СТАТИСТИЧЕСКИЕ МЕТОДЫ И АНАЛИЗ ДАННЫХ

Заметки по теории классификации

А.И.Орлов

(Москва)

Проанализированы различные направления в теории классификации. Для этой области дана классификация математических методов. Рассмотрена проблема естественной классификации, обоснован тезис: критерий естественности – устойчивость. В вероятностной теории кластер-анализа предложен метод статистической проверки обоснованности объединения кластеров. Найдено асимптотическое распределение «прогностической силы» алгоритма классифицирования, разработан метод проверки возможности пересчета на модель линейного дискриминантного анализа.

Ключевые слова: методы классификации, объекты нечисловой природы, кластер, естественность классификации, прогностическая сила, устойчивость, проверка гипотез, асимптотические распределения.

При внедрении современных статистических методов в практику социологических исследований, при разработке

соответствующих программных продуктов невозможно обойтись без классификации этих методов. Естественно начать с вида обрабатываемых данных. В [1] предложено делить прикладную статистику на четыре области:

- статистика случайных величин (одномерная статистика);
- многомерный статистический анализ;
- статистика временных рядов и случайных величин;
- статистика объектов нечисловой природы.

В первой области элемент выборки – число, во второй вектор, в третьей – функция, в четвертой – объект нечисловой природы. Термин «объект нечисловой природы» относится к элементам невекторного пространства. Их нельзя складывать, умножать на числа, в отличие от чисел, векторов и функций. Примерами являются бинарные отношения (упорядочения, разбиения на классы, толерантности); множества, нечеткие множества; результаты измерений в номинальной и порядковой шкалах (т.е. по качественным признакам), в частности булевы вектора; тексты и т.д.

Математический аппарат статистики объектов нечисловой природы базируется на использовании расстояний (мер близости, показателей различия) в пространствах таких объектов. Это вызвано отсутствием здесь операций суммирования, на которых основано большинство методов других областей статистики.

Любые методы, использующие только расстояния (меры близости, показатели различия) между объектами, также следует относить к данной области статистики, поскольку они могут работать с объектами произвольного пространства, если в нем задана метрика или ее аналоги.

Таким образом, весьма многие математические методы классификации можно связать со статистикой объектов нечисловой природы.

Работы в этой развитой области прикладной математики (ей посвящено несколько тысяч статей и книг) координирует подкомиссия «Статистика объектов нечисловой при-

роды» Научного Совета АН СССР по комплексной проблеме «Кибернетика». В 1985 г. она совместно с Институтом социологии АН СССР подготовила и опубликовала сборник статей [2], включающий обзоры основных направлений данной области статистики. Отметим статью [3], рассказывающую о статистике объектов нечисловой природы «в целом», и [4], обсуждающую возможность введения в ней различных мер близости. Опубликован еще целый ряд обзоров как общего характера [5,6], так и по отдельным направлениям – статистике разбиений [7], бернуллиевских векторов[1] [8] и другим. Специфические методы анализа летописей описаны в [9]; результаты представляют несомненный интерес для историков, поскольку демонстрируют необходимость пересмотра традиционной хронологии.

В настоящей статье рассматривается важное направление статистики объектов нечисловой природы – методы классификации, основанные на расстояниях между объектами.

1. Направления в теории классификации

Какие научные исследования относить к этой теории? Из прагматических соображений целесообразно принять, что сюда входят исследования, во-первых, отнесенные самими авторами к этой теории; во-вторых, связанные с ней общностью тематики, хотя бы их авторы и не упоминали термин «классификация». Это предполагает ее сложную внутреннюю структуру.

В современной теории классификации можно выделить два относительно самостоятельных направления. Одно из них опирается на опыт таких наук, как биология, геогра-

[1] Бернуллиевский вектор – это вероятностная модель булева вектора с независимыми координатами; причем вероятности того, что на определенном месте стоит 1, вообще говоря, различны.

фия, геология; библиотечное дело также анализировалось в этом направлении. Типичные объекты рассмотрения – классификация химических элементов (таблица Менделеева), биологическая систематика, универсальная десятичная классификация (УДК). Опыт этого направления с гносеологических позиций обобщен в [10], соответствующий математический аппарат в [11,12].

Другое направление опирается на опыт технических исследований, медицины, социологии, экономики. Типичные задачи – техническая и медицинская диагностика, разбиение на группы отраслей промышленности, тесно связанных между собой, выделение групп однородной продукции. Наиболее развитая область – распознавание образов, или дискриминантный анализ [13]. Краткое осмысление опыта и современного состояния этого направления дано в [14], сводка алгоритмов классификации имеется в [15]. Оно опирается на математические модели; для проведения расчетов интенсивно используется ЭВМ. Однако относить его к математике столь же нецелесообразно, как астрономию или квантовую механику. Рассматриваемые математические модели можно и нужно изучать на формальном уровне, и такие исследования проводятся. Но направление в целом сконцентрировано на решении конкретных задач прикладных областей и вносит вклад в технические или экономические науки, медицину, социологию, но, как правило, не в математику. Использование математических методов как инструмента исследования нельзя относить к чистой математике.

