

Применение метода STEPCLASS к задачам классификации с иерархической структурой

Аннотация. Рассматривается применение метода STEPCLASS к задачам многокритериальной/многопризнаковой классификации, имеющим иерархическую структуру. Описываются два реализованных в STEPCLASS варианта сведения задачи большой размерности к задаче с иерархической структурой – декомпозиция по признакам и декомпозиция по классам. Приводятся примеры иерархической структуризации задач порядковой и номинальной классификации с применением метода STEPCLASS.

Ключевые слова: вербальный анализ решений, многокритериальная/многопризнаковая классификация, иерархическая структура, задачи большой размерности, декомпозиция по признакам, декомпозиция по классам.

Введение

Метод STEPCLASS [1], основанный на принципах вербального анализа решений (ВАР) [2], исходно разработан в целях выявления полной (с точностью до знаний эксперта) и согласованной (в определенном смысле) совокупности классифицирующих правил эксперта для номинальной многопризнаковой классификации, когда и классы, и значения признаков, описывающих объекты, подлежащие классификации, не могут быть упорядочены. Невозможность упорядочения классов определяется характером решаемой задачи. Так, в случае выявления правил эксперта для дифференциальной диагностики заболеваний, представляющей собой типичную задачу многопризнаковой классификации, классы соответствуют разным заболеваниям, т.е. каждый класс составляют объекты с одним и тем же диагнозом, и, соответственно, не могут быть упорядочены.

Впоследствии метод STEPCLASS был расширен на задачи номинально-порядковой многопризнаковой классификации [3, 4] и многокритериальной/многопризнаковой порядковой классификации [5]. Если говорить в терминах дифференциальной диагностики, задача номи-

нально-порядковой классификации, возникает, когда эксперт может упорядочить значения некоторых или всех признаков по их типичности для каждого диагноза независимо от значений других признаков. Задача порядковой классификации отличается от задачи номинально-порядковой классификации тем, что классы могут быть упорядочены. Например, применительно к диагностике какого-то одного заболевания классы могут составлять объекты с одинаковыми уровнями уверенности эксперта в их соответствии такому диагнозу или с одинаковыми уровнями тяжести такого заболевания. И в том, и в другом случае такие уровни могут быть упорядочены (например, большая, средняя и небольшая уверенность или тяжелое, среднее и легкое состояние, соответственно).

В рамках многокритериального анализа решений (MCDA) и теории принятия решений при многих критериях (MCDM) наиболее активно разрабатываются методы решения задачи порядковой классификации, которую в западной литературе называют многокритериальной сортировкой. К числу наиболее известных подходов, лежащих в основе таких методов, относятся теория полезности [6-8], подход, основанный на относительном превосходстве [9-11],

¹Работа частично поддержана Российским фондом фундаментальных исследований (проект 13-07-00168).

теория приближенных множеств с учетом доминирования [12-14] и ее расширение, требующее согласованности переменных [15], а также теория размытых множеств [16, 17]. Для решения задачи номинальной классификации известно относительно небольшое число методов на базе MCDA/MCDM. К ним относятся такие методы, как многокритериальная фильтрация [18], PROAFTN [19], TRINOMFC [20], PROCFTN [21], процедуры k-ближайшего сходства [22], NEXClass [23, 24] и SMAA-Classification [25]. Заметим, что задача номинально-порядковой классификации ставится и решается только на базе ВАР [3, 4, 26, 27]. Также заметим, что даже в случае номинальной классификации во всех перечисленных выше подходах, кроме ВАР, используется терминология, принятая в MCDA/MCDM, т.е. объекты, подлежащие классификации, чаще всего называются *альтернативами* (как в задачах выбора) и описываются оценками по многим *критериям*, а классификационные правила основаны на *предпочтениях лица, принимающего решения (ЛПР)*. Если задача порядковой классификации заключается в разбиении объектов на «лучшие», «средние» и «худшие» (разумеется, что число классов может быть больше трех и их наименования могут быть другими) с точки зрения ЛПР такая терминология вполне приемлема. Однако, как отмечалось выше, такая задача возникает и при классификации объектов в соответствии со знаниями эксперта. Кроме того, как показывается ниже в подразделе 1.1, в задаче порядковой классификации иногда уместно одновременно использовать термины «критерии» и «эксперт». В [28, 29] объясняется, почему использование такой терминологии некорректно для задач номинальной и номинально-порядковой классификации, которые решаются только на основе знаний эксперта. В таком случае термины «альтернативы» и «критерии» разумно заменить терминами «объекты» и «признаки», соответственно. Далее, для простоты, будем использовать такие термины как синонимы, если особо не указывается иное.

Все вышеперечисленные подходы и методы принадлежат к категории контролируемого обучения (*supervised learning*, т.е. обучения с участием ЛПР/эксперта). Они основаны на предварительном задании классов либо посредством типичных для них объектов, либо так называемых граничных объектов, и классифи-

кация осуществляется по принципам, которые, хотя и различны для разных методов, но близки к используемым в теории распознавания образов. Все эти подходы и методы мотивированы, до некоторой степени, ограниченной возможностью непосредственного активного участия ЛПР/эксперта в решении задачи, что отчасти обусловлено существенными затратами времени на выявление его предпочтений/знаний. Поэтому в таких подходах основное внимание уделяется не организации эффективного диалога с ЛПР/экспертом для выявления его предпочтений/знаний, а способам вывода классифицирующих правил на основе их некоторой выборки, полученной из тех или иных источников (включая, частично, ЛПР/эксперта). Математический аппарат, используемый для такого вывода, делает результаты решения задач классификации труднообъяснимыми (если вообще объяснимыми) не только для пользователя результатов решения, но и для самого ЛПР/эксперта. Кроме того, многие известные методы классификации плохо справляются (если вообще справляются) с решением задач большой размерности. Как отмечается в [30], чем больше размерность задачи (определяемая числом признаков и их значений), тем с меньшей уверенностью можно говорить о достижимости полноты классифицирующих правил и точности выводимых из нее заключений.

