

Интерактивная процедура снижения размерности признакового пространства в задачах многокритериальной классификации *

А. Б. Петровский¹, Г. В. Ройзензон²

Представлена интерактивная процедура последовательного снижения размерности признакового пространства, которая позволяет упростить порядковую классификацию многокритериальных альтернатив и уменьшить трудоёмкость её построения.

Введение

Одной из основных задач теории принятия решений является разделение множества альтернатив на несколько заранее заданных упорядоченных групп (классов решений). В реальных ситуациях альтернативы описываются многими разнообразными признаками, и число таких признаков может быть достаточно велико (десятки и сотни). Непосредственная классификация альтернатив, описываемых большим числом признаков, на основе предпочтений лица, принимающего решение (ЛПР), является трудоёмкой процедурой, которая требует существенных временных затрат ЛПР и разработки специальных методов его опроса. При решении задачи классификации большой размерности ЛПР зачастую применяет различные упрощённые стратегии с использованием только части критериев, что отрицательно сказывается на построении границ классов решений, выработке решающих правил, дальнейшем анализе полученных результатов [Ларищев, 2006].

Дополнительные трудности появляются в случае слабо структурируемых проблем, сочетающих количественные и качественные зависимости, для которых построение объективных моделей либо невозможно в принци-

* Работа поддержана программами фундаментальных исследований президиума РАН «Фундаментальные проблемы информатики и информационных технологий» и ОНИТ РАН «Фундаментальные основы информационных технологий и систем», Российским фондом фундаментальных исследований (проекты 06-07-89352, 07-01-00515, 07-07-13546, 08-01-00247).

¹ 117312, Москва, проспект 60-лет Октября, 9, ИСА РАН, pab@isa.ru

² 117312, Москва, проспект 60-лет Октября, 9, ИСА РАН, rgv@isa.ru

пе, либо сопряжено со значительными сложностями. Примерами таких задач могут служить конкурсный отбор научных проектов, оцененных несколькими экспертами по многим качественным критериям [Петровский и др., 1990], оценка банковских кредитов по степени риска [Асанов и др., 2001а].

Снижение размерности признакового пространства позволяет упростить решение задач порядковой классификации. В работе предлагается новый подход к сравнению и классификации многопризнаковых объектов по их свойствам, в котором большое число исходных характеристик альтернатив последовательно агрегируется в небольшое число критериев, имеющих небольшие шкалы оценок, отражающих предпочтения ЛПР. Процедура агрегирования признаков использует различные методы вербального анализа решений и имеет блочный характер, за счет чего существенно сокращается трудоемкость построения решающих правил классификации и появляется возможность объяснения полученного результата.

1. Порядковая классификация многокритериальных альтернатив

Задача многокритериальной порядковой классификации формулируется следующим образом. Задано множество альтернатив A_1, \dots, A_p , оцененных по многим критериям K_1, \dots, K_m . Каждый критерий K_i имеет упорядоченную дискретную шкалу $X_i = \{x_i^1, \dots, x_i^{g_i}\}$, $i = 1, \dots, m$. Заданы упорядоченные классы (категории) C_1, \dots, C_q . Требуется разбить исходную совокупность многопризнаковых объектов по классам. Рассмотрим некоторые из методологических подходов к решению этой задачи.

В методе ELECTRE TRI [Roy et al., 1993] классифицируется множество альтернатив, оцененных по многим критериям, имеющим балльные шкалы и разные веса. Классификация альтернатив производится в соответствии с границами классов, которые строятся с использованием специальных индексов конкорданса (согласия) и дискорданса (несогласия) при попарном сравнении альтернатив. Значения этих индексов вычисляются в процессе решения проблемы. Веса критериев назначаются ЛПР, что является субъективной и не имеющей строгих обоснований процедурой.

Интерактивная процедура классификации, в которой предпочтения ЛПР описываются линейной функцией полезности, содержится в [Köksalan et al., 2003]. Функция полезности задается сверткой многих числовых критериев в виде «взвешенной суммы», где определение весов исходных показателей является достаточно серьезной проблемой. При большом числе критериев построение функции полезности сопряжено с большими трудозатратами ЛПР. Применение методов, использующих взвешенную свертку критериев, для решения задач классификации большой размерности не позволяет дать объяснения полученных результатов, поскольку невозможно восстановить исходные данные по агрегированным показателям.