2. Математические методы классификации

В 60-х годах XX века внутри прикладной статистики достаточно четко оформилась область, посвященная методам классификации (см., например, [13-15]). Несколько модифицируя формулировки М.Дж.Кендалла и А.Стьюарта 1966 г.

(см. русский перевод [16, с.437]), выделим три подобласти: дискриминация, кластеризация, группировка. В дискриминантном анализе классы предполагаются заданными – плотностями вероятностей или обучающими выборками. Задача состоит в том, чтобы вновь поступающий объект отнести в один из этих классов. У понятия «дискриминация» имеется много синонимов: распознавание образов, диагностика, автоматическая классификация с учителем, статистическая классификация и т.д.

При кластеризации и группировке целью является выявление и выделение классов. Задача кластер-анализа «состоит в выяснении по эмпирическим данным, насколько элементы "группируются" или распадаются на изолированные "скопления", "кластеры"»[16, с.467]. Иными словами, задача – выявление «естественного»разбиения на классы, свободного от субъективизма исследователя, а цель – выделение групп однородных объектов, сходных между собой, при резком отличии этих групп друг от друга.

При группировке, наоборот, «мы хотим разбить элементы на группы независимо от того, естественны ли границы разбиения или нет» [16, с.437]. Цель состоит в выявлении групп однородных объектов, сходных между собой (как в кластер-анализе), однако «соседние» группы могут не иметь резких различий (в отличие от кластер-анализа). Границы между группами условны, зависят от субъективизма исследователя, как при лесоустройстве проведение просек (границ участков) зависит от специалистов лесного ведомства, а не от свойств леса.

Задачи кластеризации и группировки принципиально различны, хотя для их решения могут применяться одни и те же алгоритмы. Проблема состоит в том, чтобы понять, разрешима ли задача кластер-анализа для конкретных данных или возможна только их группировка, поскольку они достаточно однородны и не разбиваются на резко разделяющиеся между собой кластеры.

В задачах кластеризации и группировки основное – метрика, расстояния между объектами, меры близости, сходства, различия. Хорошо известно, что для любого заданного разбиения объектов на группы и любого $\varepsilon>0$ можно указать метрику такую, что расстояния между объектами из одной группы будут меньше ε, а между объектами из разных групп – больше $1/\varepsilon$. Тогда любой разумный алгоритм кластеризации даст именно заданное распределение. Некоторые подходы к выбору расстояния в задачах классификации рассмотрены в [4].

«Термином „классификация" обозначают, по крайней мере, три разные вещи: процедуру построения классификации, построенную классификацию и процедуру ее использования» [10, с.6]. Статистический анализ полученных, в частности экспертами, классификаций – часть статистики бинарных отношений и тем самым – статистики объектов нечисловой природы. Процедура использования классификации, т.е. отнесения вновь поступающего объекта к одному из классов – предмет дискриминантного анализа. Как отмечалось [17], он является частным случаем общей схемы регрессионного анализа. Однако есть ряд специфических постановок.

Часто рекомендуют сначала проводить классификацию, а потом регрессионный анализ (в классическом смысле). Однако при этом нельзя опираться на нормальную модель, так как не будет нормальности в кластерах [18].

В прикладных исследованиях применяют различные методы дискриминантного анализа, основанные на вероятностных моделях (см., например, [19]), а также с ними не связанные, т.е. эвристические, в частности алгоритм «Кора»[20]. Независимо от «происхождения», каждый подобный алгоритм должен быть исследован как на вероятностных моделях, так и на массивах реальных данных. Цель исследования – выбор наилучшего в определенной области применимости, включение его в стандартный пакет программ, учебные пособия. Но для этого надо уметь сравнивать алгоритмы по

качеству. Часто используют показатель «вероятность правильной классификации» (при обработке конкретных данных – «частота правильной классификации»). В [21] показано, что он некорректен. Здесь же предложен другой – оценка «расстояния Махаланобиса» между классами и найдено его асимптотическое распределение (см. ниже).

Процедуры построения классификации делятся на вероятностные и детерминированные. К первым относятся задачи расщепления смесей [15]. Как при выборе степени полинома в регрессии [22,23], встает вопрос оценки числа элементов смеси. Применение критерия Уилкса дает результаты типа [22,23], демонстрирующие несостоятельность обычно используемых оценок [18].