Метод STEPCLASS, как и все прочие методы, основанные на парадигме ВАР, требует активного участия эксперта на всех этапах решения многопризнаковой задачи классификации. Процедура выявления классифицирующих правил организована в STEPCLASS посредством моделирования объектов, подлежащих классификации, в виде комбинаций значений признаков и их запроса экспертом для анализа и классификации, благодаря чему эксперт работает в режиме, близком к своей повседневной профессиональной деятельности. Это позволяет, с одной стороны, уменьшить когнитивную нагрузку на эксперта и затраты его времени, а с другой – сделать результаты решения прозрачными и хорошо объяснимыми и для самого эксперта, и для конечных пользователей. Кроме того, в методе STEPCLASS, как и в прочих методах ВАР, реализовано несколько способов сокращения числа объектов, непосредственно анализируемых экспертом, и, при этом, гарантируется построение полных (с точностью до

знаний эксперта) и непротиворечивых баз знаний даже для задач многопризнаковой классификации большой размерности.

В статье рассматривается применение метода STEPCLASS для решения задач, имеющих, по своему характеру, иерархическую структуру. Также показывается реализованный в STEPCLASS подход к решению задач большой размерности посредством их сведения к иерархии подзадач меньшей размерности. Приводятся примеры решения таких прикладных задач на основе метода STEPCLASS.

1. Структура задачи многопризнаковой классификации

В [28,29] показывается, что основными элементами структуры задачи многопризнаковой классификации являются:

- 1) наименования классов, к которым могут принадлежать объекты данной проблемной области;
- 2) наименования признаков, значениями которых описываются объекты данной проблемной области;
- 3) набор значений (шкала) каждого признака;
- 4) матрицы допустимости значений каждого признака классам.

Если задача классификации является порядковой, в ее структуру дополнительно включается [5]:

- 5) бинарное отношение линейного порядка на множестве классов, соответствующее предпочтениям ЛППР или, если задача основана на знаниях эксперта, классы упорядочиваются либо в соответствии с уровнями уверенности эксперта в наличии у объектов некоторого общего свойства, либо в соответствии с уровнями проявления у них такого свойства;

- 6) бинарное отношение квазипорядка на шкале каждого критерия/признака, соответствующее предпочтениям ЛППР или знаниям эксперта о типичности значений признаков некоторому общему свойству.

Если задача классификации является номинально-порядковой, бинарное отношение, указанное в пункте (5), не используется, а вместо бинарного отношения квазипорядка, упомянутого в пункте (6), на шкале каждого признака, значения которого эксперт может упорядочить по типичности для каждого класса независимо от значений других признаков, строятся соот-

ветствующие бинарные отношения квазипорядка [3, 4].

Метод STEPCLASS включает две формальные процедуры структуризации задачи многопризнаковой классификации: непосредственная структуризация и структуризация на примерах. В [28,29] показывается, как каждая из этих процедур посредством специальных наводящих вопросов помогает эксперту вспомнить и перечислить элементы структуры задачи. Однако, поскольку структуризация задачи не свойственна профессиональной деятельности эксперта, ни одна из этих процедур не гарантирует полноты сформированных перечней наименований классов, признаков и их возможных значений, соответственно. Следовательно, выявленная структура либо может оказаться окончательной (достаточной для классификации), либо потребует расширения и корректировки, в том числе на этапе выявления классифицирующих правил.

Признаки, перечисленные экспертом в режиме непосредственной структуризации или выведенные из примеров объектов для каждого класса, приведенных им в режиме структуризации на примерах, могут быть «сложными». Это означает, что, на практике, значения таких признаков не проявляются у объектов, подлежащих классификации, непосредственно, а зависят от совокупности значений других признаков. В таком случае, конечный пользователь системы, основанной на знаниях эксперта, будет располагать не значениями таких признаков, а некоторой базовой информацией, из которой такие значения должны быть определены. Следовательно, после того как набор признаков и их возможных значений будет сформирован экспертом посредством любой из вышеуказанных процедур структуризации, в отношении каждого признака эксперту задается вопрос о том, как пользователь сможет определять значения признаков — непосредственно или на основе других.

Если для какого-то признака такая информация нужна, эксперту предлагается применить одну из вышеуказанных процедур структуризации, чтобы определить признаки и их возможные значения для классов, наименованиями которых являются значения такого признака. Тем самым, формируется подзадача для такого признака. Такие подзадачи строятся для каждого «сложного» признака и в результате получается иерархическая структура, состоящая из основной задачи классификации и подзадач для ее

признаков. Заметим, что подзадачи также могут включать «сложные» признаки, и, в таком случае для них создаются свои подзадачи, и так далее. В двух следующих подразделах приводятся примеры задачи порядковой классификации с такой иерархической структурой.

1.1. Задача оценки качества письменных переводов с иностранных языков

Типичным примером задачи многокритериальной порядковой классификации является оценка качества письменных переводов с иностранных языков. Актуальность такой задачи определяется несколькими факторами, среди которых следует отметить, с одной стороны, конкуренцию между переводческими фирмами в поиске и привлечении заказчиков, и, соответственно, стремление к завоеванию и поддержанию репутации благодаря высокому качеству переводов, с другой стороны - большое число лиц, претендующих на наличие у них способности выполнять письменные переводы, что зачастую неоправданно.

Как отмечается в [35], написание текста – трудная работа. Она весьма отличается от речи, которая, как бы ни была сложна, легка для понимания. Люди с раннего детства учатся говорить без сколько-нибудь формального обучения. Однако умение писать требует долгих лет обучения, и даже после этого многие люди никогда не научатся делать это хорошо. Действительно, даже самые хорошие переводчики допускают ошибки, иногда весьма серьезные. Нет никакого смысла искать переводчиков, которые всегда идеально выполняют свою работу. Поэтому для такой задачи эксперт (высококвалифицированный редактор) сформулировал три наименования классов: “хороший” перевод, практически не требующий исправления, “удовлетворительный” перевод, требующий незначительного исправления и “плохой” перевод, требующий значительного исправления или полной переделки (возможно, другим переводчиком).

