Д.Ю. Кочин

Извлечение решающих правил из границ классов при решении задач порядковой классификации

Аннотация. Задача порядковой классификации заключается в разнесении экспертом полного множества альтернатив на несколько классов. На набор классов и оценки по критериям распространяется требование порядка – классы и шкалы оценок по критериям должны быть упорядочены в соответствии с целью классификации. Множество альтернатив представлено в виде всевозможных комбинаций наборов оценок по всем критериям. В силу упорядоченности задачи классы, полученные после ее решения, могут быть заданы только своими границами. Остальные альтернативы, принадлежащие классу, будут находиться в отношении доминирования хотя бы с одной из граничных альтернатив, т. е. между границами. Существует гипотеза, что путем анализа границ можно выявить решающие правила, близкие к тем, которые эксперт неявно использует при построении классификации. В статье вводится формальное определение решающего правила и предлагается алгоритм для представления границы классов в виде набора решающих правил.

Ключевые слова: порядковая классификация, вербальный анализ решений, решающие правила, границы классов.

Введение

Принято различать два типа знаний: декларативные и процедуральные. К декларативным знаниям относят описания фактов, изложение теорий, наблюдений. Учебники и научные монографии являются примерами декларативных знаний.

Процедуральные знания можно также назвать умениями, навыками. Человек овладевает процедуральными знаниями, когда он не только знает теорию, но и умеет применить ее на практике. Человека, в совершенстве овладевшего процедуральными знаниями в какой–либо области, принято называть экспертом. Путь от новичка в какой-то профессиональной области до эксперта, находящегося на вершине профессионального мастерства, требует не менее 10 лет интенсивной практики [11]. Как показали исследования, этот отрезок времени является примерно одинаковым для столь разных областей человеческой деятельности, как медицина, игра в шахматы, сочинение музыки, спорт и т.д. [10].

Высокая востребованность в специалистах и большая длительность периода их становления делает актуальной проблему создания интеллектуальных обучающих систем (ИОС), позволяющих передавать знания от эксперта к новичку [9, 13].

Среди моделей современной когнитивной психологии доминирует так называемый информационный подход [14], рассматривающий человека как систему переработки информации [18]. Знание в рамках этого подхода рассматривается как комплекс реально существующих элементов (символов, образов), хранящихся в памяти человека, которые обрабатываются мозгом подобно программе в компьютере и являются источником интеллектуального поведения. В данном случае не важно, из чего состоит символ сам по себе (как он реализован на более низком уровне – уровне межнейронных связей), а под самим словом «символ» имеется в виду некоторый внутренний образ (паттерн). При этом знание рассматривается как набор взаимосвязанных и относительно статичных элементов, которые можно хранить, извлекать, модифицировать, передавать экспертной системе или другому человеку. В рамках данного подхода был разработан ряд успешных вычислительных моделей памяти (EPAM [19]), процессов мышления и обучения (SOAR) [17], ACT-R

[8], эксперименты с которыми показали адекватность их поведения поведению человека в психологических экспериментах.

В рамках информационного подхода можно сформулировать следующие проблемы передачи экспертных знаний:

1. Извлечение экспертных знаний, т.е. построение базы знаний, с помощью которой возможно некоторым формальным способом получать решения, совпадающие с решениями эксперта.

2. Обучение новичков знаниям, представленным в построенной базе знаний, после которого качество решений новичков будет близко к качеству решений эксперта.

Решение указанных проблем тесно связано с характером рассматриваемой предметной области. Принято различать хорошо определенные или хорошо структурируемые области знаний, к которым относится, например, решение типовых задач математики, физики, программирования, и недостаточно определенные или слабоструктурируемые области, такие, например, как медицинская диагностика. В отличие от хорошо определенных областей в решении указанных проблем для слабоструктурируемых областей знаний, где «качественные, трудноформализируемые и неопределенные факторы имеют тенденцию доминировать» [18], существуют значительные трудности.

Отличительной чертой многих проблем слабоструктурируемых областей знаний является отсутствие надежных количественных моделей, поэтому решение подобных проблем обычно поручают экспертам – наиболее квалифицированным и опытным специалистам. Поскольку специалисты-эксперты являются основными носителями профессиональных знаний и навыков в своей области, то большое значение имеют исследования особенностей поведения экспертов при принятии решений.