Задачи построения классификации целесообразно разбить на два типа: с четко разделенными кластерами (задачи кластер-анализа) и с условными границами, непрерывно переходящими друг в друга классами (задачи группировки). Такое деление полезно, хотя в обоих случаях могут применяться одинаковые алгоритмы. Их бесконечно много, что нетрудно доказать.

Действительно, рассмотрим алгоритм средней связи [15, с.102]. Он основан на мере близости $d(x, y)$ между объектами x и y. Легко проверить, что величина $d^{\alpha}(x, y)$ также является мерой близости между x и y и порождает новый алгоритм. Если α пробегает отрезок, то получается бесконечно много алгоритмов.

Каким из них пользоваться при обработке данных? Дело осложняется тем, что мер близости различных видов существует весьма много [4]. Именно в связи с обсуждаемой проблемой в [24] рассматривается разделение между кластер-анализом и задачами группировки.

Если классы реальны [18], естественны, то любой алгоритм кластер-анализа их выделит. В [10, с.206-209] приведен ряд критериев естественности классификации, однако их невозможно применить при обработке конкретных данных.

Ниже предлагается в качестве такого критерия рассматривать устойчивость относительно выбора алгоритма кластер-анализа.

Проверить устойчивость можно, применив к данным несколько подходов, например столь не похожие алгоритмы, как «ближнего соседа» и «дальнего соседа» [15]. Если полученные результаты содержательно близки, то они адекватны действительности. В противном случае следует предположить, что естественной классификации не существует, задача кластер-анализа не имеет решения, и можно проводить только группировку.

Часто применяется агломеративный иерархический алгоритм «Дендрограмма» [15, с.100-103], в котором вначале все элементы рассматриваются как отдельные кластеры, а затем на каждом шагу объединяются два наиболее близких кластера.

Для работы «Дендрограммы» необходимо задать правило вычисления расстояния между кластерами. Оно вычисляется через расстояние $d(x, y)$ между элементами x и y. Поскольку $d^{\alpha}(x, y)$ при $0<\alpha<1$ также расстояние, то, как правило, существует бесконечно много различных вариантов этого алгоритма. Представим себе, что они применяются для обработки одних и тех же реальных данных. Если при всех α получается одинаковое разбиение элементов на кластеры, т.е. результат работы алгоритма устойчив по отношению к изменению α [25], то имеем «естественную» классификацию. В противном случае результат зависит от субъективно выбранного исследователем параметра α, т.е. задача кластер-анализа неразрешима (предполагаем, что выбор α нельзя специально обосновать). Задача группировки в этой ситуации имеет много решений. Из них можно выбрать одно по дополнительным критериям.

Следовательно, получаем эвристический критерий: если решение задачи кластер-анализа существует, то оно находится с помощью любого алгоритма. Целесообразно использо-

вать наиболее простой. Так, для классификации социально-психологических характеристик способных к математике школьников [26] мы использовали алгоритм [27]. На программирование и счет на ЭВМ ушло около полугода. Недавно с помощью одного из вариантов «дендрограммы» – алгоритма «ближайшего соседа» [15, с. 101] – кластер-анализ был проведен вручную за 1,5 часа. Результаты практически совпали!

3. Проблема естественной классификации

Существуют различные точки зрения на эту проблему [10, с. 202-213]. На Всесоюзной школе-семинаре «Использование математических методов в задачах классификации» (Пущино, 13-20 апреля 1986 г.), в частности, были высказаны мнения, что естественная классификация:

- закон природы (Г.Б.Бокий);
- основана на глубоких закономерностях, тогда как искусственная классификация – на неглубоких (Б.Г.Миркин);
- для конкретного индивида та, которая наиболее быстро вытекает из его тезауруса (Ю.П.Дробышев);
- это классификация ботинок по размеру (Л.Г.Малиновский);
- удовлетворяет многим целям; цель искусственной классификации задает человек (И.Д.Мандель);
- классификация с точки зрения потребителя продукции (М.Х.Клин);
- классификация, позволяющая делать прогнозы (С.И.Сухонос);
- имеет критерием устойчивость (А.И.Орлов).

Приведенные высказывания[1] и соответствующий материал в [10] уже дают представление о больших расхождениях в понимании «естественной классификации». Этот тер-

[1] Записаны автором со слов выступавших.

мин является нечетким, как, впрочем, и многие другие. В [28] подробно обоснована нечеткость естественного языка и тот факт, что «мы мыслим нечетко», что однако не слишком мешает нам функционировать. Кажущееся рациональным требование выработать строгие определения, а потом развивать науку – невыполнимо. Следовать ему – значит отвлекать силы от реальных задач. При системном подходе (в интерпретации [11])к теории классификации становится ясно, что строгие определения можно надеяться получить на последних этапах построения теории. Мы же сейчас находимся на первых. Поэтому, не давая определения понятию «естественная классификация», обсудим, как проверить на «естественность» классификацию, полученную расчетным путем.

