Построение решающих деревьев минимальной стоимости для попарного сравнения объектов

Майсурадзе А. И.
useraim@mail.ru
Москва, Московский государственный университет

Рассматривается задача построения коллективной функции сходства для пар объектов распознавания при заданном наборе функций сходства. При этом считается, что функции сходства из набора для пар объектов еще не вычислены и имеют определенную стоимость вычисления, а результирующая коллективная функция сходства должна удовлетворять заданным прецедентным ограничениям.

Постановка задачи синтеза функции сходства

При решении многих задач интеллектуального анализа данных нередко предлагается проводить сравнение исследуемых объектов или вводить на них некоторые функции сходства. Во многих прикладных областях такие функции сходства возникают естественным путём, превосходя по качеству попытки признакового описания объектов. В других случаях функции сходства отражают экспертные знания о предметной области. Но практически всегда естественные или экспертные функции сходства невозможно или малоэффективно непосредственно использовать для решения поставленной задачи интеллектуального анализа данных.

В то же время в указанных ситуациях может быть доступна прецедентная информация — известно желаемое значение функции сходства для некоторых пар объектов. В частности, искомая функция сходства может быть формализована как бинарное отношение на объектах, для которого известна истинность или ложность на некоторых парах объектов. Кроме того, иногда на искомую функцию сходства налагаются универсальные ограничения. В рассматриваемом ниже прикладном примере искомая функция сходства должна быть симметричным бинарным отношением, значения которого известны для всех пар объектов из фиксированного списка.

Таким образом, в описанных выше ситуациях может быть поставлена задача синтеза функции сходства по заданному набору естественных или экспертных функций сходства, удовлетворяющей заданным прецедентным и универсальным ограничениям. Указанная задача практически является классической задачей обучения по прецедентам, для решения которой могут быть привлечены многочисленные алгоритмические модели восстановления регрессии или классификации. Существенной особенностью рассматриваемой задачи является тот факт, что, в отличие от тра-

дициионных признаков, вычисление функции сходства из набора во многих прикладных областях является весьма трудоёмкой операцией.

Указанная трудоёмкость создает сложности как на этапе настройки, так и в ходе использования настроенной функции сходства. В данной работе (исходя из содержания прикладного примера) выбор делается в пользу функций, которые в среднем как можно быстрее обрабатывают новые объекты, но могут потребовать существенных затрат времени при первоначальной настройке.

Описание модели на базе решающих деревьев

Большинство стандартных моделей классификации объектов по признаковым описаниям устроены следующим образом: отказ от использования некоторого признака может произойти только в ходе настройки, а при обработке каждого очередного объекта распознавания требуются значения всех оставленных признаков. В отличие от этого, основываясь на приведенной выше формализации задачи, требовалось выбрать такую модель алгоритмов классификации, которая при обработке очередного объекта распознавания запрашивает значения лишь некоторых признаков, причем при обработке разных объектов могут потребоваться разные признаки. Одной из наиболее распространенных и изученных моделей, обладающих указанным свойством, является модель решающих деревьев.

Отметим, что в рассматриваемом в работе подходе «объектом распознавания» является пара исходных объектов. Таким образом, на вход модели подается «признаковое описание» пары исходных объектов — вектор значений экспертных функций сходства из заданного набора. При этом реальное вычисление значений происходит тогда, когда значение действительно потребуется (lazy evaluation). Каждому вычислению экспертной функции сходства можно приписать стоимость (например, пропорционально тродоёмкости вычислений). Традиционные решающие деревья производят ветвление, сравнивая в узле значение только одной экспертной функции сходства с фиксированным порогом. Следовательно, каждой обработке настроенной моделью очередного объекта распознавания можно приписать суммарную стоимость вычисления. (Возможны два подхода: дублировать стоимость повторно вычисляемой экспертной функции сходства, либо кэшировать значения.) Кроме того, стоимость можно приписать ошибкам распознавания на прецедентах (штрафы). Таким образом, при заданном наборе прецедентов (пар исходных объектов) каждое решающее дерево описывается парой стоимостей: затраты на вычисления и штраф за ошибки. Если стоимости заданы в сравнимых единицах измерения, то можно перейти к суммарной стоимости дерева.

Это позволяет поставить задачу поиска решающего дерева минимальной стоимости для заданного набора прецедентов.

К положительным сторонам описанного подхода следует отнести тот факт, что минимизируемый функционал ограничивает рост дерева. Полученное дерево как бы не нуждается в «обрезке», т. к. размер дерева и точность распознавания уже согласованы. К отрицательным сторонам следует отнести высокую трудоёмкость настройки.

Прикладной пример

В работе [1] была рассмотрена задача биометрической идентификации личности по форме ладони. Аппроксимация силуэта ладони многоугольной фигурой и построение её скелета позволили разработать для сравнения ладоней целый ряд функций близости, описываемых полуметриками. Вновь поступающее для идентификации изображение требовалось сравнить со всей базой эталонов (по несколько эталонов для каждого человека). К сожалению, вычисление наилучших по качеству идентификации функций сходства по парам изображений требовало больших затрат времени, неприемлемых в реальных прикладных системах идентификации. Полный же отказ от таких функций приводил к недопустимой деградации качества распознавания. В то же время, правдоподобной выглядела гипотеза, что многие пары можно оценить, пользуясь только функциями сходства, не требующими существенных затрат времени. Предложенный в настоящей работе подход позволит существенно повысить качество идентификации за приемлемое для реальных прикладных систем время.

Работа выполнена при поддержке РФФИ, проект № 07-01-00211-а.

Литература

[1] *Местецкий Л. М., Мехедов И. С.* Комбинированное правило ближайших соседей при классификации формы ладоней // Тез. докл. межд. конф. Интеллектуализация обработки информации (ИОИ-2006). — Симферополь: Крымский научный центр НАН Украины, 2006. — С. 142–144.