Серьезной проблемой в извлечении экспертных знаний является тот факт, что в большинстве случаев эксперты не могут явно сформулировать правила, которыми они пользуются при принятии решений. Те же правила, которые удается получить в явном виде, охватывают лишь наиболее простые случаи из тех, которые эксперт способен правильно распознать. Это позволяет говорить о том, что в результате многолетней интенсивной практики у экспертов неявно формируются правила распознавания, по крайней мере, часть которых находится на подсознательном уровне [15]. Именно подсознательный характер навыков экспертов стал причиной возникновения значительных трудностей при построении экспертных систем, а извлечение экспертных знаний было названо «узким местом» (bottleneck) искусственного интеллекта [12].

1. Порядковая классификация

Во многих практических случаях задача извлечения экспертных знаний может быть представлена как задача классификации, так как экспертное знание часто состоит в отнесении объектов (альтернатив, состояний) к классам решений. Так, например, инженер анализирует сбой в сложной технической системе и определяет возможный тип неисправности. Врач изучает состояние пациента и ставит диагноз, выбирая из нескольких возможных типов заболеваний. Элементы, составляющие некоторую совокупность, подлежащую классификации, могут иметь разнообразную природу. В частности, это могут быть различные физические объекты, варианты выбора, состояния некоторого объекта. Далее в контексте задачи классификации мы будем употреблять термины “объект”, “альтернатива”, “состояние”, “случай”, “вектор” как синонимы.

Способ отнесения рассматриваемого объекта к тому или иному классу решений очень часто не может быть явно описан экспертом, в силу невербализуемости стратегии его поведения. Тем не менее, эти невербализуемые навыки эффективно и быстро применяются, когда эксперт решает классификационные задачи в своей предметной области.

Задача экспертной классификации в постановке [3, 16] предполагает определение множества критериев (признаков), которыми описывается каждый объект. Для каждого критерия задается конечное множество допустимых оценок – шкала критерия. Если в некоторой задаче множество оценок по одному или более критериям бесконечно, то соответствующая шкала преобразуется в конечную посредством разбиения исходного бесконечного множества оценок на конечный набор интервалов.

Декартово произведение шкал всех критериев представляет собой пространство всех гипотетически возможных объектов (состояний),

описываемых в рамках данной задачи. Требуется на основе экспертных знаний построить классификацию в указанном пространстве состояний, т.е. сформировать правила отнесения каждого объекта к одному из заранее определенных классов.

Принято различать задачи номинальной и порядковой классификации (classification and sorting). В первом случае объекты должны быть отнесены к номинальным, т.е. неупорядоченным классам решений. Во втором случае классы решений упорядочены по степени выраженности некоторого свойства, например, степени тяжести заболевания (тяжелая, средняя, легкая), качеству кредитного проекта (высшая категория, средняя категория, сомнительная категория, убытки). В этом случае оценки по критериям также следует упорядочивать по их характерности для заданных классов. Преобразованием структуры задачи можно свести задачу номинальной классификации к нескольким задачам порядковой классификации [6]. Например, задача номинальной классификации с M классами решений $C_1 \ldots C_M$ может быть решена как совокупность M задач порядковой классификации, где в задаче i используются классы решений "$\in C_i$" и "$\notin C_i$". Основное внимание далее будет уделено задаче порядковой экспертной классификации.

Приведем формальную постановку задачи многокритериальной порядковой экспертной классификации.

Дано:

1. G – свойство, отвечающее целевому критерию задачи (наличие и степень тяжести заболевания, критичность неисправности в технической системе, ценность кредитного проекта и т.д.).

2. $K = \{K_1, K_2, \ldots, K_N\}$ – множество критериев (признаков), по которым оценивается каждый объект исследования.

3. $S_q = \{k_1^q, k_2^q, \ldots, k_{\omega_q}^q\}$, $q = 1, \ldots, N$ – множество оценок по критерию K_q; ω_q – число градаций на шкале критерия K_q; оценки $k_1^q, k_2^q, \ldots, k_{\omega_q}^q$ упорядочены по убыванию характерности для свойства G. Т. е. на каждом множестве S_q определено рефлексивное антисимметричное транзитивное отношение Q_q (необязательно связное) такое, что $\left(k_i^q, k_j^q\right) \in Q_q \Rightarrow i \le j$.