Можно выделить два критерия «естественности», по поводу которых имеется относительное согласие:

А. Естественная классификация должна быть реальной, соответствующей действительному миру, лишенной внесенного исследователем субъективизма;

Б. Естественная классификация должна быть важной или с научной точки зрения (давать возможность прогноза, предсказания новых свойств, сжатия информации и т.д.), или с практической.

Пусть классификация проводится на основе информации об объектах, представленной в виде матрицы «объект-признак» или матрицы попарных расстояний (мер близости). Пусть алгоритм классификации дал разбиение на кластеры. Как можно получить доводы в пользу естественности этой классификации? Например, уверенность в том, что она – закон природы, может появиться только в результате длительного ее изучения и практического применения. Это соображение относится и к другим из перечисленных в [10] и выше критериев, в частности к Б (важности). Сосредоточимся на критерии А (реальности).

Понятие «реальности» кластера требует специальног обсуждения (оно начато в работе [18]). Рассмотрим существо

различий между понятиями «классификация» и «группировка». Пусть, к примеру, необходимо деревья, растущие в определенной местности, разбить на группы находящихся рядом друг с другом. Ясна интуитивная разница между несколькими отдельными рощами, далеко отстоящими друг от друга и разделенными полями, и сплошным лесом, разбитым просеками на квадраты с целью лесоустройства. Однако формально определить эту разницу столь же сложно, как определить понятие «куча зерен», чем занимались еще в Древней Греции [28].

Переформулируем сказанное в терминах «кластер-анализа» и «методов группировки». Выделенные с помощью первого подхода кластеры реальны, а потому могут рассматриваться как кандидаты в «естественные». Группировка дает «искусственные» классы, которые не могут быть «естественными».

Выборку из унимодального распределения можно, видимо, рассматривать как «естественный», «реальный» кластер. Применим к ней какой-либо алгоритм классификации («Форель», «ближнего соседа» и т.п. [15]). Он даст разбиение на классы, которые, разумеется, не являются «реальными», поскольку отражают прежде всего свойства алгоритма. Как отличить такую ситуацию от противоположной, когда имеются реальные кластеры и алгоритм классификации более или менее точно их выделяет? Критерий истины – практика, но это длительный процесс. Поэтому представляет интерес критерий, оценивающий «реальность» выделяемых с помощью алгоритма классификации кластеров одновременно с его применением.

Такой показатель существует – это критерий устойчивости. Устойчивость – понятие широкое [25]. Поскольку значения признаков всегда измеряются с погрешностями, то «реальное» разбиение должно быть устойчиво (т.е. не меняться или меняться слабо) при малых отклонениях исходных данных. Алгоритмов классификации существует бесконечно

много, и «реальное» разбиение должно быть устойчиво по отношению к переходу к другому алгоритму. Другими словами, если «реальное» разбиение возможно, то оно находится с помощью любого алгоритма автоматической классификации. Следовательно, критерием естественности классификации может служить совпадение результатов работы двух достаточно различающихся алгоритмов, например «ближнего соседа» и «дальнего соседа» [15].

Нами рассмотрены[1] два типа «глобальных» критериев «естественности классификации», касающихся разбиения в целом. «Локальные» критерии относятся к отдельным кластерам. Простейшая постановка такова: достаточно ли однородны два кластера (две совокупности) для их объединения? Если оно возможно, то кластеры не являются «естественными». Преимущество этой постановки в том, что она допускает применение статистических критериев однородности двух выборок. В одномерном случае (классификация по одному признаку) разработано большое число подобных критериев – Смирнова, омега-квадрат (Лемана-Розенблатта), Вилкоксона, Ван-дер-Вардена, Стьюдента и др. [29]. Имеются критерии и для многомерных данных [24]. Для одного из видов объектов нечисловой природы [14] – люсианов – статистические методы выделения «реальных» кластеров развиты в [30].

Что касается глобальных критериев, то для изучения устойчивости по отношению к малым отклонениям исходных данных естественно использовать метод статистических испытаний и проводить расчеты по «возмущенным» данным. Некоторые теоретические утверждения, касающиеся влияния «возмущений» на кластеры различных типов, приведены в [18].

Несколько алгоритмов классификации были применены нами к данным [26]: алгоритмы «ближайшего соседа»,

[1] Программная реализация осуществлена в разработанных под научным руководством автора пакетах ППАНД и ДИСАН.

«дальнего соседа» и Куперштоха-Миркина-Трофимова [27]. С содержательной точки зрения, полученные разбиения отличались мало[1] . Поэтому есть основания считать, что с помощью этих алгоритмов действительно выявлена «реальная» структура данных.