Классификация многокритериальная альтернатив с применением огрубленных множеств предложена [Greco et al., 2002]. Предпочтения ЛПП выражаются с помощью наборов решающих правил, которые с разной степенью определенности относят альтернативы в заданные классы. Метод оперирует с достаточно большим числом решающих правил классификации, трудным для непосредственного анализа ЛПП, и требует обучения на специально выделенных массивах данных.

Иной подход к порядковой классификации многокритериальных альтернатив дает методология вербального анализа решений [Ларичев, 2006], предназначенная для работы с альтернативами, имеющими лишь вербальные оценки по критериям, к которым не применяются никакие количественные преобразования. Оценка и сравнение могут проводиться как для всех гипотетически возможных, так и для конкретных альтернатив. Предпочтения ЛПП проверяются на непротиворечивость, а выявленные противоречия предъявляются персоне для анализа и разъяснения. Эффективность методов классификации, разработанных в рамках вербального анализа решений, оценивается числом обращений к ЛПП, необходимых для построения полной непротиворечивой классификации.

При построении классификации большой размерности важно учитывать возможности человека. Результаты психологических экспериментов показывают [Ларичев, 2006], что при числе критериев больше 5, числе оценок на шкалах критериев больше 4 и числе классов решений более 5 люди склонны применять различные упрощенные стратегии классификации с использованием только части критериев. Подобные сложности возникают и при решении практических задач [Петровский и др., 1990].

Преодолеть указанные сложности, можно, например, за счет снижения размерности признакового пространства. Один из способов состоит в бинаризации шкал критериев [Асанов и др., 2001б]. Однако такой подход сильно упрощает описание альтернатив и классов решений, снижает выразительные возможности языка для представления предпочтений ЛПП и объяснения полученного результата, что не всегда приемлемо в конкретных приложениях. Рассмотрим другой возможный подход к снижению размерности признакового пространства, использующий последовательное иерархическое агрегирование признаков в небольшое число критериев с вербальными порядковыми шкалами.

2. Снижение размерности признакового пространства

Формально задача снижения размерности признакового пространства имеет следующий вид:

$$X_1 \times \dots \times X_m \rightarrow Y_1 \times \dots \times Y_n, n < m,$$

где X_1, \dots, X_m — исходный набор признаков, Y_1, \dots, Y_n — новый набор признаков, m — размерность исходного признакового пространства, n — размерность нового признакового пространства. Каждый из признаков имеет свою собственную шкалу $X_i = \{x_i^1, \dots, x_i^{g_i}\}$, $i = 1, \dots, m$, $Y_j = \{y_j^1, \dots, y_j^{h_j}\}$, $j = 1, \dots, n$ с упорядоченной градацией оценок.

Предлагаемый подход к агрегированию признаков базируется на предпочтениях ЛПР. Первоначально при участии ЛПР формируется набор исходных характеристик рассматриваемых объектов. В зависимости от специфики задачи эти характеристики могут быть либо заданы заранее, либо сформированы в процессе анализа проблемы. Например, при выборе трассы газопровода необходимо учитывать затраты на строительство, экологический ущерб, вероятность аварий и оценку их последствий и другие показатели [Ларичев, 2006]. Далее, основываясь на опыте и интуиции ЛПР, исходные характеристики объединяются в группы критериев, обладающих вербальными порядковыми шкалами с небольшим числом градаций (3–5). Смысловое содержание критериев и шкал оценок определяется ЛПР. Критерии должны иметь такие шкалы оценок, которые, с одной стороны, отражают агрегированные качества объектов, а с другой стороны, понятны ЛПР при окончательном выборе объекта или их классификации.