4. $Y = S_1 \times S_2 \times \ldots \times S_N$ – декартово произведение шкал критериев, которое определяет пространство состояний объектов, подлежащих классификации. Каждый объект описывается набором оценок по критериям $K_1, \ldots, K_N$ и представляется в виде векторной оценки $\boldsymbol{y} \in Y$, где $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$, y_q – одна из оценок из множества S_q.

5. $L = |Y| = \prod_{q=1}^{N} \omega_q$ – мощность множества Y.

6. $C = \{C_1, C_2, \ldots, C_M\}$ – множество классов решений, упорядоченных по убыванию выраженности свойства G. Т. е. на множестве C определено рефлексивное антисимметричное транзитивное отношение Q_C такое, что $\left(C_i, C_j\right) \in Q_C \Leftrightarrow i \le j$. Все эти классы четко определены, каждый объект может быть отнесен экспертом к одному и только к одному классу.

Введем бинарное отношение строгого доминирования векторных оценок:

$$P = \left\{ (\boldsymbol{x}, \boldsymbol{y}) \in Y \times Y \,\middle|\, \forall q = 1 \ldots N \;\; (x_q, y_q) \in Q_q \;\text{ и }\; \exists q_0 : x_{q_0} \neq y_{q_0} \right\} \quad (1)$$

Отношение доминирования для краткости можно записывать эквивалентной записью $(\boldsymbol{x}, \boldsymbol{y}) \in P \Leftrightarrow \boldsymbol{x} P \boldsymbol{y} \Leftrightarrow \boldsymbol{x} > \boldsymbol{y}$.

Требуется: на основе знаний эксперта построить разбиение множества допустимых альтернатив Y на M классов решений C_i ($C_i \cap C_j = \varnothing \; \forall i \neq j$, $\cup_i C_i \supseteq Y$*) так, чтобы выполнялось свойство непротиворечивости:

$$\forall \boldsymbol{x}, \boldsymbol{y} \in Y : \mathrm{x} \in C_i, \boldsymbol{y} \in C_j, (\boldsymbol{x}, \boldsymbol{y}) \in P \Rightarrow i \ge j \,. \quad (2)$$

Для решения этой задачи в рамках подхода Вербального Анализа Решений существует несколько методов - ОРКЛАСС, ЦИКЛ, КЛАНШ, КЛАРА и др. [2]. Основываясь на отношении доминирования и требованию непротиворечивости классификации, они позволяют построить полную классификацию, предъявляя эксперту явно лишь небольшую часть альтернатив из множества, подлежащего классификации.

Результатом работы всякого метода порядковой классификации являются границы классов, по которым любую альтернативу из классифицируемого множества можно отнести

к одному из классов. Верхняя граница класса – множество недоминируемых альтернатив из этого класса, нижняя – множество недоминирующих альтернатив. Вследствие этого определения любая альтернатива из этого класса будет находиться в отношении доминирования хотя бы с одной альтернативой из каждой границы.

2. Решающие правила

Психологические исследования показывают, что во многих задачах принятия решений людьми используется подсознательный счет и перебор объектов, стимулов и т.д. [7]. Это говорит о том, что такие операции являются простыми и естественными для человека, в отличие от, например, умножения. Поэтому естественно предполагать, что эксперты используют некоторые счетно-аддитивные структуры для представления правил принятия решений, в том числе и решающих правил в задачах классификации.

Приведенные данные позволяют предположить, что экспертные правила принятия решений могут быть формально представлены в виде небольшого числа (7±2) простых иерархических правил со счетно-аддитивной структурой. Действительно, исследования [1, 4, 5] показали, что границы классов в задачах экспертной классификации часто могут быть описаны небольшим количеством правил, имеющих простую двухуровневую структуру. Так, например, в бинарном случае, когда все критерии имеют двоичные шкалы оценок ($\omega_q = 2$), одним из решающих правил может быть следующее: «*альтернатива* ***y*** *принадлежит классу* C_k *, если она имеет первую оценку по первому и третьему критерию, и не менее трех первых оценок по остальным критериям*». В общем случае правило имеет вид двухуровневого дерева, в корне которого зафиксированы оценки по ключевым признакам, а на втором уровне находятся сочетания оценок по второстепенным, дополнительным признакам.