Идея устойчивости как критерия «реальности» иногда реализуется неадекватно. Так, для однопараметрических алгоритмов в [31] предлагается выделять разбиения, которым соответствуют наибольшие интервалы устойчивости по параметру, т.е. наибольшие приращения параметра между очередными объединениями кластеров. Для данных [26] это предложение не дало полезных результатов – были получены различные разбиения: три алгоритма – три разбиения. И с теоретической точки зрения предложение [31] несостоятельно, что нетрудно показать.

Рассмотрим алгоритм «ближайшего соседа», использующий меру близости $d(x, y)$, и однопараметрическое семейство алгоритмов с мерой близости $d^{\alpha}(x, y)$, $\alpha>0$, также являющихся алгоритмами «ближайшего соседа». Тогда дендрограммы, полученные с помощью этих алгоритмов, совпадают при всех α, поскольку при их реализации происходит лишь сравнение мер близости между объектами. Другими словами, дендрограмма, полученная с помощью алгоритма «ближайшего соседа», является адекватной в порядковой шкале (измерения меры близости $d(x, y)$), т.е. сохраняется при любом строго возрастающем преобразовании этой меры [25]. Однако выделенные по методу [31] «устойчивые разбиения» меняются. В частности, при достаточно большом α «наиболее объективным» по [31] будет разбиение на два кластера! Таким об-

---

1 Фактически анализировались иерархические деревья разбиений, поскольку все три алгоритма включали одномерные параметры, смысл которых – расстояние между объединяемыми на очередном шагу кластерами.

разом, разбиение, выдвинутое в [31] как «устойчивое», на самом деле оказывается весьма неустойчивым.

4. Вероятностная теория кластер-анализа

Как и для прочих методов прикладной статистики, свойства алгоритмов кластер-анализа необходимо изучать на вероятностных моделях [18]. Это касается, например, условий естественного объединения двух кластеров. Робастная процедура проверки допустимости объединения предложена в [32], непараметрическая – в [24].

Вероятностные постановки нужно применять, в частности, при перенесении результатов, полученных по выборке, на генеральную совокупность [1]. Вероятностная теория кластер-анализа и методов группировки различна для исходных данных типа таблиц «объект × признак» и матриц сходства. Для первых параметрическая теория называется «расщеплением смесей» [15, гл.2]. Непараметрическая теория основана на непараметрических оценках плотностей вероятностей и их мод [15, 17, 33].

Если исходные данные – матрица сходства $\|d(x, y)\|$, то необходимо признать, что развитой вероятностно-статистической теории пока нет. Подходы к ее построению обсуждались в [18]. Одна из основных проблем – проверка «реальности» кластера (ср.[34]). Предположим, что результаты наблюдений можно рассматривать как выборку из некоторого распределения с монотонно убывающей плотностью при увеличении расстояния от некоторого центра. Примененный к подобным данным какой-либо алгоритм кластер-анализа порождает некоторое разбиение. Ясно, что оно – чисто формальное: выделенным таксонам не соответствуют никакие «реальные» классы. Другими словами, задача кластер-анализа не имеет решения, а алгоритм дает лишь группировку. При об-

работке реальных данных мы не знаем вида плотности. Проблема состоит в том, чтобы определить результат работы алгоритма (реальные кластеры или формальные группы)[1].

Частный случай этой проблемы – проверка обоснованности объединения двух кластеров $\{a_1, a_2, ..., a_k\}$ и $\{b_1, b_2, ..., b_m\}$, например при использовании «Дендрограммы». Ряд авторов высказывали следующую идею. Пусть есть две совокупности мер близости: внутри кластеров $d(a_i, a_j)$, $1 \le i < j \le k$, $d(b_\alpha, b_\beta)$, $1 \le \alpha < \beta \le m$, и между кластерами $d(a_i, b_\alpha)$, $1 \le i \le k$, $1 \le \alpha \le m$. Эти совокупности предлагается рассматривать как независимые выборки и проверять гипотезу о совпадении их функций распределения. Если гипотеза не отвергается, объединение кластеров считается обоснованным; в противном случае – объединять нельзя, алгоритм прекращает работу. В [35] для проверки однородности использовался критерий Вилкоксона U, а в [36] – Лемана-Розенблатта типа ω^2.

В рассматриваемом подходе есть две некорректности. Во-первых, меры близости не являются независимыми случайными величинами. Во-вторых, не учитывается, что объединяются кластеры не произвольные фиксированные, а полученные в результате работы алгоритма, и их состав оказывается случайным [18, разд. 4]. От первой из этих некорректностей можно частично избавиться.