При применении процедуры структуризации на примерах для первого из перечисленных выше классов эксперт ввел следующее правило: «если смысл передан хорошо, содержание передано хорошо, язык и стиль хорошие, то хороший перевод, практически не требующий исправления». Тем самым, были выявлены три критерия: «передача смысла», «передача содержания» и «язык и стиль», а также соответ-

ствующие значения этих критериев, допустимые для такого класса. Анализируя этот пример, эксперт добавил, что для этого класса допустимым является также значение «язык и стиль удовлетворительные» для критерия «язык и стиль». В результате применения этой процедуры для двух остальных классов, новые критерии не появились, но для каждого из них был сформирован набор значений, допустимых для каждого класса, включая первый.

На вопрос, как пользователь (редактор) сможет определить качество передачи смысла, эксперт ответил, что для этого необходимо оценить точность и полноту перевода. Для критерия «точность перевода» были сформулированы две оценки: «перевод отражает смысл оригинала», «перевод не отражает смысл оригинала», а для критерия «полнота» – три оценки: «все элементы смысла переданы», «небольшое число элементов смысла не передано» и «много непереданных элементов смысла». В результате, для «сложного» критерия «передача смысла» была создана подзадача с тремя классами (соответствующими его значениям): «смысл передан хорошо», «смысл передан удовлетворительно» и «смысл передан плохо» и двумя критериями: «точность» и «полнота».

Аналогично, для «сложного» критерия «передача содержания» была создана подзадача с тремя классами (соответствующими его значениям): «содержание передано хорошо», «содержание передано удовлетворительно» и «содержание передано плохо» и двумя критериями: «логика» и «фактические и/или концептуальные ошибки».

Что касается третьего критерия «язык и стиль», для получения оценки по нему необходимо определить качество перевода по таким критериям, как «гладкость текста перевода», «терминология», «идиоматичность словосочетаний» и «грамматика». В свою очередь, для оценки гладкости текста перевода необходима информация о связи между предложениями и легкости чтения, а для оценки грамматики – о правописании и пунктуации. В результате, получилась иерархическая структура задачи, показанная на Рис. 1.

Эксперт мог формировать правила оценки качества переводов с применением системы STEPCLASS в любом порядке таких подзадач, включая основную (корневую) задачу. Совокупность таких правил составила базу знаний системы поддержки принятия решений (СППР)

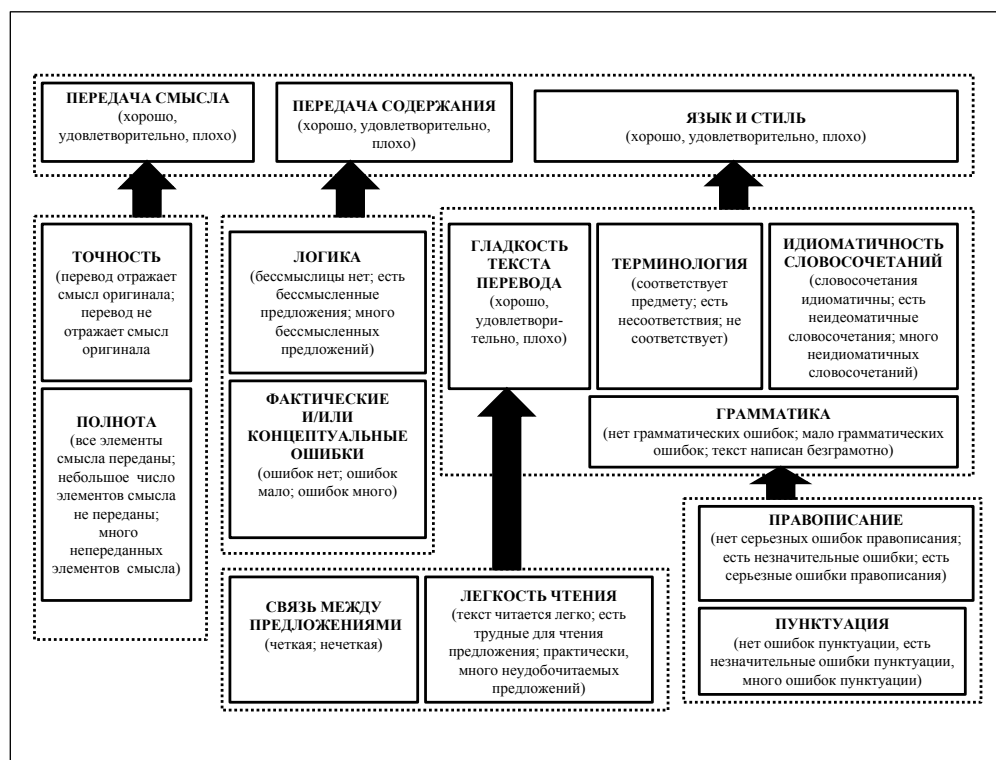


Рис. 1. Иерархическая структура задачи оценки качества письменных переводов

«КАЧЕСТВО ПЕРЕВОДА». Для оценки качества конкретного перевода ее пользователю нужно ввести информацию в следующем порядке: «точность», «полнота»; «логика», «фактические и/или концептуальные ошибки»; «связь между предложениями», «легкость чтения»; «идиоматичность словосочетаний», «правописание», «пунктуация», и «терминология». После ввода оценок по критериям «точность» и «полнота» СППР найдет в базе знаний правило для сочетания таких оценок, определит класс, к которому такое сочетание относится, и использует его впоследствии как значение критерия «передача смысла».

Аналогично, после ввода пользователем оценок по критериям «логика» и «фактические и/или концептуальные ошибки» СППР, на основе соответствующего правила определит значение критерия «передача содержания». После ввода оценок по критериям «связь между предложениями» и «легкость чтения» СППР автоматически найдет в базе знаний правило, которое позволит определить значение критерия «гладкость перевода», а после ввода оценок по критериям «правописание» и «пунктуация» – значение критерия «грамматика». Затем, на основе таких значений в сочетании со значени-

ями критериев «терминология» и «идиоматичность словосочетаний», СППР автоматически определит значение критерия «язык и стиль», и, наконец, для сочетания соответствующих значений критериев «передача смысла», «передача содержания» и «язык и стиль» в базе знаний будет найдено правило для оценки качества такого перевода.

1.2. Задача оценки кредитоспособности предприятий на основе финансовых коэффициентов

В системе STEPCLASS в задачах с иерархической структурой переход от подзадач нижнего уровня к подзадаче непосредственно верхнего уровня может строиться не только на основе правил эксперта «если..., то...», но и на основе заданных им формул. Для этого система включает редактор формул, анализатор формул и транслятор формул.