Итак, в результате работы всякого метода порядковой классификации мы имеем границы классов. Например, при разбиении альтернатив, оцененных по пяти критериям, на два класса простое решающее правило «*К классу* 1 *относятся альтернативы, у которых не менее трех высших оценок*» порождает следующую классификацию. Верхние границы обозначены индексом U, нижние – L:

$B^U(C_1)$ = {11111},

$B^L(C_1)$ = {22111, 21211, 12211, 21121, 12121, 11221, 21112, 12112, 11212, 11122},

$B^U(C_2)$ = {22211, 22121, 21221, 12221, 22112, 21212, 12212, 21122, 12122, 11222},

$B^L(C_2)$ = {22222},

Как видно, в результате применения экспертом лишь одного простого правила, получается 20 граничных элементов.

Можно ли по виду границы класса установить правило, которое ее породило? В данном случае – очень легко. Дело в том, что нижняя граница первого класса представляет собой всевозможные перестановки из трех лучших оценок и двух худших. Фактически, все десять элементов этой границы можно записать одним выражением $P_5^{3(1),2(2)}$, которое означает множество перестановок из пяти элементов, 3 из которых – первая оценка, и остальные 2 – вторая оценка. Легко видеть, что такая запись весьма близка к исходному смыслу решающего правила и представляет собой более формализованную его запись.

Подобным образом можно каждой нижней границе (кроме самой последней, которая не нужна для разделения классов) сопоставить одно или несколько таких правил. Тогда, имея какую-нибудь альтернативу, можно пройтись по этим правилам сверху вниз (от классов, соответствующих более высокому качеству, к классам, соответствующим менее высокому качеству), проверяя, удовлетворяет она правилам данного класса или нет. Если альтернатива удовлетворяет или эта альтернатива не удовлетворяет правилам предыдущих классов, значит, она относится к этому классу.

Дадим формальную постановку задачи формирования набора решающих правил. Пусть имеется некоторое произвольное множество разных альтернатив, оцененных по N критериям. Требуется описать эту совокупность альтернатив минимальным числом правил вида:

$$ab*** + P_n^{k_1[x_1]\cdots k_m[x_m]}, \text{ кроме } \{abcde,\cdots,abpqr\}, \quad (3)$$

так, чтобы каждая альтернатива попадала ровно в одно правило.

Назовем запись (3) *шаблоном* правила, который описывает некоторое множество альтерна-

тив. У этих альтернатив может быть некоторая общая часть, например, у всех альтернатив по первому и второму критерию выставлены оценки *a* и *b* (соответствует значениям ключевых признаков решающего правила). Запись *ab***** именно это и означает. Вообще же, зафиксированы могут быть оценки по любым критериям и по любому их количеству. Например, если фиксированы оценки по всем критериям, то шаблон описывает единственную альтернативу. Первую часть шаблона правила, т. е. *ab***** будем называть *фиксированной частью* правила.

К критериям, по которым в фиксированной части проставлены звездочки *, относится вторая часть шаблона $P_n^{k_1[x_1]\cdots k_m[x_m]}$, которую будем называть *перестановочной частью* правила. Вторая часть шаблона задает параметры перестановок, которые осуществляются на местах, помеченных *, и соответствует сочетаниям значений дополнительных признаков. Здесь n равно числу звездочек, k_i – числу оценок x_i, участвующих в перестановке. Например, множество $P_4^{2[1],2[2]}$ задает все перестановки из двух единиц и двух двоек, т.е. шесть элементов {1122, 1212, 1221, 2112, 2121, 2211}. Следовательно, шаблон $*2*3** + P_4^{2[1],2[2]}$ задает тоже шесть элементов {121322, 122312, 122321, 221312, 221321, 222311}. Можно заметить, что в перестановках из множества $P_4^{2[1],2[2]}$ участвовали только две оценки – 1 и 2, а при объединении с фиксированной частью *2*3** в получившихся альтернативах участвуют уже 3 оценки. Тройка добавилась из фиксированной части правила. Вообще, перестановочная и фиксированная части правила независимы друг от друга, их связь состоит только в том, что число звездочек в фиксированной части должно быть равно общему числу оценок, участвующих в перестановках.