Теорема 1 [24]. *Пусть $a_1, a_2, ..., a_k, b_1, b_2, ..., b_m$ – независимые одинаково распределенные случайные величины (со значениями в произвольном пространстве). Пусть случайная величина $d(a_1, a_2)$ имеет все моменты. Тогда при $k, m \to \infty$ распределение статистики*

$$\frac{8\sqrt{3}U - 3(k+m)(k+m-1)\,(k(k-1)+m(m-1))}{2(k+m)\sqrt{km(k^2+m^2)}},$$

[1] Подробнее см. [18].

где U – сумма рангов элементов первой выборки в объединенной; первая выборка составлена из внутрикластерных расстояний (мер близости) $d(a_i, a_j)$, $1 \le i < j \le k$ и $d(b_\alpha, b_\beta)$, $1 \le \alpha < \beta \le m$, а вторая – из межкластерных $d(a_i, b_\alpha)$, $1 \le i \le k$, $1 \le \alpha \le m$;

сходится к нормальному распределению с математическим ожиданием 0 и дисперсией 1.

5. О сравнении алгоритмов классифицирования по результатам обработки реальных данных

Изложение ведется на примере социологического прогноза (ср.[21]).

При построении социологической информационно-исследовательской системы (СИИС) прогноза исходов голосований возникает задача сравнения прогностических правил «по силе». Прогностическое правило – это алгоритм, позволяющий по характеристикам депутата или избирателя прогнозировать результат голосования. Если прогноз дихотомичен («за» или «против»), то правило является алгоритмом классификации (в другой терминологии – классифицирования), при котором избиратель относится к одному из двух классов – согласно прогнозу голосующих «за» либо «против». Прогностические правила могут быть извлечены из социологической литературы и практики. Каждое из них обычно формулируется в терминах небольшого числа признаков, но наборы признаков сильно меняются от правила к правилу. Поскольку в СИИС должно фиксироваться лишь ограниченное число признаков, то возникает проблема их отбора. Естественно отбирать лишь те их них, которые входят в наборы, дающие наиболее «надежные» прогнозы. Для придания точного смысла термину «надежный» необходимо иметь способ

сравнения алгоритмов классификации по прогностической «силе».

Результаты обработки реальных данных с помощью алгоритма классификации в рассматриваемом случае двух классов описываются долями: правильной классификации в первом классе æ; правильной классификации во втором классе λ; классов в объединенной совокупности π_i, $i = 1,2$; $\pi_1 + \pi_2 = 1$.

Величины æ, λ, π_1, π_2 определяются ретроспективно.

Нередко как показатель качества алгоритма классификации (прогностической «силы») используют долю правильной классификации

$$\mu = \pi_1 æ + \pi_2 \lambda.$$

Однако μ определяется через характеристики π_1, π_2, частично заданные исследователем (например, на них влияет тактика отбора выборки респондентов). В аналогичной медицинской задаче величина μ оказалась больше для тривиального прогноза (у всех больных течение заболевания будет благоприятно), чем для использованного в [20] алгоритма, применение которого с медицинской точки зрения вполне оправдано. Поэтому μ нецелесообразно использовать как показатель качества алгоритма классифицирования.

Применение теории статистических решений требует знания потерь от ошибочной классификации, а в большинстве социологических задач определить потери принципиально невозможно.

Для выявления информативного набора признаков целесообразно использовать метод пересчета на модель линейного дискриминантного анализа [21], согласно которому статистической оценкой прогностической «силы» является

$$\delta^* = \Phi\,(d^*/2),\ d^* = \Phi^{-1}(æ) + \Phi^{-1}(\lambda),$$

где Ф – функция нормального распределения вероятностей с математическим ожиданием 0 и дисперсией 1, а Φ^{-1} – обратная ей функция.

Если классы описываются выборками из многомерных нормальных совокупностей с одинаковыми матрицами ковариаций, а для классификации применяется линейный дискриминантный анализ Р.Фишера, величина d^* представляет собой статистическую оценку расстояния Махаланобиса между совокупностями, независимо от порогового значения, определяющего конкретное решающее правило [37]. В общем случае показатель δ^* вводится как эвристический.

Пусть алгоритм классификации применяется к совокупности, состоящей из m объектов первого класса и n объектов второго класса.

Теорема 2. *Пусть* $m, n \to \infty$. *Тогда для всех* x

$$P\left\{\frac{\delta^* - \delta}{A(æ, \lambda)} < x\right\} \to \Phi(x),$$

где δ – *прогностическая «сила» алгоритма классификации;*

$$A^2(æ,\lambda) = \frac{1}{4}\left\{\left[\frac{\phi(d^*/2)}{\phi(\Phi^{-1}(æ))}\right]^2 \frac{æ(1-æ)}{m} + \left[\frac{\phi(d^*/2)}{\phi(\Phi^{-1}(\lambda))}\right]^2 \frac{\lambda(1-\lambda)}{n}\right\};$$

$\phi(x) = \Phi'(x)$ – *плотность стандартного нормального распределения вероятностей.*

С помощью теоремы 2 по æ и λ обычным образом определяют доверительные границы для прогностической «силы» δ.