Покажем, как работает такой принцип с использованием приведенного в [5] упрощенного примера задачи оценки кредитоспособности предприятий, подающих заявки на получение кредита.

Для структуризации этой задачи были использованы данные из [36] с некоторыми изме-

Табл. 1. Критерии и их значения, допустимые для классов

Критерии	Значения (шкалы) критериев	Допустимость для классов		
		Высокая кредитоспособность	Средняя кредитоспособность	Низкая кредитоспособность
Коэффициент текущей ликвидности	$K_{ТЛ} \geq 0.2$	1	1	0
	$0.15 \leq K_{ТЛ} < 0.2$	1	1	1
	$K_{ТЛ} < 0.15$	0	1	1
Коэффициент срочной ликвидности	$K_{СЛ} \geq 1.0$	1	1	1
	$0.5 \leq K_{СЛ} < 1.0$	1	1	1
	$K_{СЛ} < 0.5$	1	1	0
Коэффициент абсолютной ликвидности	$K_{АЛ} \geq 2.0$	1	1	0
	$1.0 \leq K_{АЛ} < 2$	1	1	1
	$K_{СЛ} < 1.0$	0	1	1
Коэффициент автономии	$K_A \geq 0.7$	1	1	1
	$0.5 \leq K_A < 0.7$	1	1	1
	$K_A < 0.5$	1	1	1

нениями. Соответствующие наименования классов, критерии и их оценки приводятся в Табл.1. Наименования классов и значения каждого критерия приводятся в порядке убывания предпочтительности для ЛПП (например, начальника кредитного департамента банка).

Коэффициенты текущей ликвидности, срочной ликвидности, абсолютной ликвидности и автономии рассчитываются по следующим формулам:

$$K_{ТЛ} = \frac{A_1 + A_2 + A_3}{\Pi_1 + \Pi_2}; K_{СЛ} = \frac{A_1 + A_2}{\Pi_1 + \Pi_2};$$

$$K_{АЛ} = \frac{A_1}{\Pi_1 + \Pi_2} \text{ и } K_A = \frac{\Pi_4}{A_1 + A_2 + A_3 + A_4},$$

где: A_1 – наиболее ликвидные активы (денежные средства предприятия и краткосрочные финансовые вложения);

A_2 – быстро реализуемые активы (дебиторская задолженность и прочие активы);

A_3 – медленно реализуемые активы (запасы без статей «Прочие запасы и затраты» и «Расходы будущих периодов»), а также статьи из раздела I баланса «Долгосрочные финансовые вложения», уменьшенные на величину статьи «Вложения в уставные фонды других предприятий»;

A_4 – труднореализуемые активы: статьи раздела I актива баланса «Внеоборотные активы»;

Π_1 – наиболее срочные пассивы: кредиторская задолженность, прочие пассивы, а также ссуды, не погашенные в срок;

Π_2 – краткосрочные пассивы: краткосрочные кредиты и заемные средства;

Π_4 – постоянные пассивы: статьи раздела III баланса «Капитал и резервы» и отдельные статьи раздела V баланса, не вошедшие в предыдущие группы: «Доходы будущих периодов» и «Резервы предстоящих расходов».

На Рис. 2 показывается иерархическая структура задачи оценки кредитоспособности предприятий на основе финансовых коэффициентов.

Для построения базы знаний СППР «ОЦЕНКА КРЕДИТОСПОСОБНОСТИ» ЛПП сформировал для основной задачи правила, позволяющие отнести комбинации значений этих четырех критериев (коэффициентов) к одному из трех классов, соответствующих уровням кредитоспособности. Затем, с помощью редактора формул ЛПП ввел правила в виде формул расчета коэффициентов, а также их параметров ($A_1 - A_4$, Π_1 , Π_2 и Π_4) и указал, тем самым, какие данные должен ввести пользователь для этого.

Например, для расчета параметра A_1 пользователь должен ввести данные таких статей баланса предприятия, как «Денежные средства» и «Краткосрочные финансовые вложения», которые СППР «ОЦЕНКА КРЕДИТОСПОСОБНОСТИ» просуммирует по формуле, введенной ЛПП. После того как пользователь введет значения по прочим запрошенным СППР статьям баланса, СППР автоматически вычислит

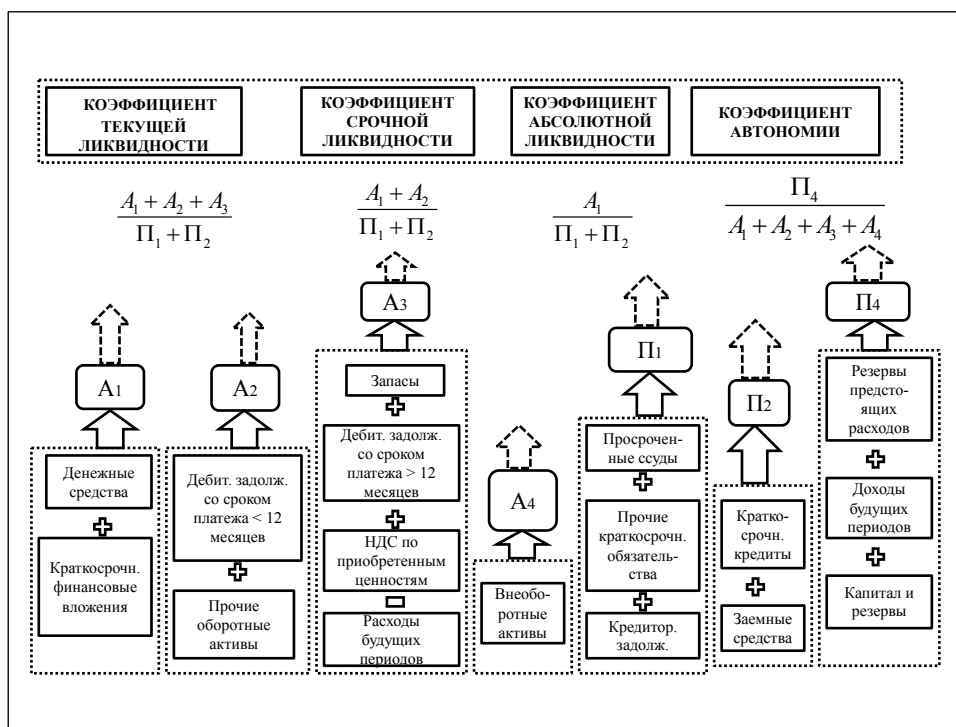


Рис. 2. Иерархическая структура задачи оценки кредитоспособности предприятий на основе финансовых коэффициентов

значения остальных соответствующих параметров, а затем – значения коэффициентов, являющихся критериями в данной задаче. Наконец, для найденной комбинации значений коэффициентов СППР найдет в базе знаний правило, по которому определит уровень кредитоспособности данного предприятия.