Третья часть шаблона работает в том случае, если множество альтернатив, описываемых шаблоном, содержит не все возможные перестановки, т. е. является не полной перестановкой, но для полноты ей не хватает небольшого числа элементов. Тогда в третьей части просто перечисляются отсутствующие элементы.

При выявлении решающих правил классификации требуется разложить совокупность альтернатив именно на минимальное число правил, исходя из гипотезы, что решающие правила содержатся в кратковременной памяти эксперта, имеющей ограниченный объем. Частным случаем разложения является простое перечисление всех элементов исходной совокупности альтернатив, так как любая альтернатива есть также и правило с фиксированными оценками по всем критериям. Поиск минимального разложения направлен на уменьшение количества правил, описывающих исходную совокупность граничных альтернатив.

Например, для множества альтернатив {31221, 22221, 21321, 12321, 21131, 12131, 11231, 33312, 11322, 11132} минимальное разложение включает следующие пять правил:

31221	$**32* + P_3^{2[1],1[2]}$	22221	$***3* + P_4^{3[1],1[2]}$	33312

3. Построение минимального набора решающих правил

Для построения минимального набора решающих правил введем несколько определений.

Определение 1. Составом оценок альтернативы называется вектор $\boldsymbol{s} = (k_1, \ldots, k_n)$, где $n = \max_q(\omega_q)$, k_i – количество *i*-ых оценок в альтернативе.

Определение 2. Перестановочной частью правила, порожденную составом оценок $\boldsymbol{s} = (k_1, \ldots, k_n)$, назовем перестановочную часть $P_{\mathbf{S}} = P_m^{k_1[1]\cdots k_n[n]}$, где m – число ненулевых компонент вектора $\boldsymbol{s}$.

Утверждение 1. Правило вида (3) описывает множество альтернатив с одинаковыми составами оценок.

Доказательство. Третья часть правила, которая может только сузить множество альтернатив, не противоречит утверждению, поэтому можно ее не рассматривать. Фиксированные оценки присутствуют у всех альтернатив, а перестановочные части у всех альтернатив имеют одинаковый состав. Таким образом, и все альтернативы в описываемом множестве имеют одинаковый состав.

Согласно этому утверждению, пытаться найти правило следует только для множества альтернатив с одинаковым составом оценок.

Утверждение 2. Множество альтернатив с одинаковым составом оценок не содержит аль-

тернатив, находящихся в отношении доминирования (1).

Доказательство. Предположим противное, т.е. существуют две альтернативы $\boldsymbol{y}_1$ и $\boldsymbol{y}_2$, причем $\boldsymbol{y}_1$ доминирует $\boldsymbol{y}_2$. Тогда все оценки альтернативы $\boldsymbol{y}_1$ по абсолютной величине не больше соответствующих оценок альтернативы $\boldsymbol{y}_2$, и существует оценка, которая строго меньше. Следовательно, сумма оценок альтернативы $\boldsymbol{y}_1$ меньше, чем у альтернативы $\boldsymbol{y}_2$. Но по условию, эти альтернативы имеют одинаковый состав оценок, и, следовательно, суммы их оценок должны быть одинаковы. Полученное противоречие и доказывает утверждение.

Утверждение 3. Если мощность множества альтернатив с одинаковым составом $\boldsymbol{s} = (k_1, \ldots, k_n)$ равна

$$\frac{m!}{k_1!\cdot\ldots\cdot k_n!},$$

где m – число ненулевых компонент вектора $\boldsymbol{s}$, то этому множеству соответствует правило, состоящее только из перестановочной части $P_{\boldsymbol{s}}$.

Доказательство. Действительно, все альтернативы во множестве альтернатив разные, у всех альтернатив одинаковый состав, и их число равно комбинаторному числу перестановок m оценок в составе $\boldsymbol{s}$. По принципу Дирихле получаем, что это множество суть все перестановки $P_{\boldsymbol{s}}$, т. е. задаются правилом $P_{\boldsymbol{s}}$.

Утверждение 4. Если мощность множества альтернатив с одинаковым составом $\boldsymbol{s} = (k_1, \ldots, k_n)$ меньше

$$\frac{m!}{k_1!\cdot\ldots\cdot k_n!},$$

где m – число ненулевых компонент вектора $\mathbf{s}$, то этому множеству соответствует правило, состоящее из перестановочной части $P_{\boldsymbol{s}}$, *кроме* {*некоторое множество альтернатив*}.

