

СНИЖЕНИЕ РАЗМЕРНОСТИ ПРИЗНАКОВОГО ПРОСТРАНСТВА В ЗАДАЧАХ МНОГОКРИТЕРИАЛЬНОЙ КЛАССИФИКАЦИИ

Алексей Петровский, Григорий Ройзензон

Аннотация: Представлен метод последовательного снижения размерности признакового пространства, который позволяет упростить процедуру порядковой классификации многокритериальных альтернатив и уменьшить ее трудоемкость.

Ключевые слова: многокритериальная порядковая классификация, вербальный анализ решений, агрегирование признаков, снижение размерности признакового пространства

ACM Classification Keywords: H.4.2 Information Systems Applications: Types of Systems: Decision Support.

Conference: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008

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 достаточно большим числом признаков, на основе предпочтений лица, принимающего решение (ЛПР), является достаточно трудоемкой процедурой, требующей значительных временных затрат и разработки специальных методов опроса ЛПР. При решении задачи классификации большой размерности ЛПР зачастую применяет различные упрощенные стратегии с использованием только части критериев, что негативно сказывается на построении границ классов решений, выработке решающих правил, дальнейшем анализе полученных результатов [Ларичев, 2006].

Дополнительные трудности появляются в случае слабо структурируемых задач классификации, сочетающих количественные и качественные зависимости, для которых построение объективных моделей либо невозможно в принципе, 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., 2001a].

Одним из способов преодоления описанных выше сложностей является снижение размерности признакового пространства. 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.

Многокритериальная порядковая классификация 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 имеет упорядоченную дискретную шкалу $X_i = \{x_i^1, \dots, x_i^q\}$, $i=1, \dots, m$. 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 and Bouyssou, 1993] классифицируется множество альтернатив, оцененных по многим критериям, имеющим балльные шкалы и разные веса. Классификация строится с использованием специальных индексов конкорданса (согласия) и дискорданса (несогласия) при попарном сравнении альтернатив. Значения этих индексов формируются ЛПР не в какой-то фиксированный момент, а в процессе анализа 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]. Математический аппарат, используемый в этой процедуре, базируется на свертке 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 объяснение полученных результатов, поскольку невозможно восстановить исходные данные по агрегированным показателям.

Классификация многокритериальная альтернатив с применением огрубленных множеств предложена в [Greco et al., 2002]. Предпочтения ЛПР 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 ЛПР, и требует для своего построения обучения на специально выделенных массивах данных.

Принципиально иной подход к многокритериальной порядковой классификации 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 ЛПР проверяются на непротиворечивость, а выявленные противоречия предъявляются персоне для анализа и разъяснения. Эффективность методов классификации ОРКЛАСС и ЦИКЛ, разработанных в рамках вербального анализа решений, оценивается числом обращений к ЛПР, 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 оценок на шкалах критериев [Асанов и др., 2001б]. Однако такой подход сильно упрощает описание альтернатив и классов решений, снижает выразительные возможности языка для представления предпочтений ЛПР и объяснения полученного результата, что не всегда согласовывается со спецификой конкретной предметной области. Рассмотрим другой возможный подход к снижению размерности признакового пространства, использующий последовательное иерархическое агрегирование признаков в небольшое число критериев с вербальными порядковыми шкалами.

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 $X_i = \{x_i^1, \dots, x_i^g\}$, $i=1, \dots, m$, $Y_j = \{y_j^1, \dots, y_j^h\}$, $j=1, \dots, n$ с упорядоченной градацией оценок.

Предлагаемый подход к агрегированию признаков базируется на предпочтениях ЛПР. Первоначально при участии ЛПР формируется исходный набор 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, будут отражать агрегированные качества объектов, а с другой стороны, будут понятны ЛПР при окончательном выборе объекта или их классификации.

Метод ИСКРА (Иерархическая Структуризация Критериев и Атрибутов) включает следующие шаги [Ройзензон, 2005а]. Составляется перечень всех базовых показателей (например, список технических характеристик объектов), которые составляют 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 отнесены к тому или иному составному критерию. Для формирования шкал оценок по составным критериям ЛПР может воспользоваться несколькими процедурами.

Наиболее простым и легко воспринимаемым ЛПР способом конструирования порядковой шкалы для составного критерия является эвристический подход, в котором используются однотипные наборы порядковых вербальных шкал базовых показателей. Исходные оценки объединяются в обобщенную оценку по принципу: все лучшие оценки по базовым показателям образуют одну лучшую оценку по составному критерию, все средние оценки – средние оценки, все худшие оценки – одну худшую оценку.

Более сложные процедуры построения шкал критериев предполагают применение методов вербального анализа решений [Ларичев, 2006], в которых необходимо рассматривать множество всех возможных кортежей оценок в признаковом пространстве, образованном декартовым произведением значений оценок на шкалах критериев. Метод ЗАПРОС позволяет построить единую порядковую шкалу составного критерия, формируя ее из оценок по отдельным базовым показателям. Метод ОРКЛАСС предназначен для построения полной непротиворечивой порядковой классификации многопризнаковых объектов, в качестве которых в нашем случае выступают наборы оценок по базовым показателям, образующим составной критерий.