Для агрегирования признаков и снижения размерности признакового пространства воспользуемся методом ИСКРА (Иерархическая Структуризация Критериев и Атрибутов), который включает следующие шаги [Ройзензон, 2005а]. Составляется перечень всех базовых показателей (например, список технических характеристик объектов), которые образуют нижний уровень иерархической системы показателей. Для каждого базового показателя формируется шкала, которая может иметь числовые (точечные, интервальные) или вербальные оценки. Шкалы оценок базовых показателей могут совпадать с обычно используемыми на практике, либо конструироваться специально.

Далее ЛПР, по своему усмотрению, определяет число, состав и содержание критериев следующего уровня иерархии. В качестве критерия можно выбрать один из базовых показателей или несколько характеристик, объединенных в составной критерий. ЛПР устанавливает, какие базовые показатели будут считаться самостоятельными критериями, а какие будут включены в тот или иной составной критерий. Для формирования шкал оценок составных критериев можно использовать несколько процедур.

Наиболее простым и легко воспринимаемым ЛПР способом конструирования порядковой шкалы для составного критерия является эвристический метод, основанный на сечении многомерного признакового пространства параллельными гиперплоскостями. Каждый слой, состоящий из однотипных комбинаций базовых показателей, представляет одну из оценок по составному критерию, а число таких сечений (оценок) определяет

ся ЛПР из содержательных соображений. Например, исходные оценки объединяются в обобщенную оценку по принципу: все лучшие оценки по базовым показателям образуют одну лучшую оценку по составному критерию, все средние оценки — несколько средних оценок, все худшие оценки — одну худшую оценку. Отметим, что актуальность использования стратифицированного подхода в задачах принятия решений подчеркивалась в работе [Готов и др., 1984].

Более сложные процедуры построения шкалы составного критерия предполагают применение методов вербального анализа решений [Ларичев, 2006], в которых в качестве рассматриваемых альтернатив выступают все возможные кортежи оценок в признаковом пространстве, образованном декартовым произведением шкал оценок базовых показателей. Метод ЗАПРОС позволяет построить единую порядковую шкалу составного критерия, формируя ее из значений оценок по отдельным базовым показателям. Методы ОРКЛАСС и ЦИКЛ позволяют построить полную непротиворечивую порядковую классификацию всех наборов оценок по базовым показателям, где классы образуют шкалу оценок составного критерия.

Процедура агрегирования показателей имеет последовательный характер, т. е. полученные группы критериев могут быть, в свою очередь, объединены в новые группы (следующий уровень иерархии) и так далее. При конструировании шкал составных критериев на разных этапах процедуры могут использоваться различные подходы. Например, один из агрегированных критериев можно сформировать при помощи метода сечений, а другой — при помощи многокритериальной порядковой классификации.

3. Построение правил классификации

В случае порядковой классификации многокритериальных альтернатив процедуру агрегирования показателей можно представить в виде иерархической структуры, состоящей из последовательно выполняемых однотипных блоков классификации. Блоки содержательно выделяются ЛПР в зависимости от специфики решаемой задачи.

Каждый блок классификации i -го уровня иерархии включает некоторый набор признаков и один составной критерий. В качестве объектов классификации выступают наборы градаций оценок на шкалах признаков. Классами решений i -го уровня служат градации оценок на шкале составного критерия. В блоке классификации $(i + 1)$ -го уровня иерархии составные критерии i -го уровня считаются признаками, множество градаций оценок которых представляет собой новые объекты классификации в сокращенном признаковом пространстве, а классами решений будут теперь градации оценок на шкале составного критерия $(i + 1)$ -го уровня.

Процедура повторяется до тех пор, пока не останется единственный составной критерий верхнего уровня, шкала оценок которого образует искомые упорядоченные классы решений C_1, \dots, C_q . Тем самым устанавливается соответствие между классами решений C_1, \dots, C_q и совокупностью исходных показателей — множеством $X_1 \times \dots \times X_m$ всех возможных комбинаций градаций оценок на шкалах критериев $X_i = \{x_i^1, \dots, x_i^{g_i}\}$, $i = 1, \dots, m$ критериев K_1, \dots, K_m и находятся границы классов что позволяет легко построить классификацию реальных альтернатив A_1, \dots, A_p , оцененных по многим критериям.

