Об алгоритме классификации на основе полного решающего дерева

Дюкова Е. В., Песков Н. В.
djukova@ccas.ru, peskov@ccas.ru
Москва, ВЦ РАН

Решающие деревья — это один из наиболее популярных инструментов для решения задач классификации по прецедентам в случае, когда исследуемые объекты описываются в признаковом пространстве.

Синтез решающего дерева представляет собой итерационный процесс. Как правило, для построения очередной вершины дерева выбирается признак, наилучшим образом удовлетворяющий некоторому критерию ветвления. По значениям этого признака и осуществляется ветвление, далее указанная процедура повторяется для каждой из ветвей. Описанный подход обуславливает основные достоинства метода, а именно, решающее правило строится быстро, получается достаточно простым и хорошо интерпретируемым.

Однако, в случае, когда два или более признака удовлетворяют рассматриваемому критерию ветвления в равной или почти равной мере, выбор одного из этих признаков осуществляется практически случайным образом. При этом в зависимости от выбранного признака построенные деревья могут существенно отличаться как по составу используемых признаков, так и по своим распознающим качествам.

В докладе рассматривается следующий подход к решению указанной проблемы. При возникновении ситуации, когда два или более признака удовлетворяют критерию в равной мере, предлагается проводить ветвление по каждому из этих признаков независимо. Полученная в результате конструкция названа полным решающим деревом.

Данный подход продемонстрирован на примере усовершенствования алгоритма построения допустимого разбиения (далее АДР). Критерий ветвления в АДР представляет собой набор условий с разным приоритетом. При выборе очередной вершины сначала проверяется условие с наибольшим приоритетом. Ищется признак с наименьшим номером, для которого это условие выполняется. Если ни один признак не удовлетворяет рассматриваемому условию, то проверяется следующее по порядку условие с более низким приоритетом.

Проведенное экспериментальное исследование на реальных задачах показало, что точность распознавания при использовании полного решающего дерева существенно выше точности распознавания АДР. Кроме того, было проведено сравнение новой модели, названной ПРД, с алгоритмом C4.5, широко используемого в промышленных системах анализа

данных. На рассмотренных задачах новая модель ведет себя не хуже алгоритма C4.5, а на некоторых существенно превосходит C4.5.

Работа выполнена при поддержке РФФИ, проекты № 07-01-00516 и № 06-07-89299 и гранта Президента РФ по поддержке ведущих научных школ НШ № 5833.2006.1.

Литература

[1] *Донской В. И., Башта А. И.* Дискретные модели принятия решений при неполной информации. — Симферополь: Таврия, 1992. — С. 33–74.