Доказательство. Так как мощность множества альтернатив строго меньше $\frac{m!}{k_1!\cdot\ldots\cdot k_n!}$, то они не образуют полную перестановку. Если мы добавим недостающие элементы, то по утверждению 3 получим правило. Вычтя из этого правила добавленные элементы, получим правило, описывающее исходное множество альтернатив.

Очевидно, что любое множество альтернатив с одинаковым составом записывается одним правилом. Это напрямую следует из утверждения 4. При этом все недостающие до полной перестановки альтернативы будут записаны в третью часть правила. И если недостающих альтернатив много, то в записи полученного правила будет перечисляться слишком много элементов. Такое правило интереса не представляет, так как ищутся только краткие правила. Например, множество {1122, 1221} по утверждению 4 записывается правилом $P_4^{2[1],2[2]}$, **кроме {1212, 2112, 2121, 2211}**. Как видно, вместо перечисления только двух элементов {1122, 1221} получилось перечисление четырех **{1212, 2112, 2121, 2211}**. Т. е. не удалось получить компактную запись исходного множества, хотя оно и записано только одним правилом. Таким образом, число недостающих элементов стоит ограничить. Предлагается следующее ограничение: до полной перестановки не хватает не более k_c элементов и k_c составляет от общего числа элементов, описываемых перестановочной частью правила, не более k_p процентов. В реализованной автором компьютерной системе было принято $k_c = 3$ и $k_p = 25\%$.

Пусть имеется множество альтернатив с одинаковым составом $\vec{\mathbf{s}}$. Рассмотрим теперь матрицу $\boldsymbol{C} = \| c_{ij} \|$, где c_{ij} – число раз, когда i-тая оценка выставлена по j-тому критерию. Например, для множества {11122, 11212, 11221, 12112, 12121, 12211}

$$\mathbf{C} = \begin{Vmatrix} 6 & 3 & 3 & 3 & 3 \\ 0 & 3 & 3 & 3 & 3 \end{Vmatrix}.$$

Для удобства матрицу **C** можно представить в виде таблицы:

C :	1	2	3	4	5
1	6	3	3	3	3
2	0	3	3	3	3

где в горизонтальном заголовке указаны позиции, а в вертикальном – оценки. Элементы таблицы – числа заданных оценок на заданных позициях.

В данном примере оценка 1 встречается на первом месте 6 раз, т. е. все альтернативы из характеризуемого этой матрицей множества имеют оценку 1 по первому критерию. Поэтому, их можно записать как 1****. Отбросим теперь от всех альтернатив первый критерий и получим множество из шести альтернатив, каждая из которых состоит из четырех критериев.

Это множество удовлетворяет условию утверждения 3, а стало быть, записывается правилом $****+P_4^{2[0],2[1]}$. Возвращаясь к исходному множеству, получим, что оно записывается правилом $1****+P_4^{2[0],2[1]}$. В общем случае, если элемент c_{ij} матрицы **C** равен числу альтернатив, входящих в описываемое множество, то оценка по критерию j является общей для всех альтернатив, и ее можно вывести из рассмотрения.

Таким образом, если множество альтернатив B с одинаковым составом записывается единственным правилом вида (3), мы можем его найти по следующему алгоритму:

Алгоритм 1 – поиск единственного правила.

1. Составить матрицу **C**.
2. Исключить из рассмотрения критерии $\{j : c_{ij} = \mathrm{card}(B)\}$, общие для всех альтернатив.
3. Множество альтернатив с уменьшенным числом критериев подставить в условие утверждения 4 с учетом ограничений по k_c и k_p.
4. Если условие не удовлетворяется, то множество альтернатив B одним правилом не записывается.
5. В противном случае, объединяя результат применения утверждения 4 с исключенными на шаге 2 критериями с фиксированными оценками, получить искомое правило.

Общий алгоритм – ***алгоритм 2*** – заключается в переборе всех возможных подмножеств множества B и применения к ним алгоритма 1 поиска единственного правила. Выбираются все варианты разбиения B на минимальное число подмножеств, описываемых одним правилом, т. е. описание множества B минимальным числом правил.

