УДК 62-50

КЛАССИФИКАЦИЯ ОБЪЕКТОВ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ СПЕЦИАЛИСТА ПРИ ПРОЕКТИРОВАНИИ ПРОФЕССИОНАЛЬНЫХ И ОБРАЗОВАТЕЛЬНЫХ СТАНДАРТОВ

В.В. Никитин[(1)], С.В. Мальцева[(1)], А.А. Дорофеюк[(2)], А.С. Мандель[(2)], А.Л. Чернявский[(2)]

[(1)] *Государственный университет — Высшая школа экономики, г. Москва;*
[(2)] *Институт проблем управления им. В.А. Трапезникова РАН, г. Москва*

Показано, что задача формирования обобщающих понятий, характеризующих сферу профессиональной деятельности специалиста, сводится к задаче диагонализации некоторой матрицы связей. Предложена человеко-машинная процедура её решения.

ВВЕДЕНИЕ

Одна из центральных задач управления процессами подготовки профессиональных кадров заключается в разработке государственных профессиональных и образовательных стандартов. Решение этой задачи представляет собой многоэтапный процесс анализа, оценивания и обработки больших массивов информации [1].

Важный начальный этап этого процесса состоит в определении объектов профессиональной деятельности (ОПД) специалиста, относительно которых «выстраиваются» его профессиональная деятельность и профессиональные компетенции. Образовательные программы, в свою очередь, должны быть построены таким образом, чтобы обеспечивать формирование у обучающихся заданного набора компетенций.

Формирование множества ОПД для определенной предметной области и их классификация представляют собой слабо формализуемую и трудоемкую задачу. Это, в первую очередь, сказывается на формировании содержания образования в высокотехнологичных, динамически развивающихся отраслях, для которых характерны частое появление новых объектов и связанных с ними понятий, существование большого числа синонимичных понятий, быстрое устаревание некоторых типов объектов, изменения в описании и интерпретации отдельных понятий. Особенно это проявляется на комплексных объектах, которые являются сложной агрегацией множества простых объектов и определяются обобщающими понятиями. Важно, что именно эти понятия, как правило, используются в качестве определений ОПД в стандартах высшего профессионального образования.

Сегодня общепризнанным способом идентификации таких сложных объектов служат онтологии, на основе которых можно определить не только набор более простых понятий, составляющих сложное понятие, но и связи между сложными понятиями, а также степень сходства и различия между понятиями, что позволяет выделить набор обобщающих понятий, на основе которых можно идентифицировать основные ОПД. Эти понятия являются образующими понятиями для формирования набора компетенций специалиста.

Создание онтологий в различных предметных областях представляет собой одну из самых актуальных задач. Её решением на разных уровнях занимаются исследовательские центры и международные организации. Примером одного из наиболее крупных проектов может служить проект UN SPSC (United Nations Standard Products and Services Code — Стандартный классификатор товаров и услуг), разработка которого ведется экспертной группой UNDO (United Nations Development Organization) и крупной компанией «Dun & Bradstreet». Проект развивается на основе опыта использования транснациональных систем и стандартов

(SWIFT, EDIFACT и др.) [1]. В настоящее время этот классификатор насчитывает более 13 тыс. категорий товаров и услуг.

Задача формирования обобщающих понятий, используемых при проектировании профессиональных и образовательных стандартов для той или иной предметной области, может быть понята как задача формирования классов простых (исходных) ОПД. Далее под ОПД всюду понимаются именно исходные ОПД. Сложность этой задачи связана, прежде всего, с объемом исходной информации. Число ОПД исчисляется сотнями, а число классов, в которые их нужно объединить, — десятками. При решении задачи вручную эксперт вынужден резко ограничивать число рассматриваемых вариантов классификации, основываясь на своих субъективных оценках. Для повышения объективности и обоснованности классификации желательно использовать формализованные критерии и многовариантные процедуры.

1. ФОРМАЛИЗАЦИЯ ПОСТАНОВКИ ЗАДАЧИ

Вначале введем необходимые обозначения:

$V = \{v_n \mid n = 1, ..., N_V\}$ — множество ОПД, элемент множества представляет собой код ОПД;

$\chi = \{\chi_n \mid n = 1, ..., N_V\}$ — вектор наименований ОПД, элементы вектора — наименования ОПД;

$\mu = \{\mu_n|, n = 1, ..., N_V\}$ — вектор описаний ОПД, элементы вектора — текстовые описания ОПД.

