УДК 62-50

АЛГОРИТМЫ ПОСТРОЕНИЯ ХОРОШО ИНТЕРПРЕТИРУЕМЫХ КЛАССИФИКАЦИЙ[1]

А. А. Дорофеюк, А. Л. Чернявский

Институт проблем управления им. В. А. Трапезникова РАН, г. Москва

Предложен способ задания многомерной классификации, делающий ее более понятной специалисту-предметнику. Рассмотрены алгоритмы построения заданных таким образом классификаций.

ВВЕДЕНИЕ

Классификационный подход к анализу данных широко применяется для построения «сжатого описания» исходных данных [1], которое в последующем можно использовать, например, для принятия управленческих решений.

Сжатие информации достигается тем, что совокупность объектов разбивается на классы, а каждый класс представляется одним «синтетическим» объектом — центром класса и набором стандартных отклонений, характеризующих разброс объектов этого класса относительно центра. При небольшом числе классов такое описание оказывается достаточно экономным.

Однако для возможности использования результатов классификации в практических задачах важно не только то, насколько экономно эта классификация представляет исходную информацию, но и то, насколько она удобна для интерпретации в содержательных терминах. Представление в виде набора центров классов и стандартных отклонений удобно для компьютера, но мало что говорит специалисту-предметнику. До сих пор проблеме интерпретируемости классификаций уделялось недостаточное внимание. Отчасти это связано с тем, что в практических задачах классификация часто использовалась как вспомогательное средство для построения регрессионной зависимости (кусочная аппроксимация), когда интерпретируемость классификации не так существенна [2]. Работа посвящена описанию двух алгоритмов построения хорошо интерпретируемой классификации — покоординатной и спрямляющей классификаций.

В обоих алгоритмах множество значений каждого показателя разбивается на небольшое число диапазонов. Хорошо интерпретируемая классификация задается набором границ этих диапазонов. Как только границы диапазонов определены, каждый объект получает описание в виде позиционного кода, где число позиций равно числу показателей. Классом хорошо интерпретируемой классификации является совокупность объектов с идентичными кодами.

Покоординатная классификация осуществляет разбиение каждого показателя на заданное число диапазонов независимо от других показателей, т. е. искомая классификация задается сочетанием одномерных классификаций по всем показателям. Основное преимущество такой классификации — возможность использования алгоритмов оптимальной одномерной классификации [2].

Спрямляющая классификация строится как своего рода аппроксимация многомерной классификации, например, ищутся такие границы диапазонов, при которых число объектов, оказавшихся по другую сторону границы от большинства объектов своего класса, будет минимальным.

1. АЛГОРИТМЫ ПОСТРОЕНИЯ ХОРОШО ИНТЕРПРЕТИРУЕМЫХ КЛАССИФИКАЦИЙ

Для построения хорошо интерпретируемой классификации множество значений каждого используемого для классификации показателя разбивается на небольшое число диапазонов. Так, например, при разбиении на три диапазона они интерпретируются как диапазоны *высоких*, *средних* и *низких* значений показателя. Таким образом, хорошо интерпретируемая классификация задается набором границ диапазонов значений каждого показателя, участвующего в классификации. Тогда задача сводится к нахождению таких границ диапазонов параметров, которые отражают реальную структуру взаиморасположения объектов в многомерном пространстве показателей.

Как только границы диапазонов найдены, каждый объект получает описание в виде позиционного кода, имеющего вид (P_1, P_2, ..., P_k), где число позиций k равно числу используемых для классификации показателей. Такой код означает, что этот объект принадлежит классу с номером P_1 для первого показателя, P_2 — второго, и т. д. Классом хорошо интерпретируемой классификации является совокупность объектов с идентичными описаниями. Разработаны два алгоритма построения хорошо интерпретируемой классификации.

[1] Работа рекомендована к печати Программным комитетом Третьей международной конференции по проблемам управления (Москва, 20—22 июня 2006 г.).

Покоординатная классификация. Очевидно, самый простой способ построения хорошо интерпретируемой классификации — независимое (от других показателей) разбиение множества значений каждого показателя на заданное число диапазонов. В этом случае хорошо интерпретируемая классификация задается как сочетание одномерных классификаций по каждому показателю. Такую классификацию будем называть *покоординатной классификацией*.

Основное преимущество такой классификации — возможность использования алгоритмов оптимальной одномерной классификации, позволяющих получать глобальный оптимум критерия качества классификации [2]. Другие важные преимущества покоординатной классификации состоят в ее простоте и наглядности. Отметим также, что если число диапазонов по каждому показателю задано, то довольно сложная для многомерной классификации проблема выбора числа классов отпадает. Пусть, например, построена трехмерная покоординатная классификация с тремя диапазонами по каждому показателю. Это означает, что пространство объектов оказывается разбитым на $3^3 = 27$ областей, а число классов r будет таким, каким оно фактически «получится», т. е. $r = 27 - r_0$, где r_0 — число областей, в которых не оказалось ни одного объекта.

