. Предположим, что имеется экспертная информация о том, что вероятность ошибки для f_0 (дисперсия помехи, в случае регрессионного анализа), ограничена некоторой малой величиной. Для задачи распознавания это означает, что образы достаточно хорошо «разделены» в пространстве переменных.

Пусть задано максимально возможное число N объектов анализа, каждый из которых описывается некоторыми показателями, среди которых могут быть показатели как количественной, так и качественной природы. Имеется некоторый прогнозируемый показатель, который может быть либо качественным (в случае задачи распознавания), либо количественным (для задачи регрессионного анализа). Зададим некоторое число L этапов планирования. При проведении l-го этапа планирования

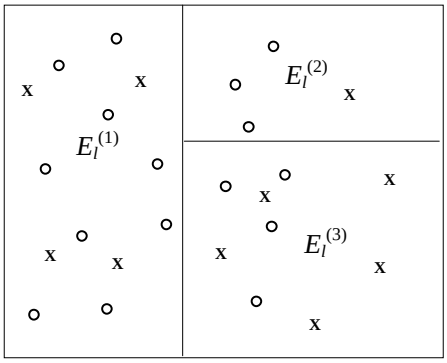

Рис. 1. Пример дерева разбиения и планируемых объектов (x — первый образ; o — второй образ)

используется эмпирическая информация, полученная на основе анализа всех экспериментов предыдущих $l-1$ этапов.

На первом этапе расстановка планируемых точек в пространстве переменных осуществляется случайным образом с использованием равномерного распределения (так как на данном этапе отсутствует информация о поведении прогнозируемой переменной). По сформированной таким образом выборке строится дерево решений. Для этого может использоваться, например, рекурсивный $\mathfrak{R}$-метод [1].

Рассмотрим разбиение области планирования на M подобластей $E_l^{(1)}, \ldots, E_l^{(m)}, \ldots, E_l^{(M)}$, соответствующее листьям дерева решений, построенного по наблюдениям предыдущих этапов (рис. 1; задача распознавания образов).

Проведение l-й группы экспериментов следует организовать так, чтобы в максимальной степени уменьшить степень расхождения с оптимальной функцией. В качестве оценки степени расхождения можно использовать байесовскую оценку вероятности ошибки [2], в которой учитывается число имеющихся подобластей разбиения и экспертная оценка степени «пересечения» образов. Можно также использовать и обычную частотную оценку. В случае задачи регрессионного анализа оценкой может служить величина среднеквадратического отклонения прогнозируемого показателя. Пусть фиксирован способ расстановки планируемых на данном этапе N_l точек, $\sum\limits_{l=1}^{L} N_l = N$, в каждой из подобластей (например, в соответствии с равномерным распределением). При этом адаптация будет заключаться в изменении набора вероятностей $\{P_l^{(1)}, \ldots, P_l^{(m)}, \ldots, P_l^{(M)}\}$ попадания планируемой точки в подобласти. Это изменение должно от-

ражать накопленную на данном этапе информацию о поведении прогнозируемой переменной. Например, если выяснилось, что прогнозируемая количественная переменная в некоторой подобласти изменяется относительно мало, то и вероятность попадания в эту подобласть должна быть уменьшена.

Рассмотрим класс стратегий планирования, для которых выполняется: $P_l^{(m)} = g_l\big(\delta^{(m)}, |E^{(m)}|\big)$, где $\delta^{(m)}$ — оценка степени расхождения с оптимальной функцией в области $E^{(m)}$, $|E^{(m)}|$ — мощность или объем данной области, $g_l(\cdot,\cdot)$ — некоторая заданная неотрицательная функция, монотонно возрастающая по каждому из аргументов, причем должно выполняться $\sum\limits_{m=1}^{M} P_l^{(m)} = 1$. Таким образом, вероятность попадания точки в подобласть увеличивается при возрастании, с одной стороны, ошибки для данной подобласти, а с другой стороны, объема подобласти. Конкретный вид функции g_l может задаваться по разному. Например, можно положить, что эта функция линейна: $g_l(a,b) = \frac{1}{Z_l}\big(\varkappa(l)a + b\big)$, где Z_l — нормировочный коэффициент, $\varkappa(l)$ — коэффициент адаптации, растущий с увеличением номера этапа планирования l («адаптация» означает, что с увеличением объема информации растет степень доверия к соответствующей оценке).

После проведения планирования для данного этапа строится новое дерево решений по сформированной выборке, и т. д.

В докладе будут продемонстрированы результаты экспериментального исследования алгоритма адаптивного планирования.

Работа выполнена при поддержке РФФИ, проект № 07-01-00331а.

Литература

[1] *Лбов Г. С., Бериков В. Б.* Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. — Новосибирск: Изд-во Ин-та математики, 2005. — 220 с.

[2] *Бериков В. Б., Лбов Г. С.* Байесовские оценки качества распознавания по конечному множеству событий // Доклады РАН — 2005. — Т. 402. — № 1. — С. 1–4.