Пусть $R = \{r_{ij}\}$, $i, j = 1, ..., N_V$, $i \neq j$, — матрица связей между ОПД, элемент r_{ij} которой служит численной характеристикой степени связи между i-м и j-м ОПД ($0 \leqslant r_{ij} \leqslant 1$).

Цель обработки матрицы связей состоит в разбиении всего множества элементов $v_1, ..., v_{N_v}$ на такие непересекающиеся подмножества (агрегаты) $q'_1, ..., q'_{N'}$, чтобы связи r_{ij} между элементами, попавшими в одно подмножество, были возможно больше, а между элементами, попавшими в разные подмножества, — возможно меньше. Такую задачу будем называть задачей классификации (агрегирования) исходного множества элементов, а полученные классы (агрегаты) $q'_1, ..., q'_{N'}$ интерпретировать как классы объектов профессиональной деятельности.

Эту же задачу можно описать и несколько иначе. Если различным образом нумеровать элементы $v_1, ..., v_{N_v}$, это будет соответствовать перестановкам строк и столбцов матрицы связей R. Рассмотрим задачу нахождения такой нумерации элементов (т. е. такого расположения строк и соответствующих столбцов матрицы R и выделения таких непересекающихся квадратных подматриц вдоль

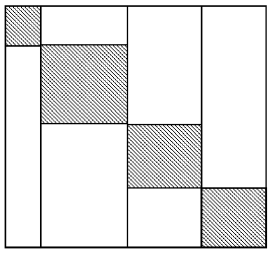

Рис. 1. Диагонализированная матрица связей

главной диагонали преобразованной подобным образом матрицы связей, чтобы элементы — числа r_{ij} — каждой выделенной подматрицы были возможно больше, а числа r_{ij}, расположенные вне этих подматриц, — возможно меньше. Будем также требовать, чтобы выделенные квадратные подматрицы полностью покрывали главную диагональ матрицы связи. Условно выделенные подматрицы с большими компонентами можно изобразить в виде квадратов, вообще говоря, разного размера (рис. 1), расположенных вдоль главной диагонали матрицы R.

Подмножество элементов, которым соответствуют строки (или столбцы) преобразованной матрицы связей R, образующие одну из выделенных подматриц, — это и есть один из классов (сильно связанных агрегатов), о которых шла речь. Имея в виду такую интерпретацию, рассматриваемую задачу называют также задачей диагонализации матрицы связей.

Приведенное описание задачи ещё не есть точная её постановка, так как последняя предполагает формальное определение того, что означает требование, чтобы в выделенных подматрицах «компоненты были возможно большими, а вне их — возможно меньшими». С этой целью обычно вводится в рассмотрение подходящий функционал (критерий качества классификации), зависящий от того, как именно разбиты элементы на N' классов $q'_1, ..., q'_{N'}$, и такой, что «хороший» в интуитивном понимании способ классификации элементов соответствует экстремальному значению функционала. Как только подходящий функционал сформулирован, задача сводится к конструированию процедуры его экстремизации. Такой подход к решению задачи мы будем называть вариационным.

Этот подход, однако, наталкивается на ряд трудностей, особенно если число классифицируемых элементов достаточно велико (порядка нескольких сотен).

Прежде всего, не удается предложить функционал, который отражал бы все аспекты нашего интуитивного представления о «хорошей» классификации элементов в самых разных задачах. Тем самым далеко не всегда результаты классификации, получаемые путем экстремизации того или иного

функционала, будут устраивать пользователя. Далее, в такого рода задачах, за весьма редкими и специфическими исключениями, не существует эффективных процедур (т. е. процедур, не сводящихся к полному перебору всех возможных вариантов), которые доставляли бы глобальный экстремум соответствующему функционалу. Поэтому, как правило, можно надеяться только на достижение его локального экстремума. Вместе с тем, с ростом числа элементов растет число локальных экстремумов функционала и тем самым уменьшается вероятность достижения такого локального экстремума функционала, который был бы достаточно близок к его глобальному экстремуму.

В связи с отмеченными недостатками вариационного подхода широкое распространение при решении задачи классификации получили алгоритмы, относительно которых нельзя сказать, какой формальный критерий качества (функционал) они оптимизируют, но которые, по самому своему характеру, отражают те или иные аспекты нашего представления о «хорошей» классификации. Такой подход к решению задачи диагонализации матрицы связи мы будем называть эвристическим.