Рассмотрим построение правил классификации на модельном примере (рис. 1). Исходное множество альтернатив описывается восемью базовыми показателями K_1, \dots, K_8 , имеющими шкалы X_i с двумя или тремя вербальными порядковыми оценками 0,1,2, где 0 обозначает лучшую оценку, 1 — среднюю (или худшую), 2 — худшую. Требуется разбить множество альтернатив на пять упорядоченных классов C_1, \dots, C_5 , где $C_1 \prec \dots \prec C_5$.



Рис. 1. Схема построения критериев и формирования шкал оценок

Например, критерий K_1 характеризует «Степень выполнения заявленных задач», которая может оцениваться, как 0 — задачи выполнены полностью, 1 — задачи выполнены частично, 2 — задачи не выполнены; критерий K_3 оценивает «Достижение поставленной цели в установленные сроки» как 0 — реальное, 1 — нереальное. Критерием верхнего уровня является «Результативность», градации оценок по шкале которого (высокая, хорошая, средняя, низкая, неудовлетворительная) определяют 5 упорядоченных классов решений C_1, \dots, C_5 .

Критерии K_1, \dots, K_8 имеют следующие шкалы: $X_1 = \{0,1,2\}$; $X_2 = \{0,1,2\}$; $X_3 = \{0,1\}$; $X_4 = \{0,1,2\}$; $X_5 = \{0,1\}$; $X_6 = \{0,1\}$; $X_7 = \{0,1\}$; $X_8 = \{0,1,2\}$. Таким образом, размерность исходного признакового пространства равна 1296. Непосредственная классификация исходного множества комбинаций оценок требует существенных трудозатрат ЛПР.

Допустим, что ЛПР решил агрегировать исходные признаки K_1, K_2, K_3 в составной критерий AK_1 ; признаки K_5, K_6, K_7 — соответственно в составной критерий AK_2 и признаки K_4, K_8 — в составной критерий AK_3 . Составные критерии AK_1, AK_2, AK_3 имеют порядковые шкалы с тремя градациями: $Y_1 = \{0,1,2\}$; $Y_2 = \{0,1,2\}$; $Y_3 = \{0,1,2\}$, где значения 0,1,2 являются вербальными оценками, определяемыми содержанием соответствующих составных критериев, и выступают как классы решений первого уровня.

Для формирования шкал составных критериев ЛПР воспользовался методом ОРКЛАСС. В результате опроса ЛПР для шкалы Y_1 получены следующие градации оценок (классы решений с границами): $y_1^1 = 0$ — класс 0 (верхняя граница: 000; нижняя граница: 100,010,001); $y_1^2 = 1$ — класс 1 (верхняя граница: 200,110,020,101,011; нижняя граница: 210,120,201,111,021); $y_1^3 = 2$ — класс 2 (верхняя граница: 220,211,121; нижняя граница: 221).

Для шкал Y_2 и Y_3 составных критериев AK_2 и AK_3 получены такие градации оценок:

- $y_2^1 = 0$ — класс 0 (верхняя граница: 000; нижняя граница: 001); $y_2^2 = 1$ — класс 1 (верхняя граница: 100,010; нижняя граница: 101,011); $y_2^3 = 2$ — класс 2 (верхняя граница: 110; нижняя граница: 111);
- $y_3^1 = 0$ — класс 0 (верхняя граница: 00; нижняя граница: 00); $y_3^2 = 1$ — класс 1 (верхняя граница: 10,01; нижняя граница: 20,11,02); $y_3^3 = 2$ — класс 2 (верхняя граница: 21,12; нижняя граница: 22).

Рассмотрим теперь наборы всех оценок по составным критериям как объекты классификации следующего второго уровня, где классами решений C_1, \dots, C_5 являются градации оценок шкалы $Z = \{z^1, z^2, z^3, z^4, z^5\}$ составного критерия верхнего уровня иерархии. Аналогичным образом, агрегируя показатели AK_1, AK_2, AK_3 , получим для шкалы Z такие оценки:

- z^1 — класс C_1 (верхняя граница: 000; нижняя граница: 000); z^2 — класс C_2 (верхняя граница: 100,010,001; нижняя граница: 110,101); z^3 — класс C_3 (верхняя граница: 200,020,011,002; нижняя граница:

211,121,202,112,022); z^4 — класс C_4 (верхняя граница: 220,212,122; нижняя граница: 221,212,122); z^5 — класс C_5 (верхняя граница: 222; нижняя граница: 222).