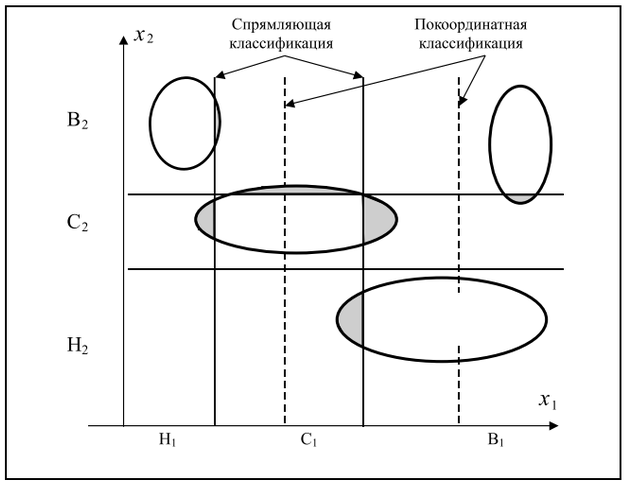
Спрямляющая классификация. Недостаток покоординатной классификации заключается в том, что она строится по одномерной проекции многомерного распределения объектов в пространстве показателей, т. е. не учитывает большую часть информации о структуре множества объектов. Второй способ построения хорошо интерпретируемой классификации позволяет устранить этот недостаток.

При этом способе основой для хорошо интерпретируемой классификации служит многомерная классификация [1]. Хорошо интерпретируемая классификация строится как своего рода аппроксимация этой многомерной классификации.

Поскольку задача состоит в том, чтобы хорошо интерпретируемая классификация наилучшим образом аппроксимировала многомерную классификацию, будем искать такие границы диапазонов, при которых число «отсеченных» объектов, т. е. объектов, оказавшихся по другую сторону границы от большинства объектов своего класса, будет минимальным.

Очевидно, что такие границы легко находятся простым перебором. Описания «отсеченных» объектов, оказавшихся по другую сторону границы диапазона, изменяются. Новое описание может совпасть с описанием одного из имеющихся классов (в этом случае будем говорить, что объект «переходит» в другой класс). В противном случае «отсеченных» объект образует новый класс. Построенную таким образом классификацию будем называть *спрямляющей классификацией*. Число возникших при построении спрямляющей классификации новых классов и число точек в них могут служить характеристикой качества аппроксимации исходной многомерной классификации спрямляющей классификацией.

Заметим, что число классов в спрямляющей классификации, как и в случае покоординатной классификации, в основном определяется фактической структурой точек в многомерном пространстве.



Хорошо интерпретируемые классификации

2. СРАВНЕНИЕ ПОКООРДИНАТНОЙ И СПРЯМЛЯЮЩЕЙ КЛАССИФИКАЦИЙ

В общем случае покоординатная и спрямляющая классификации даже при одинаковом числе диапазонов разбиения показателей могут давать разные результаты. Покажем это на следующем модельном примере. На рисунке овалами условно изображены классы двумерных объектов с равномерным распределением объектов внутри классов. Множество значений каждого из двух показателей x_1 и x_2 разбивается на три диапазона — с высокими (В), средними (С) и низкими (Н) значениями показателя.

В этом примере покоординатная классификация по показателю x_1 с разбиением на три диапазона даст разбиение, показанное штриховыми вертикальными линиями, а спрямляющая классификация — разбиение, показанное сплошными вертикальными линиями (залитые области — это области отсеченных объектов, классификация которых изменится по сравнению с исходной двумерной классификацией). Заранее нельзя сказать, какое из этих разбиений предпочтительнее, это зависит от содержательной постановки задачи.

В заключение заметим, что в прикладных работах в основном используется спрямляющая классификация. Описанные алгоритмы применялись при анализе и совершенствовании систем управления региональным здравоохранением и региональными пассажирскими автоперевозками, а также в задаче оценки и прогнозирования социального развития регионов России.

ЛИТЕРАТУРА

1. *Бауман Е. В., Дорофеюк А. А.* Классификационный анализ данных / Тр. Междунар. конф. по проблемам управления. — М.: СИНТЕГ, 1999. — Т. 1. — С. 62—77.
2. *Дорофеюк А. А., Бауман Е. В., Корнилов Г. В.* Алгоритмы оптимальной кусочно-линейной аппроксимации сложных зависимостей // Автоматика и телемеханика. — 2004. — № 10.

☎ *(495) 334-90-70, e-mail: adorof@ipu.ru*

Статья представлена к публикации членом редколлегии А. С. Манделем. ❑