Далее рассматриваются алгоритмы решения задачи диагонализации матрицы связи, реализующие как вариационный, так и эвристический подходы.

Построение классов ОПД — это лишь один из этапов проектирования профессиональных и образовательных стандартов. В связи с этим к алгоритмам классификации ОПД предъявляется дополнительное требование: на выходе они должны давать не только классификацию (т.е. совокупность классов ОПД), но и матрицу связей между классами $\Phi = \{\varphi_{ij}\}$, где N — число классов, а φ_{ij} — численные характеристики степени связи между классами ОПД, аналогичные характеристикам степени связи r_{ij} между отдельными ОПД.

2. АЛГОРИТМЫ РЕШЕНИЯ ЗАДАЧИ

Описываемые далее алгоритмы состоят из двух этапов:

1) построение численных характеристик степени связи (матрицы связей R);

2) классификация (диагонализация матрицы связей**)**.

Первый этап общий для всех алгоритмов.

2.1. Построение численных характеристик элементов множества *R* (матрицы связей)

Набор численных характеристик степеней связи между разными ОПД (элементы матрицы связей $R = \{r_{ij}\}$, $i, j = 1, ..., N_v$, $i \neq j$) определяется экспертным путем.

Связи между любыми двумя объектами могут быть:

— ассоциативные (близость объектов может быть охарактеризована качественно и (или) количественно);

— агрегатно-ассоциативные (оба объекта входят в состав некоторого агрегата, или объекта «более высокого уровня»);

— композитные (один объект является составной частью второго объекта);

— обобщающие (один объект можно рассматривать как частный случай второго объекта).

Для получения численной оценки связи между ОПД зададим следующие правила экспертного оценивания:

— если эксперт считает, что связь между объектами v_i и v_j композитная, полагаем $r_{ij} = 1$;

— если эксперт считает, что связь между объектами v_i и v_j агрегатно-ассоциативная или обобщающая, полагаем $r_{ij} = 0{,}75$;

— если эксперт считает, что связь между объектами v_i и v_j ассоциативная, он должен оценить ее числом, лежащим в диапазоне $0 < r_{ij} < 0{,}75$;

— если связь между объектами v_i и v_j отсутствует, полагаем $r_{ij} = 0$.

После экспертного оценивания по указанным правилам множество связей между объектами профессиональной деятельности $v_1, ..., v_N$ становится числовым множеством, которое задается матрицей связей $R = \{r_{ij}\}$, $i, j = 1, ..., N_v$, $i \neq j$. Именно матрица связей и есть тот эмпирический материал, который обрабатывается с помощью описываемых далее алгоритмов.

2.2. Вариационный подход к решению задачи классификации

Пусть имеются N_v элементов $v_1, ..., v_{N_v}$ и соответствующая им матрица связей $R = \{r_{ij}\}$, $i, j = 1, ..., N_v$, $i \neq j$, все элементы которой неотрицательные числа.

Пользуясь матрицей R, требуется разбить элементы на N' подмножеств (классов) $q'_1, ..., q'_{N'}$, где число N' считается заданным заранее.

В результате ряда теоретических и экспериментальных исследований установлено [2, 3], что в достаточно широком классе задач удовлетворительные результаты классификации достигаются при максимизации функционала

$$F = \sum_{k=1}^{N'} \frac{m_k}{N_v} \left[\frac{1}{m_k(m_k-1)} \sum_{i,j \in q'_k,\ i \neq j} r_{ij} \right] = \frac{1}{N_v} \sum_{k=1}^{N'} \frac{1}{m_k - 1} \sum_{i,j \in q'_k,\ i \neq j} r_{ij}, \qquad (1)$$

где m_k — число элементов в соответствующем классе. Условие $i \neq j$ введено в функционал (1) для того, чтобы степень связи элемента с самим собой (если таковая имеется) не влияла на результат классификации. Для определенности при $m_k = 1$ будем полагать $\frac{1}{m_k - 1} \sum_{i,j \in q'_k,\, i \neq j} r_{ij} = 0$. Сумма $\sum_{i,j \in q'_k,\, i \neq j} r_{ij}$ есть сумма всех степеней связи между разными элементами, попавшими в один класс q'_k. Величина $m_k (m_k - 1)$ — общее число таких величин, а число m_k / N_v — доля элементов, попавших в класс q'_k, т. е. число, характеризующее «размер» этого класса. Поэтому функционал F имеет смысл суммы взвешенных средних степеней связи внутри каждого класса, причем коэффициенты взвешивания пропорциональны размерам класса. В связи с этим максимизация функционала F приводит к более «плотным» (с большей средней степенью связи между элементами) классам большого размера за счет меньшей плотности классов малого размера.