Время работы такого алгоритма полного перебора растет экспоненциально с ростом числа элементов множества B. Поэтому для больших множеств разумно применять другой алгоритм, который обеспечивает, если и не минимальное число правил, то число правил, близкое к минимальному. Алгоритм перебирает не все возможные подмножества B, а только их часть.

Рассмотрим для примера множество B = {221212, 221221, 221122, 212122, 212212, 212221}. Для этого множества

$$\boldsymbol{s} = \begin{pmatrix} 2 \\ 4 \end{pmatrix},\ \mathbf{C}:\ \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 3 & 3 & 3 & 3 & 3 \\ 2 & 6 & 3 & 3 & 3 & 3 & 3 \end{array}.$$

Алгоритм 1 для этого множества правила не найдет. Чтобы успешно применить алгоритм 1, надо разбить это множество на подмножества. Предлагается разбивать его следующим образом. В матрице **C** для каждого j-го столбца, в котором нет числа card(B), производим разбиение множества B на два подмножества так, что в первом подмножестве по j-му критерию будет стоять оценка 1, а во втором – оценка 2. Т. е. производится разделение по оценкам для j-го критерия.

Например, после разбиения множества B по четвертому критерию получим два подмножества: {221122, 212122, 212212} и {221212, 221221, 212221}. Ни для одного из них алгоритм 1 ничего не дает, и пришлось бы разбивать каждое их этих подмножеств по критериям дальше. Но если произвести разбиение множества B по второму критерию, то получаются подмножества {221212, 221221, 221122} и {212122, 212212, 212221}. Алгоритм 1 дает для них правила, соответственно $212***+P_3^{1[0],2[1]}$ и $221***+P_3^{1[0],2[1]}$. Таким образом, ***алгоритм 3*** заключается в рекурсивном разбиении множества альтернатив по каждому из критериев и применении к подмножествам алгоритма 1 до тех пор, пока алгоритм 1 не выдаст единственное правило. Результатом работы этого алгоритма также является минимальное разбиение из тех, которые были получены в процессе его работы.

Вычислительная сложность алгоритма 3 равна $\Theta = O(X \cdot N^{\log_2 X})$, где X – число элементов множества альтернатив с одинаковым составом, которое требуется представить в виде правила, а N – число критериев. Действительно, алгоритм N раз пытается разбить исходное множество альтернатив на два подмножества по значениям каждого критерия. Предположим, что ему удается всегда разбивать множество ровно пополам. В результате, $\Theta(X) = N \cdot 2 \cdot \Theta(X/2)$. Поскольку $\Theta(1) = O(1)$, то, действуя по индукции, получаем $\Theta(2) = O(2N)$, $\Theta(4) = O(4N^2)$, $\Theta(8) = O(8N^3)$ и т.д. Нетрудно заметить, что данный ряд можно описать

непрерывной функцией $\Theta(X) = O(X \cdot N^{\log_2 X})$, что и распространяет нашу оценку на произвольное X.

Алгоритм 3 перебирает не все возможные разбиения. За счет этого он выполняется довольно быстро, но не может считаться точным. Например, множеству {212111, 122111, 221111, 111221, 111212, 111122} соответствуют два правила $111***+P_3^{1[0],2[1]}$ и $***111+P_3^{1[0],2[1]}$, но алгоритмом 3 они найдены не будут. Действительно, для нахождения этих правил необходимо разбить множество на два подмножества {212111, 122111, 221111} и {111221, 111212, 111122}, а это, как легко проверить, не может быть выполнено разбиением ни по какому критерию. Однако это не является большим недостатком, поскольку исследования [4,5] показывают, что решающие правила в большинстве своем имеют иерархическую структуру, основанную на значениях основного признака. Т.е. различные правила, входящие в иерархию, имеют различные значения основного признака.

Тем не менее, попытаться уменьшить неточность алгоритма 3 призван ***алгоритм 4***, который является синтезом алгоритмов 2 и 3. В процессе рекурсивного разбиения для подмножеств с небольшим числом элементов этот алгоритм использует алгоритм 2, иначе – алгоритм 3.

Заключение

Границы классов в задачах порядковой классификации могут быть представлены в виде небольшого набора решающих правил, которые можно вербализовать. По своей структуре решающие правила близки к тем, что применяются экспертами при проведении классификации.

