Методы быстрого поиска ближайшего аналога в большой базе изображений

Загоруйко Н. Г., Борисова И. А., Дюбанов В. В., Кутненко О. А.
zag@math.nsc.ru
Новосибирск, Институт математики СО РАН

Типичная задача при использовании больших БД состоит в поиске объекта, который является ближайшим аналогом нового (контрольного) объекта. Для решения этой задачи разработано семейство алгоритмов направленного поиска «Локатор» [1], основанных на пошаговом сокращении количества конкурирующих объектов и фокусировании внимания на тех объектах, которые имеют наибольшие шансы стать победителями в этой конкуренции.

Для снижения размерности пространства за счет исключения малоинформативных признаков разработан алгоритм направленного поиска GRAD [2]. Нами предложен новый критерий информативности, основанный на функции конкурентного сходства (FRiS-функции) [3]. Эта же функция применяется и в новом алгоритме кластеризации FRiS-Cluster [4].

Эффективность разработанных алгоритмов исследовалась в модельных тестах и при решении реальных задач. Для проведения исследований была разработана программа LOCATEST [5].

Содержание задач поиска аналога и методов их решения зависит от ответов на следующие вопросы: *Описываются ли образы значениями признаков X или парными расстояниями r между всеми образами? Если решается задача распознавания в признаковом пространстве , то какова метрика этого пространства:* L_∞, L_2 *или* L_1*? Задается ли порог d допустимых различий между объектом Z и аналогом, или нет, т. е. ищется d-аналог или abs-аналог? Требуется ли найти все t d-аналогов или один самый близкий?* Сочетания разных ответов на эти вопросы порождает 12 различных задач поиска аналога. Алгоритмы их решения описаны в [1]. Здесь для примера приведены краткие характеристики некоторых из них.

Для поиска всех аналогов, которые удовлетворяют условию $R_\infty \leqslant d$, используется алгоритм **Локатор-1**. Все K объектов БД и контрольный объект Z проецируются на одну из N координат. Если расстояние от Z до объекта O_i больше d, то i–й объект из списка конкурентов на роль ближайшего вычеркивается. Для оставшихся $k < K$ объектов та же процедура повторяется с использованием проекции на вторую координату, и т. д. Данный алгоритм сокращает время поиска приблизительно в 6 раз.

Если объекты описаны не признаками, а матрицей M парных расстояний между ними, и если требуется найти одного abs-аналога объекту Z, то применяется алгоритм **Локатор-8**. Для исключения тех объектов, которые не могут быть ближайшими аналогами, используется следующее легко доказуемое

Утверждение 1. *Объекты, удаленные от объекта O_i на расстояние $R > 2R(z_i)$, находятся по отношению к объекту Z дальше, чем объект O_i.*

Для выбора следующего локатора по i–й строке матрицы M находим объект O_j, для которого величина $F = |R(z_i) - R(j_i)|,\ j = 1, \ldots, k_l$, минимальна. Знание расстояний от точки Z до двух объектов-локаторов позволяет более точно пеленговать позицию точки Z среди оставшихся объектов. Такие шаги повторяются до тех пор, пока в списке претендентов не останется один объект. На ряде реальных задач затраты времени сокращались на 2–3 порядка.

Процесс поиска ускоряется, если в качестве локаторов выбирать не случайные объекты, а такие, которые находятся в центре локальных сгустков. Для поиска таких локаторов применяется алгоритм кластеризации FRiS-Cluster.

Недостаток существующих *мер сходства* состоит в том, что сходство рассматривается в качестве абсолютной характеристики, в то время как ответы на вопросы типа «близко–далеко?» или «похож–не похож?» зависят от ответа на вопрос «по сравнению с чем или кем?». Это свойство относительных понятий сходства и различия может быть выражено функцией F конкурентного сходства (FRiS-функцией). Если расстояния от объекта Z до двух ближайших объектов a и b равны R_a и R_b, то сходство Z с объектом a равно $F(a) = (R_b - R_a)/(R_a + R_b)$. Значения F меняются в пределах от $+1$ до -1. Если контрольный объект Z совпадает с объектом a, то $R_a = 0$ и $F(a) = 1$, а $F(b) = -1$. При расстояниях $R_a = R_b$ значения $F(a) = F(b) = 0$, что указывает на границу между образами.

При использовании алгоритма FRiS-Cluster на его первых шагах все M объектов заданного множества A принадлежат одному кластеру. В связи с этим вводится конкурирующее множество из виртуальных объектов, удаленных от каждого объекта множества A на расстояние R^*. Произвольный объект O_i множества A назначается центром первого кластера, оцениваются расстояния R_j до него от всех остальных объектов O_j и определяются значения функции $F(i) = (R^* - R_j)/(R^* + R_j)$. Вычисляется сумма значений функций сходства F_s всех объектов первого кластера со своим центром. Затем в качестве центра второго кластера выбирается объект, набравший наибольшее значение F_s в конкурен-

ции с центром первого кластера. Процесс увеличения числа кластеров k останавливается, когда достигается первый локальный максимум функции $F_s(k)$.

Если объекты разделены на классы, то для выбора признаков может быть использован алгоритм GRAD. На его первом этапе создаются вторичные признаки в виде «гранул» — наиболее информативных комбинаций из двух и трёх зависимых признаков. На множестве гранул выполняется последовательность процедур добавления (Addition) наиболее информативных и исключения (Deletion) наименее информативных признаков. Их информативность оценивается по среднему значению функции сходства F_s всех объектов обучающей выборки со своими эталонами. Преимущество этого F_s критерия по сравнению с распространенным критерием U минимума ошибок на обучении подтверждена на большом числе задач. Применение критерия F позволило решить «проблему пригодности признаков», поставленную А. Н. Колмогоровым [6].

В докладе приводятся примеры решения задач выбора ближайшего аналога в базе генетических данных и базе микрофотографий. Показано, что для ускорения поиска ближайшего аналога следует использовать различные средства: направленный перебор претендентов (алгоритмы Локатор), предварительное сокращение размерности пространства (алгоритм GRAD) и предварительный выбор числа сгустков объектов и их центров (алгоритм FRiS-Cluster).

Работа выполнена при поддержке РФФИ, грант № 05-01-00241.

Литература

[1] *Загоруйко Н. Г., Дюбанов В. В.* Семейство алгоритмов «Локатор» для быстрого поиска ближайшего аналога // СибЖИМ. — 2006. — Т. 38, № 5. — С. 54–62.

[2] *Загоруйко Н. Г., Кутненко О. А.* Алгоритм GRAD для выбора признаков // Труды VIII Межд. конф. «Применение многомерного статистического анализа в экономике и оценке качества», Москва: Изд. МЭСИ, 2006. — С. 81–89.

[3] *Загоруйко Н. Г.* Методы интеллектуального анализа данных, основанные на функции конкурентного сходства // Автометрия (в печати).

[4] *Борисова И. А.* Алгоритм таксономии FRiS-Tax // Научный вестник НГТУ (в печати).

[5] `www.dvv-2.gorodok.net`

[6] *Колмогоров А. Н.* К вопросу о пригодности найденных статистическим путем формул прогноза // Завод. лаб. — 1933. — № 1. — С. 164–167.