Вопрос о выборе функционала при реализации вариационного подхода совсем не тривиален. Оказывается, что многие функционалы, по смыслу близкие к введенному выше функционалу F, приводят к результатам, явно противоречащим нашим представлениям о «хорошей» классификации. В качестве такого рода примеров рассмотрим функционалы $F_1 = \sum_{k=1}^{N} \sum_{i,j \in q'_k,\, i \neq j} r_{ij}$, $F_2 = \sum_{k=1}^{N} \frac{1}{m_k (m_k - 1)} \sum_{i,j \in q'_k,\, i \neq j} r_{ij}$. Функционал F_1, на первый взгляд, кажется достаточно разумным, так как его максимизация означает такое разделение элементов на классы, при котором сумма всех связей внутри классов будет максимальной. Вместе с тем, величина F_1 при прочих равных условиях будет тем большей, чем больше элементов матрицы A попадет в блоки. Поэтому, если бы все элементы матрицы A были равны между собой, то при $N = 2$ максимум функционала F_1 достигался бы при таком разбиении элементов на два (непустых) класса, при котором в один класс попадет $N_v - 1$ элементов, а в другой — только один элемент (рис. 2).

Функционал F_2 имеет смысл суммы средних степеней связей внутри классов, т. е. смысл, весьма близкий к смыслу функционала F. Вместе с тем, специальные исследования показали [2], что максимизация функционала F_2 в сложных случаях приводит к плохим результатам, хотя до сих пор нет убедительного содержательного объяснения этого факта.

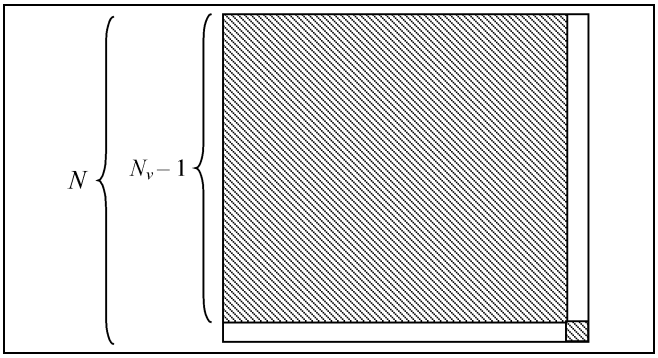

Рис. 2. Неравномерность классификации с функционалом F_1

При достаточно большом числе классов максимуму F_2 часто соответствует разбиение, в котором все классы, кроме одного, состоят из пары наиболее близких элементов. Одна из причин этого заключается в том, что любой k-й член $\frac{1}{m_k (m_k - 1)} \sum_{i,j \in q'_k,\, i \neq j} r_{ij}$ функционала F_2 не зависит от числа классов, но быстро уменьшается с ростом m_k.

Алгоритм поиска локального экстремума функционала F строится следующим образом. Пусть имеется некоторое начальное разбиение множества объектов на классы. На каждом шаге алгоритма осуществляется пробный перенос некоторого очередного элемента из того класса, в котором он находится к данному шагу, последовательно во все остальные классы, начиная с первого. При каждом таком переносе подсчитывается новое значение функционала F и сравнивается со значением этого функционала до переноса. Если при очередном пробном переносе данного элемента значение функционала возросло, то рассматриваемый элемент остается в новом классе (т. е. фактически переносится из того класса, в котором находился, в новый класс). На этом выполнение данного шага алгоритма заканчивается. Если же после пробных переносов во все другие классы значение функционала F ни разу не возросло, то рассматриваемый элемент остается в том же классе, в котором он находился до данного шага. Затем алгоритм переходит к следующему шагу, на котором осуществляются пробные переносы следующего элемента. Алгоритм останавливается после того, как просмотр всех элементов не приводит к изменению ни одного из классов. Таким образом, если считать, что «окрестностью» некоторого разбиения является совокупность всех разбиений, отличающихся от данного принадлежностью только одного элемента, то рассмотренный алгоритм доставляет функционалу экстремум, локальный по отношению к такому определению окрестности.

В качестве начального разбиения в этом алгоритме может использоваться любое разбиение элементов на заданное число L классов.

