УДК 519

Губко М.В.[1], *Лактионова М.М.* [2]

[1] Институт проблем управления им. В.А. Трапезникова РАН

[2]Московский Инженерно-Физический Институт

Алгоритм построения дерева решений

Деревья решений – хорошо зарекомендовавший себя комплекс методов классификации, распознавания и поддержки принятия решений в машинном обучении, идентификации, анализе данных, ситуационном управлении. Дерево решений должно быть компактным – это экономит затраты при ответах на вопросы; кроме того, компактные деревья обладают лучшей прогностической способностью.

Задача построения оптимального дерева решений является NP-полной. В связи с этим на практике используются многочисленные эвристические алгоритмы. Мы предлагаем простой «жадный» алгоритм построения дерева «сверху вниз», основанный на использовании предложенной в [1] нижней оценки стоимости дерева, имеющей, в отличие от известных оценок, комбинаторную природу. Ее вычисление сводится к решению набора непрерывных релаксаций задачи о минимальном покрытии. Эксперименты показывают, что каждая задача решается за время в среднем порядка $n*m$, где n – объем используемой для построения дерева обучающей выборки, а m – количество имеющихся вопросов (тестов). На примере нескольких реальных задач классификации показывается, что предлагаемый алгоритм дает не худшие результаты, чем такие популярные эвристики как CS-ID3, IDX и EG2.

Недостатком используемой нижней оценки является ее относительно высокая вычислительная сложность – порядка n^2m (для каждого из n элементов обучающей выборки за время в среднем порядка $n*m$ решается задача о покрытии). Рассмотрен вариант ее уменьшения за счет вычисления оценки на основе лишь части обучающей выборки. При этом трудоемкость ее вычисления уменьшается до $A*n*m$, где A – максимальное число элементов выборки, используемых для расчета. Полученная огрубленная оценка уже не является гарантированно нижней оценкой, однако для ее применения в алгоритме это не обязательно. В результате для типичного случая $m << n$ оценка средней трудоемкости всего алгоритма имеет порядок не более $A*m^3*n$, то есть

линейно растет по n, что принципиально для практического применения алгоритма. В докладе описываются результаты вычислительных экспериментов по определению влияния параметра A на качество результирующих деревьев.

Перспективы исследований связаны с построением по той же схеме эффективных алгоритмов поиска приближенно оптимальных иерархий для других приложений, в которых имеются нижние оценки затрат иерархии, например, в описанной в [2] модели надстройки иерархии управления над сетью технологических взаимодействий исполнителей. Для этой модели имеется простая нижняя оценка затрат иерархии, однако построение алгоритма требует разработки экономных схем перебора допустимых разбиений для довольно крупных множеств элементов (порядка нескольких тысяч).

Работа частично выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект 10-07-00129).

Литература

1. *Goubko M.V.* Lower-bound Estimate for Cost-sensitive Decision Trees // Preprints of the 18th IFAC World Congress, Milano (Italy), August 28 - September 2, 2011. P. 9005-9010.
2. *Губко М.В., Мишин С.П.* Оптимальная структура системы управления технологическими связями / Материалы международной научной конференции «Современные сложные системы управления». Старый Оскол: СТИ, 2002. С. 50–54.