Секция «Вычислительная математика и кибернетика»

Оценки предсказанной информативности логических закономерностей

Дударенко Марина Алексеевна
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: m.dudarenko@gmail.com

Алгоритмы классификации, основанные на голосовании логических закономерностей, широко используются в слабо формализованных прикладных областях благодаря возможности содержательно интерпретировать каждую закономерность. *Закономерностью* класса y называется предикат, который выделяет достаточно много (p) объектов класса y и достаточно мало (n) объектов других классов. *Логической закономерностью* называется конъюнкция пороговых условий вида $[x_j \leq \theta_j]$, где x_j — значение j-го признака, θ_j — параметр. Логические закономерности строятся по заданной обучающей выборке путём оптимизации набора признаков j и порогов θ_j либо по двум критериям $p \to \max$, $n \to \min$, либо по одному критерию информативности $H(p, n) \to \max$.

Для повышения обобщающей способности закономерностей в [1,2] предлагается оценивать значения критериев $\hat{p}$ и $\hat{n}$ на скрытой контрольной выборке и максимизировать *предсказанную информативность* $H(\hat{p}, \hat{n}) \to \max$. Для этого строится граф Хассе множества L-мерных бинарных векторов ошибок, порождаемых семейством логических закономерностей на заданной выборке из L объектов. При получении оценки предполагается, что выборка равновероятно разбивается на две подвыборки — обучающую длины l и контрольную длины k, причём контрольная выборка известна. Однако при использовании предсказанной информативности предполагается другое — что для обучения доступна вся выборка длины L, а критерии $\hat{p}$ и $\hat{n}$ оцениваются на неизвестной контрольной выборке длины K. В таком случае граф Хассе надо было бы строить по «супервыборке» длины $L + K$, но её контрольная часть неизвестна.

Целью данной работы является выяснение условий, при которых оценки предсказанной информативности, сделанные по случайной подвыборке, близки к оценкам по полной выборке. В экспериментах на модельных данных вычислялась корреляция между этими двумя величинами. Оказалось, что с ростом длины выборки корреляция быстро сходится к единице, рис.1. Скорость сходимости падает с ростом числа признаков и уровня шума, но почти не зависит от того, где расположен шум — на границе классов, на периферии или равномерно по всему пространству. Основной вывод заключается в том, что случайные подвыборки объектов в значительной степени сохраняют важнейшие структурные особенности семейства логических закономерностей, как множества бинарных векторов ошибок, за исключением случаев выборок малой длины.

Литература

1. Ивахненко А. А., Воронцов К.В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов // 15-ая Всеросс. конф. Математические методы распознавания образов, 2011. С. 48–51.

2. Vorontsov K. V., Ivahnenko A. A. Tight combinatorial generalization bounds for threshold conjunction rules // Lecture Notes on Computer Science. 4th Int'l Conf. on Pattern Recognition and Machine Intelligence, 2011. Pp 66–73.

Иллюстрации

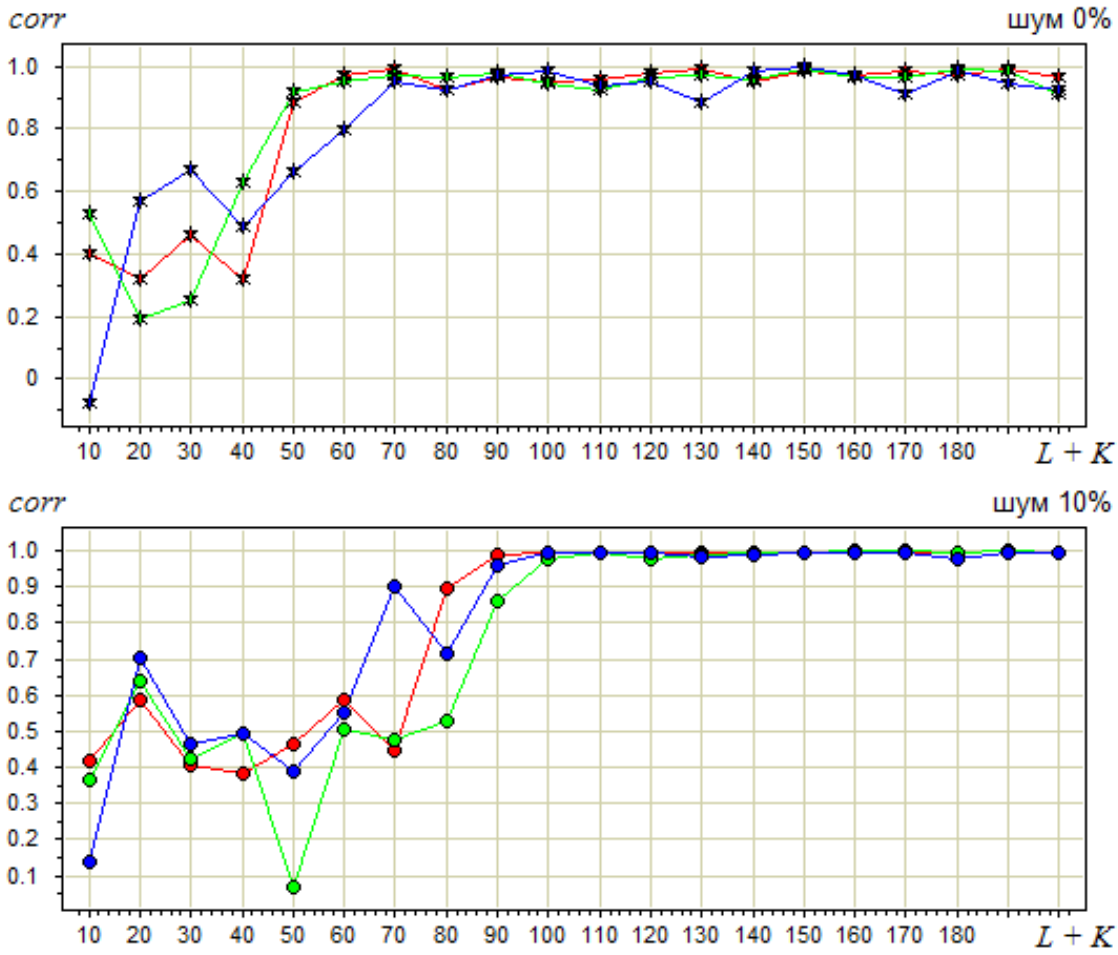

Рис. 1: Зависимость корреляции от зашумленности и длины выборки