Литература

1. Асанов, А. А. Кочин, Д. Ю. Метод выявления решающих правил в задачах экспертной классификации. // Искусственный интеллект, №2, 2002. — с. 20-31.
2. Кочин Д.Ю. Адаптивный поиск границ классов в задачах порядковой классификации // Двенадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2010). Труды конференции. — Т. 2. — М.: Физматлит, 2010. — С. 31-39.
3. Ларичев О. И., Мечитов А. И., Мошкович Е. М., Фуремс Е. М. Выявление экспертных знаний. — М.: Наука, 1989. — с. 128.
4. Ларичев, О. И., Структуры экспертных знаний в задачах классификации. // Доклады Академии Наук, т. 336, № 6, 1994. — с. 750-752.
5. Ларичев О. И., Болотов А.А. Система ДИФКЛАСС: построение полных и непротиворечивых баз экспертных знаний в задачах дифференциальной диагностики. // НТИ, Сер. 2, Информ. процессы и системы, № 9, — М.: ВИНИТИ, 1996. — с. 9-15
6. Нарыжный Е. В., Построение интеллектуальных обучающих систем, основанных на экспертных знаниях. Диссертация на соискание ученой степени кандидата технических наук, — М.: ИСА РАН, 1998.
7. Солсо Р. Когнитивная психология. — М.: Тривола, 1996.
8. Anderson, J. R. The Architecture of Cognition. // Harvard University Press, 1983.
9. Brehaut J.C., Stiell I.G., Graham I.D. Will a New Clinical Decision Rule Be Widely Used? the Case of the Canadian C-Spine Rule // Academic Emergency Medicine, Volume 13, Number 4, 2006. — pp. 413-420.
10. Ericsson, K. A. The Acquisition of Expert Perfomance: An Introduction to Some of the Issues. // The Road to Excellence: The Acquisition of Expert Perfomance in the Arts and Sciences, Sports and Games. — Hillsdale, NJ: Lawrence Erlbaum Associates, 1996. — pp. 1-51.
11. Ericsson, K. A., Lehnmann A. C. Expert and Exceptional Perfomance: Evidence of Maximal Adaptation to Task Constraints // Annual Review of Psychology, 47, 1996. — pp. 273-305.
12. Feigenbaum E. A., McCorduck P. // The 5-th Generation. — Addison-Wesley, Mass, 1983. — p. 266.
13. Fryer-Edwards K., Arnold R.M., Baile W., Tulsky J.M., Petracca F., Back A., Reflective Teaching Practices: An Approach to Teaching Communication Skills in a Small-Group Setting. // Academic Medicine, Vol. 81, No. 7, July 2006. — pp. 638-644.
14. Hunt E., COGNITIVE SCIENCE: Definition, Status and Questions. // Annual Review of Psychology, 40, 1989.
15. Kihlstrom J. F. The Cognitive Unconscious. // Science. Vol. 237, 1987. — pp. 1445-1452.
16. Larichev O. I., Moshkovich H. M., Furems E. M., Mechitov A. I., Morgoev V. K. // Knowledge Acquisition for the construction of the full and contradiction free knowledge bases, Iec ProGAMMA, Croningen, The Netherlands, 1991. — p. 240.
17. Newell A. Unified Theories of Cognition. // Cambridge, MA: Harvard University Press, 1990.
18. Newell A., Simon H. A., Human Problem Solving. // Englewood Cliffs, NJ: Prentice-Hall Inc.,1972.
19. Simon H.A., Gobet F., Richman H.B., Staszewski J.J. Goals, representations and strategies in a concept attainment task: The EPAM model. // D.L. Medin (Ed.), The psychology of learning and motivation: Vol. 37. — San Diego, CA: Academic Press, 1997. — pp. 265-290.

Кочин Дмитрий Юрьевич. Научный сотрудник ИСА РАН. Окончил Московский физико-технический институт в 2001 году и аспирантуру ИСА РАН в 2005 году. Кандидат технических наук. Автор более 25 научных работ и соавтор одной монографии. Область научных интересов: экспертная классификация, неявное обучение, принятие решений по многим критериям, построение систем поддержки диагностических решений. E-mail: dco@mail.ru