2. Задачи многопризнаковой классификации большой размерности

Эффективность существующих подходов к извлечению знаний страдает от так называемого «проклятия размерности» [37]. Как объясняется в [30], во многом это обусловлено наличием в структуре задачи избыточных и нерелевантных признаков.

Метод STEPCLASS свободен от избыточности признаков и отсутствия у них релевантности, поскольку такие признаки определяются самим экспертом на этапе структуризации (и, в случае необходимости, добавляются на этапе классификации), и, даже если эксперт задаст на этапе структуризации «лишние» признаки, он никогда не будет их использовать на этапе классификации.

Кроме того, благодаря особой организации диалога с экспертом для выявления его классифицирующих правил, метод STEPCLASS позволяет решать задачи большей размерности, чем остальные методы, основанные на принципах ВАР и предназначенные для решения задач порядковой классификации (ORCLASS [26, 31], CLARA [32], DIFCLASS [33] и CYCLE [34]) и номинально-порядковой классификации (NORCLASS [26, 27]). Во всех этих методах объект, подлежащий классификации, предъявляется эксперту в форме его описания значениями всех признаков одновременно, что, с одной стороны, увеличивает когнитивную нагрузку для эксперта, поскольку ему приходится анализировать объект сразу с учетом всех признаков (в том числе и избыточных), а с другой стороны, требует предъявления эксперту большого числа объектов для непосредственной классификации. В системе STEPCLASS для каждого классифицируемого объекта также известны значения всех признаков, но эксперт сам запрашивает только те признаки, значения которых он хочет узнать. Тем не менее, если эксперт на этапе структуризации определяет большое число признаков и большое число значений каждого признака, число гипотетиче-

ски возможных объектов, подлежащих классификации, может достигать нескольких миллионов или даже миллиардов (число таких объектов равно мощности Декартова произведения шкал признаков). Таким образом, практически неосуществимо построить полную базу знаний для такой задачи за приемлемое время, несмотря на то, что для построения полной базы знаний непосредственно классифицируется лишь часть объектов. Однако на практике эксперт не работает с полным набором признаков одновременно; значения некоторых из них он анализирует либо независимо от других, либо с учетом предварительного решения, принятого на основе анализа комбинаций значений других признаков. Это соображение реализовано в системе STEPCLASS посредством декомпозиции задачи большой размерности и сведения, тем самым, ее структуры в иерархическую структуру подзадач приемлемой размерности.

В STEPCLASS используется два способа декомпозиции задачи классификации: декомпозиция по признакам и декомпозиция по классам.

Декомпозиция по признакам выполняется следующим образом: эксперт анализирует набор признаков, определенных им на этапе структуризации, чтобы разделить их на подгруппы, характеризующие разные аспекты, которые ему необходимо знать, чтобы классифицировать объекты. Другими словами, такие подгруппы формируются так, чтобы эксперт мог сформировать правила для классификации всех гипотетически возможных комбинаций значений признаков каждой подгруппы. В отличие от задач классификации, исходно имеющих иерархическую структуру, в случае задач большой размерности, иерархическая структура подзадач строится не сверху вниз (от основной задачи к подзадачам), а снизу вверх (от подзадач к основной задаче). Наименования признаков и их значения, определенные экспертом на этапе классификации, используются в подзадачах нижнего уровня. Наименованиями признаков верхнего уровня становятся те аспекты, для которых формировались соответствующие подзадачи, а значениями каждого такого признака — наименование классов соответствующей подзадачи.

Декомпозиция по классам используется либо с самого начала, когда на этапе структуризации выясняется, что у некоторых признаков нет значений, допустимых для части классов, либо

после выявления классифицирующих правил эксперта (возможно с использованием декомпозиции по признакам), если эксперт не смог однозначно разделить некоторую или всю совокупность объектов на исходно названные им классы, и отнес их сразу к нескольким классам. И в том, и в другом случае, эксперт осуществляет классификацию объектов внутри каждой группы классов, и может использовать для этого как исходно названные им признаки (имеющие хотя бы одно допустимое значение для каждого класса такой группы), так и дополнительные признаки, необходимые для дифференциации объектов внутри каждой группы.

В двух следующих подразделах приводятся примеры декомпозиции задач номинальной классификации большой размерности для каждого из вышеуказанных случаев.

2.1. Задача дифференциальной диагностики бронхиальной астмы у детей

Для формирования базы знаний СППР «ЭСБАД» на этапе структуризации врач-эксперт на основе предварительного методологического анализа различных заболеваний, имеющих сходные с бронхиальной астмой клинические и лабораторно-функциональные проявления, определил 12 следующих заболеваний [38]: аллергический ларинготрахеит, аллергический бронхит, экзогенный аллергический альвеолит, хроническая пневмония, синдром Вильямса-Кемпбелла, муковисцидоз, инородное тело трахеобронхиального дерева, опухоли трахеи и бронхов, коклюш, дыхательные нарушения нейрогенного и психогенного характера, хронический бронхит и облитерирующий бронхиолит. Такие заболевания использовались в данной задаче в качестве наименований классов.

Посредством процедуры непосредственной структуризации был сформирован набор из 34 признаков с 5-10 значениями каждый. Выявление классифицирующих (диагностических) правил эксперта в задаче такой размерности было не только практически неосуществимо за реальное время, но и нецелесообразно. Эту совокупность признаков эксперт объединил в группы, соответствующие разделам унифицированной истории болезни пульмонологического больного: жалобы, осмотр (клинические данные), анамнез, данные клинического анализа крови, данные рентгенологического исследования и данные функции внешнего дыхания.