Таким образом, реальные альтернативы, имеющие оценки по исходным критериям, непосредственно относятся при классификации к сформированным классам решений. Отметим что для построения конечных классов решений C_1, \dots, C_5 потребовалось получить от ЛПР ответы на следующие числа вопросов: при формировании шкал Y_1, Y_2, Y_3 составных критериев AK_1, AK_2, AK_3 — соответственно на 16, 6 и 7 вопросов, при формировании шкалы Z агрегированного критерия — на 22 вопроса. Это существенно меньше, чем при использовании других методов многокритериальной порядковой классификации.

Используем теперь для конструирования порядковой шкалы составного критерия метод отсекающих гиперплоскостей. С геометрической точки зрения данная процедура состоит в «нарезке» многомерного параллелепипеда (прямоугольника) на группы наборов исходных признаков в зависимости от числа признаков, образующих составной критерий.

Возможные градации оценок для шкалы критерия AK_1 представлены на рис. 2. К классу 0 (оценка $y_1^1 = 0$) относится следующая комбинация оценок: (000),(001),(010),(100), к классу 2 (оценка $y_1^3 = 2$) — комбинация оценок (121),(211),(221),(220), а к классу 1 (оценка $y_1^2 = 1$) — все остальные комбинации оценок (011),(021),(101),(111),(201),(110),(200),(020),(210),(120).

Возможные градации оценок для шкалы критерия AK_2 представлены на рис. 3. Класс 0 ($y_2^1 = 0$) образует комбинация всех лучших оценок (000),



Рис. 2. Схема формирования шкалы оценок для составного критерия AK_1



Рис. 3. Схема формирования шкалы оценок для составного критерия AK_2

класс 2 ($y_2^3 = 2$) — комбинация всех худших оценок (111) и класс 1 ($y_2^2 = 1$) — все остальные комбинации оценок (001),(011),(101),(100),(010),(110).

Возможные градации оценок для шкалы критерия AK_3 представлены на рис. 4. Класс 0 ($y_3^1 = 0$) составляют все лучшие оценки (00); класс 1 ($y_3^2 = 1$) — комбинации оценок (01),(10),(02),(11),(20); класс 2 ($y_3^3 = 2$) — комбинации оценок (12),(21),(22).

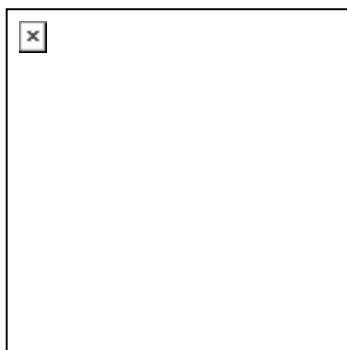


Рис. 4. Схема формирования шкалы оценок для составного критерия AK_3

Аналогичным образом, агрегируя показатели AK_1 , AK_2 , AK_3 , имеем (см. рис. 5): класс C_1 (z^1) состоит из всех лучших оценок (000); класс C_2 (z^2) — из комбинаций оценок (100),(010),(001),(002),(101),(011),(200),(110),(020); класс C_3 (z^3) — из комбинаций оценок (102),(012),(201),(111),(021),(210), (120); класс C_4 (z^4) — из комбинаций оценок (202),(112),(022),(211),(121),(220), (212),(122),(221); класс C_5 (z^5) — из всех худших оценок (222).

Заключение

Предложен новый подход к решению задачи порядковой классификации альтернатив, оцененных по многим критериям с вербальными шкалами. Применение процедуры иерархического агрегирования исходных показателей позволяет значительно снизить размерность признакового пространства, что существенно сокращает время, затрачиваемое ЛПП на построение полной и непротиворечивой классификации.