Как проверить обоснованность пересчета на модель линейного дискриминантного анализа? Допустим, что классификация состоит в вычислении прогностического индекса y и сравнении его с некоторым порогом c; объект относят к первому классу, если $y \le c$, ко второму, если $y > c$. Возьмем два значения порога c_1 и c_2. Если пересчет на модель обоснован,

то прогностические «силы» для обоих правил совпадают: $\delta(c_1) = \delta(c_2)$. Эту статистическую гипотезу можно проверить.

Пусть $æ_1$ – доля объектов первого класса, для которых $y \le c_1$, а $æ_2$ – доля объектов второго класса, для которых $c_1 < y \le c_2$. Аналогично пусть λ_2 – доля объектов второго класса, для которых $c_1 < y \le c_2$, а λ_3 – для которых $y > c_2$. Тогда оценки расстояний Махаланобиса имеют вид

$$d^*(c_1) = \Phi^{-1}(æ_1) + \Phi^{-1}(\lambda_2 + \lambda_3),$$
$$d^*(c_2) = \Phi^{-1}(æ_1 + æ_2) + \Phi^{-1}(\lambda_3).$$

Теорема 3. *Если* $m,n \to \infty$, $\delta(c_1) = \delta(c_2)$, *то при всех* x

$$P\left\{\frac{d^*(c_1) - d^*(c_2)}{B} < x\right\} \to \Phi(x),$$

где $B^2 = \frac{1}{m}T(æ_1; æ_2) + \frac{1}{n}T(\lambda_3; \lambda_2)$;

$$T(x;y) = \frac{x(1-x)}{\phi^2(\Phi^{-1}(x))} + \frac{(x+y)(1-x-y)}{\phi^2(\Phi^{-1}(x+y))} - \frac{2x(1-x-y)}{\phi(\Phi^{-1}(x))\,\phi(\Phi^{-1}(x+y))}.$$

Из теоремы 3 вытекает метод проверки рассматриваемой гипотезы: при выполнении неравенства

$$\left|\frac{d^*(c_1) - d^*(c_2)}{B}\right| \le \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

она принимается на уровне значимости, приблизительно равной α, в противном случае – отвергается.

Теоремы 1-3 получены автором в 1985 г. [21, 24]. Позже статистическими методами проверки качества классификации занимались Е.В.Кулинская и Г.А.Сатаров [38-40].

Ряд иных задач классификации рассмотрен автором в [18, 33].

Литература

1. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики. М.: ВНИИСИ, 1987.
2. Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985.
3. *Орлов А.И.* Общий взгляд на статистику объектов нечисловой природы //[2].
4. *Раушенбах Г.В.* Меры близости и сходства //[2].
5. *Орлов А.И.* Статистика объектов нечисловой природы //Статистика. Вероятность. Экономика. М.: Наука, 1985.
6. *Орлов А.И.* Статистика объектов нечисловой природы //Тр. I Всемирного Конгресса Общества им. Бернулли «Математическая статистика, теория вероятностей, комбинаторика и ее применения». Вып. 1. М.: МИАН СССР, Советский Комитет Общества им. Бернулли, 1988.
7. *Маамяги А.В.* Некоторые задачи статистического анализа классификаций. Таллинн: Изд-во АН ЭССР, 1982.
8. *Орлов А.И., Рыданова Г.В.* О некоторых результатах статистики объектов нечисловой природы //Материалы I Всесоюз.школы-семинара «Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях (3-6 июня 1985 г., Пущино)». Пущино: НИВЦ АН СССР, 1986.
9. *Фоменко А.Т.* Новая эмпирико-статистическая методика обнаружения параллелизмов в датировании дубликатов //Проблемы устойчивости стохастических моделей. М.: ВНИИСИ, 1984.
10. *Розова С.С.* Классификационная проблема в современной науке. Новосибирск: Наука, 1986.
11. *Шрейдер Ю.А., Шаров А.А.* Системы и модели. М.: Радио и связь, 1982.
12. *Воронин Ю.А.* Теория классифицирования и ее приложения. Новосибирск: Наука, 1985.