2.3. Эвристический подход к решению задачи классификации

Эффективность эвристических алгоритмов решения задачи классификации зависит от ее сложности. Если элементы действительно группируются в «плотные» классы, а связь между любыми элементами из разных классов существенно меньше, чем между элементами из одного класса, то такие алгоритмы дают хорошее решение задачи. Однако встречающаяся в реальных задачах ситуация редко бывает столь идеальной, так что лишь «в среднем», «как правило» элементы из одного класса сильнее связаны, чем элементы из разных классов. И чем сложнее задача, т. е. чем сильнее она отличается от «идеальной», тем труднее выделить классы и тем сложнее для этого должен быть алгоритм. Рассмотрим один из наиболее распространенных эвристических алгоритмов — иерархический алгоритм классификации «Объединение» [2].

Пусть два подмножества q_p и q_s элементов включают в себя, соответственно, m_p и m_s элементов. Будем измерять степень связи, или степень «близости» между этими двумя подмножествами величиной

$$K(q_p, q_s) = \frac{1}{m_p m_s} \sum_{i \in q_p} \sum_{j \in q_s} r_{ij}. \qquad (2)$$

Каждый шаг алгоритма заключается в объединении в один класс двух наиболее «близких» (в смысле (2)) друг к другу классов, полученных в результате предыдущих шагов алгоритма, так что на каждом шаге число построенных алгоритмом классов уменьшается на единицу. Работа алгоритма продолжается до тех пор, пока не будет получено заранее заданное число N классов.

В процессе выполнения каждого шага в связи с изменением классов следует также пересчитывать величины $K(q_r, q_t)$. Пусть, например, на некотором шаге объединяются классы q_p и q_s в один класс, который обозначим через q_u. Если q_r и q_t не есть q_u, то соответствующая величина $K(q_r, q_t)$ на данном шаге не изменяется. Если же один из классов, например q_r, это и есть новый класс, то, как легко видеть,

$$K(q_u, q_t) = \frac{1}{(m_p + m_s)m_t} \sum_{i \in q_p \cup q_s} \sum_{j \in q_t} r_{ij} = \frac{m_p K(q_p, q_t) + m_s K(q_s, q_t)}{m_p + m_s}.$$

В качестве начального разбиения для работы алгоритма «Объединение» можно взять N_v классов, содержащих каждый по одному элементу. В свою очередь классы, получаемые в результате работы алгоритма «Объединение», можно использовать в качестве начального разбиения для работы описанного ранее вариационного алгоритма.

3. КЛАССИФИКАЦИЯ ОБЪЕКТОВ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ

В настоящей работе описанные алгоритмы применялись следующим образом. Классификация ОПД проводилась при разных значениях N': множество V разбивалось на 8, 9, 10, 11 и 12 классов с помощью алгоритма «Объединение». При этом каждое из полученных разбиений задавалось в качестве начального разбиения для вариационного алгоритма, который и строил окончательное разбиение на данное число классов. Построенные классификации предъявлялись экспертам пользователя, которые выбирали наилучшую из пяти классификаций, т. е. окончательное число классов N, и уточняли составы классов. Элементы матрицы связей φ_{ij} между классами $q_1, \ldots, q_n$ выбранной пользователем классификации определялись по формуле (2): $\varphi_{ij} = \frac{1}{m_i m_j} \sum_{s \in q_i} \sum_{p \in q_j} r_{sp}$.

ЗАКЛЮЧЕНИЕ

Предложенные методы выделения классов объектов профессиональной деятельности специалистов позволяют перейти к автоматизированным процедурам проектирования профессиональных и образовательных стандартов. Получение границ профессиональной деятельности специалиста повышает объективность результатов разработки профессиональных и образовательных стандартов.

ЛИТЕРАТУРА

1. *Никитин В.В.* Информационно-методические обеспечение формирования перечня направлений и специальностей в области информационно-коммуникационных технологий. — М.: МАКС Пресс, 2006. — 272 с.
2. *Браверман Э.М., Мучник И.Б.* Структурные методы обработки эмпирических данных. — М.: Наука, 1983. — 302 с.
3. *Бауман Е.В., Дорофеюк А.А.* Классификационный анализ данных // Труды Междунар. конф. по проблемам управления. — М., 1999. — Т. 1. — С. 62—77.

☎ *(495) 771-32-38, (495) 334-90-70,*
e-mail: vnikitin@hse.ru, adorof@ipu.ru

Статья представлена к публикации членом редколлегии Ф.Т. Алескеровым. ❑