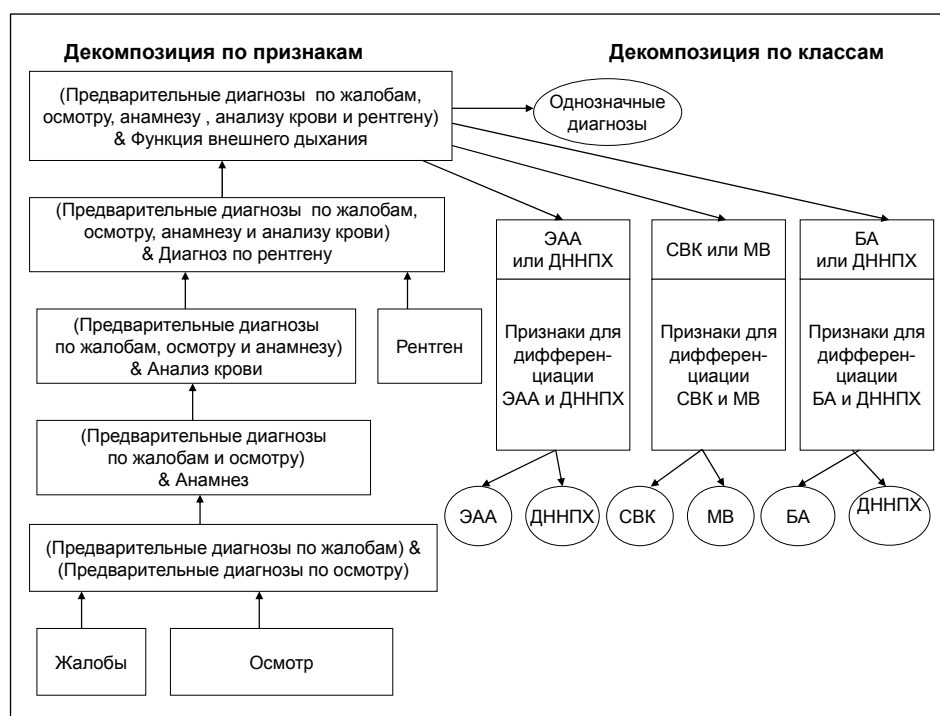


Рис. 3. Структура задачи диагностики бронхиальной астмы у детей

Это позволило построить иерархическую структуру задачи (Рис. 3), которая решалась экспертом, начиная с подзадач нижнего уровня. Сначала была построена база знаний для подзадачи «Жалобы» с семью признаками, в которой к исходно названным классам добавлялись слова «Предварительный диагноз по жалобам». Поскольку система STEPCLASS позволяет добавлять наименования классов на этапе классификации, а также в связи с тем, что признаков только этой подзадачи эксперту было явно недостаточно для дифференциальной диагностики большинства комбинаций значений признаков, эксперт мог относить такие комбинации сразу к нескольким заболеваниям (классам) с использованием союза «или» (каждая такая совокупность классов считалась новым классом).

Затем эксперт аналогичным образом сформулировал диагностические правила для подзадачи «Осмотр» (5 признаков). После этого ему было предложено решить подзадачу с двумя признаками: «Предварительный диагноз по жалобам» и «Предварительный диагноз по осмотру», значениями которых были наименования классов, образовавшихся в результате решения подзадач «Жалобы» и «Осмотр», соответственно, а к наименованиям исходно названных и новых классов, добавлялись слова «Предварительный диагноз по жалобам и осмотру». Перечень таких

классов составил значения одноименного признака в подзадаче непосредственно верхнего уровня, который был добавлен к исходно названным признакам группы «Анамнез» (6 исходных признаков), и так далее. После того как эксперт сформировал правила в задаче самого верхнего уровня, оказалось, что для некоторых объектов так и не был поставлен однозначный диагноз. Такими неоднозначными диагнозами были, например (Рис.3) экзогенный аллергический альвеолит (ЭАА) или дыхательные нарушения нейрогенного и психогенного характера (ДННПХ), синдром Вильямса-Кемпбелла (СВК) или муковисцидоз (МВ), и бронхиальная астма (БА) или дыхательные нарушения нейрогенного и психогенного характера. Для выявления правил эксперта, позволяющих дифференцировать такие диагнозы, была применена декомпозиция по классам с использованием дополнительных признаков. Например, для дифференциации бронхиальной астмы и дыхательных нарушений нейрогенного и психогенного характера эксперт решал подзадачу с такими дополнительными признаками, как наличие физической травмы в анамнезе, скарификационная проба, уровень общего иммуноглобулина Е, уровень специфического иммуноглобулина Е, наличие неврологической патологии при обследовании; эффект от применения бронхолитических пре-

паратов, эффект от применения кортикостероидных препаратов и эффект от неврологической терапии. Для дифференциации синдрома Вильямса-Кемпбелла (СВК) и муковисцидоза эксперт использовал такие признаки, как начало и течение заболевания, наличие в семейном анамнезе мертворожденных или выкидышей, наличие патологии желудочно-кишечного тракта, хлориды пота, копрологическое исследование, определение гена муковисцидоза и данные бронхоскопии.

2.2. Задача диагностики генетических типов карбонатных пород

При формировании базы знаний СППР «REEF» для диагностики карбонатных отложений [39] на этапе структуризации эксперт ввел перечень из 87 наименований классов – литологических разновидностей, а также более 100 признаков с 5 – 10 значениями.

Таким образом, число гипотетически возможных объектов, подлежащих диагностике, равное числу всех возможных комбинаций значений признаков, исчислялось десятками миллиардов. Однако, как показал анализ проблемной области, проведенный экспертом [40], задача диагностики генетических типов карбонатных отложений может быть представлена в виде иерархии подзадач меньшей и, вообще говоря, приемлемой для эксперта размерности, посредством декомпозиции по классам и последующей декомпозиции по признакам.