13. *Горелик А.Л., Скрипкин В.А.* Методы распознавания. Учебное пособие для вузов. М.: Высшая школа, 1984.
14. *Орлов А.И.* Математические методы классификации, статистика объектов нечисловой природы и медико-биологические исследования //Доклады МОИП 1984 г. Общая биология. Цитогенетический и математический подход к изучению биосистем. М.: Наука, 1986.
15. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. М.: Статистика, 1974.
16. *Кендалл М.Дж., Стьюарт А.* Многомерный статистический анализ и временные ряды. М.: Наука, 1976.
17. *Орлов А.И.* Некоторые неклассические постановки в регрессионном анализе и теории классификации //Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях. М.: Наука, 1987.
18. *Орлов А.И.* Некоторые вероятностные вопросы теории классификации //Прикладная статистика. М.: Наука, 1983.
19. *Шорников Б.С.* Классификация и диагностика в биологическом эксперименте. Проблема оценки и классификации интерьерных признаков человека. М.: Наука, 1979.
20. *Гельфанд И.М., Алексеевская М.А., Губерман Ш.А., Сыркин А.Л., Головня Л.Д., Извекова М.А.* Прогнозирование исхода инфаркта миокарда с помощью программы «Кора-3» //Кардиология. 1977. Т.17. № 6, 7.
21. *Орлов А.И.* О сравнении алгоритмов классификации по результатам обработки реальных данных //Общая биология. Новые данные исследований структуры и функций биологических систем. Доклады МОИП, 1985. М.: Наука, 1987.
22. *Орлов А.И.* Оценка размерности модели в регрессии //Алгоритмическое и программное обеспечение прикладного статистического анализа. М.: Наука, 1980.

23. *Орлов А.И.* Асимптотика некоторых оценок размерности модели в регрессии //Прикладная статистика. М.: Наука, 1983.
24. *Орлов А.И.* Некоторые вероятностные вопросы кластер-анализа //Общая биология. Новые данные исследований структуры и функций биологических систем. Доклады МОИП, 1985. М.: Наука, 1987.
25. *Орлов А.И.* Устойчивость в социально-экономических моделях. М.: Наука, 1979.
26. *Орлов А.И., Гусейнов Г.А.* Математические методы в изучении способных к математике школьников //Исследования по вероятностно-статистическому моделированию реальных систем. М.: ЦЭМИ АН СССР, 1977.
27. *Куперштох В.Л., Миркин Б.Г., Трофимов В.А.* Сумма внутренних связей как показатель качества классификации //А и Т. 1976. № 3.
28. *Орлов А.И.* Математика нечеткости //Наука и жизнь. 1982. № 7.
29. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистика. М.: Наука, 1983.
30. *Орлов А.И.* Парные сравнения в асимптотике Колмогорова //Экспертные оценки в задачах управления. М.: ИПУ, 1982.
31. *Плоткин А.А.* Устойчивость разбиения как критерий оптимальности построенной классификации //Статистические методы анализа экспертных оценок. М.: Наука, 1977.
32. *Шурыгин А.М.* Статистический кластер-критерий //Алгоритмическое и программное обеспечение прикладного статистического анализа. М.: Наука, 1980.
33. *Орлов А.И.* Классификация объектов нечисловой природы на основе непараметрических оценок плотности //Проблемы компьютерного анализа данных и моделирования. Минск: Белорусск. Гос. ун-т, 1991.

34. *Любищев А.А.* Проблемы формы, систематики и эволюции организмов. М.: Наука, 1982.
35. *Райская Н.Н., Гостилин Н.Н., Френкель А.А.* Об одном способе проверки обоснованности разбиения в кластерном анализе //Всесоюз.конф. «Применение многомерного статистического анализа в экономике и оценке качества продукции». Тез. докл. Тарту, 1977.
36. *Бала Ю.М., Фуки В.Б., Рог А.И., Савченко Т.И., Савченко А.В.* О возможности автоматизации процесса дифференциальной диагностики атеросклеротического кардиосклероза и ревматических пороков сердца, осложненных мерцательной аритмией //Кардиология, 1977. Т.17. №7.
37. *Орлов А.И.* Махаланобиса расстояние //Математическая энциклопедия. Т.3. М.: Советская Энциклопедия, 1982.
38. *Кулинская Е.В.* Об эмпирических индексах качества классификации и их реализации в пакете программ CLAMS для IBM PC//Всесоюз. симпозиум с международным участием «Теория и практика классификации и систематики в народном хозяйстве». Тез.докл. М.: ВИНИТИ, 1990.
39. *Кулинская Е.В., Сатаров Г.А.* Проверка гипотез о качестве классификации в пакете программ CLAMS для IBM PC //Всесоюз. симпозиум с международным участием «Теория и практика классификации и систематики в народном хозяйстве». Тез. докл. М: ВИНИТИ, 1990.
40. *Kulinskaya E.V., Satarov G.A.* Testing the Hypotesis about the Quality of Classification in the Data-Analysis System Clams //Statistical Data Analysis. Abstracts of International Conference. Sofia, 1990.