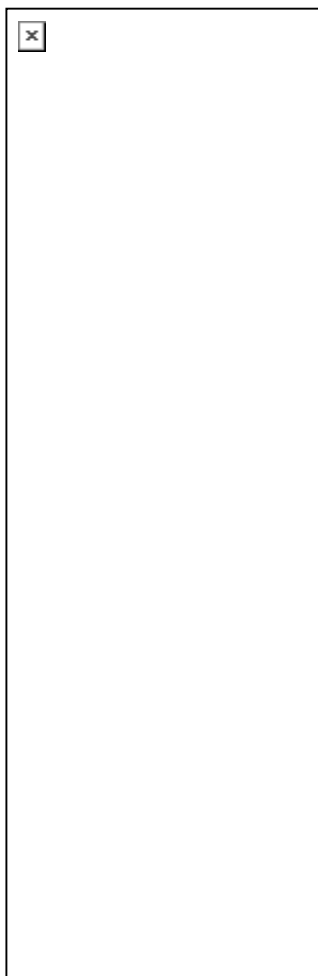


Рис. 4. Схема формирования шкалы оценок для составного критерия верхнего уровня иерархии

Важной особенностью разработанной процедуры является возможность сформировать разные наборы критериев, с тем, чтобы сравнить полученные результаты для разных вариантов классификации с целью оценки качества сделанного выбора.

Предлагаемый подход обеспечивает систематизацию имеющейся информации, анализ причин принятия окончательного решения, получение его обоснования. Разделение всех критериев на отдельные группы позволяет «распараллелить» решение задачи, что также дает ощутимую экономию времени.

Процедура снижения размерности признакового пространства была использована при решении практических задач оценки банковских кредитов в зависимости от степени риска [Асанов и др., 2001a] и многокритериального выбора вычислительных кластеров [Ройзензон, 2005b].

Литература

1. [Асанов и др., 2001a] Метод многокритериальной классификации ЦИКЛ и его применение для анализа кредитного риска / А. А. Асанов, П. В. Борисенков, О. И. Ларичев, Е. В. Нарыжный, Г. В. Ройзензон // *Экономика и математические методы*. 2001. Т. 37, № 2. С. 14–21.
2. [Асанов и др., 2001b] Асанов А., Подлипский О. Опыт построения большой базы экспертных знаний // *Методы поддержки принятия решений: Сборник трудов Института системного анализа Российской академии наук* / Под ред. О. И. Ларичева. М.: URSS, 2001. С. 42–50.
3. [Глотов и др., 1984] Глотов В. А., Павельев В. В. Векторная стратификация. М.: Наука, 1984. С. 94.
4. [Ларичев, 2006] Ларичев О. И. Вербальный анализ решений / Под ред. А. Б. Петровского. М.: Наука, 2006. С. 181.
5. [Петровский и др., 1990] Петровский А. Б., Шепелев Г. И. Система поддержки принятия решений для конкурсного отбора научных проектов // *Проблемы и методы принятия уникальных и повторяющихся решений. Сборник трудов* / Под ред. С. В. Емельянова, О. И. Ларичева. № 10. М.: ВНИИСИ, 1990. С. 25–31.
6. [Ройзензон, 2005a] Ройзензон Г. В. Способы снижения размерности признакового пространства для описания сложных систем в задачах принятия решений // *Новости искусственного интеллекта*. 2005. № 1. С. 18–28.
7. [Ройзензон, 2005b] Ройзензон Г. В. Многокритериальный выбор вычислительных кластеров // *Методы поддержки принятия решений: Сборник трудов Института системного анализа Российской академии наук* / Под ред. С. В. Емельянова, А. Б. Петровского. М.: URSS, 2005. Т. 12. С. 68–94.
8. [Greco et al., 2002] Greco S., Matarazzo B., Slowinski R. Rough sets methodology for sorting problems in presence of multiple attributes and criteria // *European Journal of Operational Research*. 2002. Vol. 138. No. 2. P. 247–259.
9. [Köksalan et al., 2003] Köksalan M., Ulu C. An interactive approach for placing alternatives in preference classes // *European Journal of Operational Research*. 2003. Vol. 144. No. 2. P. 429–439.
10. [Roy et al., 1993] Roy B., Bouyssou D. Aide Multicritere a la decision: Methodes et cas. Paris: Economica, 1993. P. 695.