Иерархическая структура задачи строилась следующим образом: в подзадачах верхних уровней классами являются группы или подгруппы литотипов, устанавливаемые только по части признаков, отражающих компонентный и вещественный состав, с менее детализированными шкалами их значений, чем требуется для точной диагностики конкретного литотипа. С такой целью для каждого признака все значения, допустимые хотя бы для одного литотипа, входящего в группу, являющуюся классом в данной подзадаче, сводились в одно «укрупненное» значение этого признака. Так, шкалы признаков, выделенных в структуру корневой подзадачи, были укрупнены таким образом, чтобы комбинации их значений могли бы быть отнесены экспертом к таким агрегированным группам литотипов, как, например, известняк, доломит, мергель, глина, песчаник, карбонатная брекчия и т.п. Для каждой из этих групп

строились подзадачи следующего уровня, диагностируемые объекты которых разбивались на более мелкие подгруппы. Например, объекты, отнесенные к группе «Известняк», разбивались на следующем уровне на подгруппы: «известняк с макрокомпонентами», «пелитоморфный или тонкокристаллический известняк» и «комковатый известняк». В свою очередь, объекты первой из этих подгрупп далее подразделялись на биогермный известняк, оолитовый известняк, обломочный известняк, биоморфный известняк, детритовый известняк, шламовый известняк и т.п. и так до тех пор, пока в очередной подзадаче эксперту не удавалось сформировать правила для диагностируемых литотипов.

В каждой подзадаче могли использоваться как признаки, уже входящие в структуру подзадач предыдущих уровней, но уже с более детализированными шкалами, так и еще не учтенные признаки. При этом из всего множества значений признаков, сформированного на этапе предварительной структуризации, в подзадачу включались только признаки, имеющие значения, допустимые хотя бы для одного класса, входящего в ту их подгруппу, для которой построена эта подзадача.

Разработанный вариант декомпозиции общей задачи в виде дерева подзадач позволял установить диагноз не более чем за семь шагов, т.е. посредством решения от трех до семи промежуточных подзадач, в каждой из которых используется от двух до 15 признаков.

Еще одна особенность задачи диагностики карбонатных отложений заключалась в том, что эксперт формировал правила, используя как первичные, непосредственно наблюдаемые или измеряемые значения признаков, так и «сложные» признаки, значения которых могут быть получены из первичных признаков только в результате вычислений по специальным формулам (например, доли компонентов в их суммарном содержании). Однако пользователь системы может располагать лишь значениями первичных признаков. Поэтому при структуризации задачи применялась также декомпозиция по признакам, которая использовалась не столько для уменьшения размерности задачи, сколько для обеспечения возможности включения в базу знаний формул для вычисления значений «сложных» признаков по значениям первичных.

В подзадаче, формируемой для «сложного» признака, классами являются его значения,

преобразованные из вербальных формулировок в количественную интервальную шкалу, а признаками – первичные признаки, по значениям которых вычисляется значение, сравниваемое со шкалой агрегированного признака для выбора соответствующего интервала. Благодаря этому пользователю СППР «REEF» достаточно вводить только значения первичных признаков, по которым автоматически будут вычисляться значения соответствующих «сложных» признаков, используемых в классифицирующих (диагностических) правилах эксперта.

Аналогичный подход к декомпозиции и структуризации был использован для построения СППР «SAND» для диагностики генетических типов песчаных тел в терригенных толщах [39].

Заключение

Метод STEPCLASS, основанный на парадигме ВАР, предназначен для выявления полной (с точностью до знаний эксперта) и согласованной совокупности правил эксперта в задачах многокритериальной/многопризнаковой классификации. Одной из его особенностей является формальный подход к структуризации задачи классификации с возможностью расширения ее структуры на этапе классификации. Структуризация задачи заключается не только в выявлении таких ее важных параметров, как наименования классов, признаков и их возможных значений, выяснении допустимости значений признаков для классов и, если возможно, упорядочении классов и/или значений признаков по типичности для классов, но и определении подзадач, необходимых для решения основной задачи. Задача классификации может иметь, по своему характеру, иерархическую структуру, когда для определения значений одних признаков требуется знать значения других признаков. В системе, реализующей метод STEPCLASS, такая иерархическая структура строится на этапе предварительной структуризации, а переход от подзадач нижнего уровня к подзадачам непосредственно верхнего уровня может осуществляться либо в соответствии с правилами эксперта «если..., то ...», либо по введенным экспертом математическим формулам с использованием редактора, анализатора и транслятора формул. Кроме того, метод STEPCLASS позволяет эффективно выявлять классифицирующие правила эксперта в задачах

большой размерности (определяемой числом признаков и их значений) посредством сведения таких задач к задачам с иерархической структурой с использованием декомпозиции по признакам и декомпозиции по классам.

Литература

1. Фуремс Е.М., Гнеденко Л.С. STEPCLASS – система извлечения экспертных знаний и проведения экспертизы для решения диагностических задач// Научно-техническая информация. Серия 2. 1996, № 9, с. 16-20.
2. Ларичев О.И., Мошкович Е.М. Качественные методы принятия решений, М.: Физматлит. 1996.
3. Фуремс Е.М. Модифицированный метод экспертной номинально-порядковой классификации// Искусственный интеллект и принятие решений. 2010, №4, с. 81-93.
4. Furems Eugenia M. Dominance-Based Extension of STEPCLASS for Multiattribute Nominal Classification// IJITDM, 2013, 12(5), p. 905–925.
5. Фуремс Е.М. Многокритериальная порядковая классификация на основе метода STEPCLASS// Искусственный интеллект и принятие решений, 2012, № 4, С. 95-100.
6. Von Neumann J., Morgenstern O. Theory of Games and Economic Behavior, 2nd ed., Princeton: Princeton University Press, 1947.
7. Кини Р., Райфа Х. Принятие решений при многих критериях: предпочтения и замещения. М.: Радио и связь. 1981.
8. Jacquet-Lagrange E., Siskos Y. Assessing a set of additive utility functions for multicriteria decision making: The UTA method// EJOR, 1982, No. 10, p. 151-164.
9. Roy B., Bouyssou D. Aide multicritère à la décision: méthodes et cas, Paris: Economica, 1993.
10. Mousseau V., Slowinski R., Zielniewicz P. A user-oriented implementation of the ELECTRE-TRI method integrating preference elicitation support// Computers & Operations Research, 2000, No. 27, p. 757-777.
11. Figueira J., Mousseau V., Roy B. Electre methods// In: Figueira J., Greco S., Ehrgott M. (eds) Multiple Criteria Decision Analysis: State of the Art Surveys, Boston: Springer-Verlag, 2005, p. 133-162.
12. Greco S., Matarazzo B., Slowinski R. Rough sets theory for multicriteria decision analysis// EJOR, 2001, 129(1), p. 1-47.
13. Greco S., Matarazzo B., Slowinski R. Dominance-Based Rough Set Approach as a Proper Way of Handling Graduality in Rough Set Theory// T. Rough Sets, 2007, No. 7, p. 36-52.
14. Błaszczyński J., Greco S. and Slowinski R. Inductive discovery of laws using monotonic rules// Eng. Appl. of AI, 2012, 25(2), p. 284-294.
15. Błaszczyński J., Greco S., Slowinski R. Multi-criteria classification – A new scheme for application of dominance-based decision rules// EJOR, 2007, No. 181, p. 1030–1044.
16. Zadeh L., Fuzzy sets// Information Control, 1965, No. 8, p. 338-353.
17. Fernandez E., Navarro J.. A new approach to multi-criteria sorting based on fuzzy outranking relations: The THESEUS method// EJOR, 2011, 213(2), p. 405-413.