Процедура агрегирования показателей может иметь 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 простого эвристического метода, а другой – при помощи многокритериальной порядковой классификации.

Построение решающих 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 состоит из некоторого набора признаков и одного составного критерия. В качестве объектов классификации выступают все градации 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 взаимно-однозначное 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 $X_i = \{x_i^1, \dots, x_i^p\}$, $i=1, \dots, m$ 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 решающих правил классификации на модельном примере. Исходное множество альтернатив описывается восемью критериями (базовыми признаками) 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 (рис. 1).

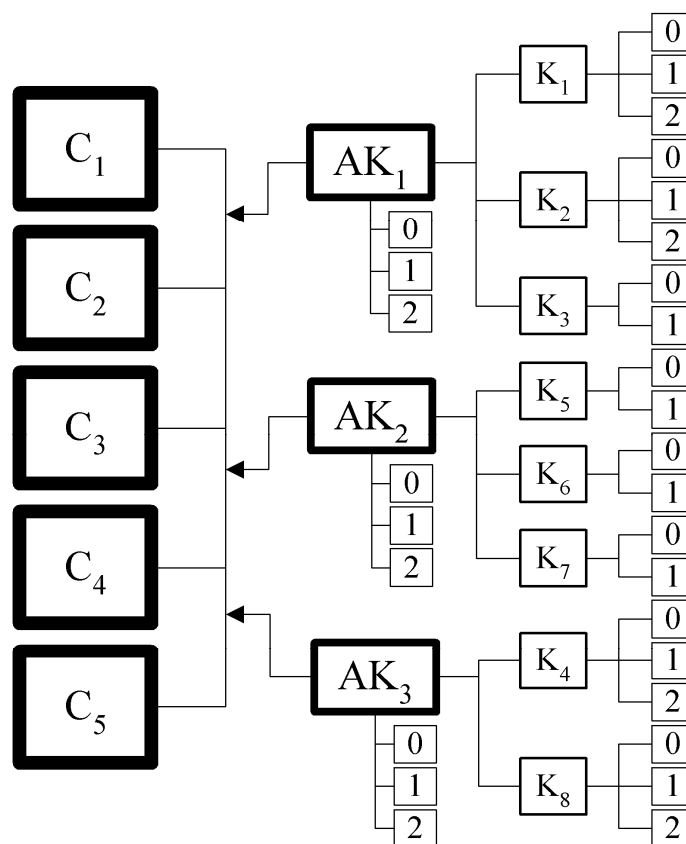


Рис. 1. Схема построения критериев и формирования шкал оценок

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. 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. Составным критерием верхнего уровня является «Результативность», градации оценок по шкале которого (высокая, хорошая, средняя, низкая, неудовлетворительная) определяют 5 упорядоченных классов решений. Непосредственная классификация исходного множества альтернатив требует существенных затрат ЛПР.

Введем три составных критерия AK_1, AK_2, AK_3 , имеющих 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 1 уровня для исходных базовых признаков (критериев). 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 . Для формирования шкал составных критериев ЛПР воспользовался одной из указанных выше процедур. В результате опроса 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 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 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 , имеем:

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 16, 6 и 7 вопросов при формировании шкал Y_1, Y_2, Y_3 составных критериев AK_1, AK_2, AK_3 соответственно и ответов на 22 вопроса при формировании шкалы Z агрегированного критерия, что существенно меньше, чем при использовании других методов многокритериальной порядковой классификации.

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, затраченное ЛПР на решение задачи. 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., 2001а] и многокритериального выбора вычислительных кластеров [Ройзензон, 2005б].

Благодарности

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

Список литературы

- [Асанов и др., 2001а] 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.
- [Асанов и др., 2001б] 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.: Эдиториал УРСС, 2001. — С. 42–50.
- [Ларичев, 2006] Ларичев О. И. Вербальный анализ решений / Под ред. А. Б. Петровского. — М.: Наука, 2006. — 181 с.
- [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.
- [Ройзензон, 2005а] Ройзензон Г. В. Способы снижения размерности признакового пространства для описания сложных систем в задачах принятия решений // Новости искусственного интеллекта. — 2005. — № 1. — С. 18–28.
- [Ройзензон, 2005б] 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.: Эдиториал УРСС, 2005. — Т. 12. — С. 68–94.
- [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. — Pp. 247-259.
- [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. — Pp. 429–439.
- [Roy et al., 1993] Roy B., Bouyssou D. Aide Multicritere a la decision: Methodes et cas — Paris: Economica, 1993. — 695 p.

Сведения об авторах

Петровский Алексей Борисович – д.т.н., зав. лабораторией Института системного анализа РАН, 117312, Россия, Москва пр-т 60 лет Октября, 9, ИСА РАН, тел. (495) 135-8503, e-mail: pab@isa.ru

Ройзензон Григорий Владимирович – научный сотрудник Института системного анализа РАН, 117312, Россия, Москва пр-т 60 лет Октября, 9, ИСА РАН, тел. (495) 135-8503, e-mail: rgv@isa.ru