18. Perny P. Multicriteria filtering methods based on concordance and non-concordance principles// *Annals of Operations Research*, 1980, 80(1), p. 137-166.
19. Belacel N. Multicriteria assignment method PROAFTN: Methodology and medical applications// *EJOR*, 2000, No. 125, p. 175-183.
20. Leger J., Martel J.-M. A multicriteria assignment procedure for a nominal sorting problematic// *EJOR*, 2002, No. 125, p. 175-183.
21. Belacel N., Boulassel M.R. Multicriteria Fuzzy Classification Procedure PROCFTN: Methodology and Medical Application// *Fuzzy Sets and Systems*, 2004, 141(2), p. 203-217.
22. Belacel N. The k-Closest Resemblance Approach for Multiple Criteria Classification problems// In: Thi Hoai, Pham Dinh Tao (eds) *Modelling, Computation and Optimization Information and Management Sciences*, London: Hermes Sciences Publishing, 2004, p. 525-534.
23. Rigopoulos G., Askounis D., Metaxiotis K. NeXClass: A Decision Support System for non-ordered Multicriteria Classification, *IJITDM*, 2010, 9(1), p. 53-79.
24. Rigopoulos G., Anagnostopoulos K. Fuzzy Multicriteria Assignment for Nominal Classification Methodology and Application in Evaluation of Greek Bank's Electronic Payment Retailers// *IJITDM*, 2010, 9(3), p. 1-18.
25. Yevseyeva I., Miettinen K., Salminen P., Lahdelma R. SMAA-Classification: A new method for nominal classification, Helsinki School of Economics, Working Paper, 2007, <http://hsepubl.lib.hse./pdf/wp/w422.pdf>.
26. Ларичев О.И., Мечитов А.И., Мошкович Е.М., Фуремс Е.М. Выявление экспертных знаний. М.: Наука. 1989.
27. 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. Groningen: Iec ProGAMMA, 1991.
28. Фуремс Е.М. Структуризация задач классификации, основанных на знаниях// *Информационные технологии и вычислительные системы*. 2007, № 3, с. 7-17.
29. Furems Eugenia M. Domain Structuring For Knowledge-Based Multiattribute Classification (A Verbal Decision Analysis Approach)// *TOP*, 2011, No. 19, p. 402-420.
30. Puppe Frank. Knowledge Reuse among Diagnostic Problem-Solving Methods in the Shell-Kit D3// *International Journal of Human-Computer Studies*, 1998, No. 49, p. 627-649.
31. Larichev O. I., Moshkovich H.M. An Approach to Ordinal Classification Problems// *International Transactions of Operational Research*, 1994, 1(3), p. 375-385.
32. Larichev O., Asanov A., Naryzhny Y. Effectiveness evaluation of expert classification methods// *EJOR*, 2002, 138(2), p. 260-273.
33. Ларичев О. И., Болотов А. А. Система ДИФКЛАСС: построение полных и непротиворечивых баз экспертных знаний в задачах дифференциальной диагностики// *Научно-техническая информация. Серия 2*. 1996, № 9, с. 9-15.
34. Асанов А.А., Кочин Д.Ю. Метод выявления решающих правил в задачах экспертной классификации// *Научно-теоретический журнал «Искусственный интеллект» Института Проблем Искусственного Интеллекта Украинской Академии Наук*. 2002, № 26 с. 20-21.
35. Mossop Brianю Revising and Editing for Translators, Manchester: St Jerome Publishing, 2001.
36. Куликов Н.И., Чайникова Л.Н. Оценка кредитоспособности предприятия-заемщика: методические указания, Тамбов: Издательство Тамбовского государственного технического университета. 2007.
37. Huang S. Dimensionality Reduction in Automatic Knowledge Acquisition: A Simple Greedy Search Approach// *IEEE Transactions on Knowledge and Data Engineering*, 2003 15(6), p. 1364-1373.
38. Furems E., Sokolova L. Experts' knowledge acquisition for differential diagnostics of bronchial asthma in children in STEPCLASS environment// *IJTPM*, 2011, 11(1), p. 68-84.
39. Порожун В.И., Фарбирович В.П., Фуремс Е.М. Построение систем диагностики генетических типов терригенных и карбонатных пород в среде STEPCLASS// *Научно-техническая информация. Серия 2*. 1996, № 9, с. 27-31.
40. Фарбирович В.П. Методологический анализ при формировании экспертной системы диагностики карбонатов// В сборнике «Геологическое изучение и использование недр: Научно-техническая информация. М.: Геоинформмарк, № 5, с. 19-27.

Фуремс Евгения Марковна. Ведущий научный сотрудник ИСА РАН. Окончила МЭСИ им. С. Орджоникидзе в 1973 году. Кандидат технических наук. Автор более 60 научных работ и соавтор трех монографий. Область научных интересов: принятие решений по многим критериям, комбинаторная оптимизация, построение систем поддержки диагностических решений, в том числе в области геологии, медицины и инвестиций. E-mail: fem@mail.ru