Структуризация объектов нечисловой природы[1]

А.А. Дорофеюк, И.В. Покровская, А.Л. Чернявский

Аннотация. Для классификации объектов нечисловой природы предлагается использовать полученную экспертным путем матрицу сходства. Классификация производится путем преобразования матрицы сходства к диагональному виду. Предлагаются вариационный и эвристический алгоритмы такого преобразования.

Ключевые слова: объекты нечисловой природы, матрица сходства, автоматическая классификация.

Введение

В настоящее время общепризнанным способом формализации разных областей знаний служит построение онтологий соответствующих предметных областей. Обычно онтологии состоят из объектов, понятий, атрибутов и отношений. Совокупность объектов, относящихся к некоторой предметной области, представляет собой нижний уровень онтологии. Первый этап построения онтологии – это этап формирования понятий, т. е. выявления структуры совокупности объектов предметной области. Понятиями называются классы объектов, формируемые по тому или иному принципу.

Для классификации объектов, которые можно описать набором числовых параметров, разработано множество подходов и алгоритмов (например, [1]). Существуют, однако, предметные области, объекты которых имеют нечисловую природу. Например, при разработке образовательных стандартов классифицировать необходимо области профессионального знания, суть которых нельзя описать при помощи количественных характеристик [2]. В этом случае классификация может производиться по принципу объединения в один класс в каком-то смысле «сходных», «близких» объектов, а степень сходства может быть определена экспертным путем.

Пусть r_{ij} - полученная тем или иным способом мера (степень) сходства объектов i и j, а $R = \{ r_{ij} \}, i, j = 1, \ldots N, i \neq j$ – матрица сходства между объектами. Тогда первый этап формирования понятий – классификацию исходной совокупности объектов предметной области – можно формально представить как процедуру преобразования матрицы R, а именно нахождения такого расположения (нумерации) строк и столбцов матрицы R и выделения таких непересекающихся подматриц вдоль главной диагонали преобразованной таким образом матрицы сходства, чтобы элементы – числа r_{ij} – каждой выделенной подматрицы были возможно

[1]Работа выполнена при частичной поддержке РФФИ, гранты 17-07-00857-а, 15-07-06713-а, 16-07-00896-а, 16-07-00895-а, 16-29-12880-офи, 16-29-12895-офи, 16-29-12943-офи.

больше, а числа r_{ij}, расположенные вне этих подматриц, – возможно меньше. Будем также требовать, чтобы выделенные квадратные подматрицы полностью покрывали главную диагональ матрицы сходства. Условно выделенные подматрицы с большими компонентами можно изобразить в виде квадратов разного размера, расположенных вдоль главной диагонали матрицы *R* (Рис. 1).

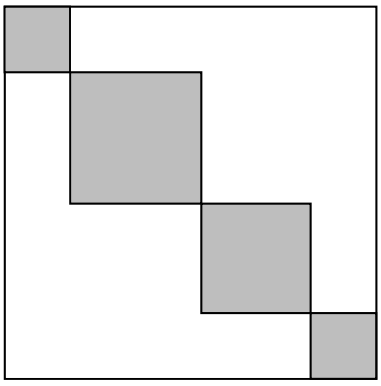

Рис. 1. Преобразованная матрица сходства

Подмножество элементов, которым соответствуют строки (или столбцы) преобразованной матрицы сходства *R,* образующие одну из выделенных подматриц, – это и есть один из классов, о которых шла речь.

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 классы, 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 задачах такого рода, 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 преобразования матрицы связи мы будем называть эвристическим.

Далее рассматриваются алгоритмы решения описанной выше задачи классификации, реализующие как вариационный, так и эвристический подходы.

Построение классов объектов – это лишь один из этапов проектирования онтологии. Во многих случаях к алгоритмам классификации предъявляется дополнительное требование: на выходе они должны давать не только классификацию (т. е. совокупность классов объектов), но и матрицу сходства между классами $\Phi = (\varphi_{ij})$, где φ_{ij} – численные характеристики степени сходства между классами объектов, аналогичные характеристикам степени связи r_{ij} между отдельными объектами.

1. Алгоритмы решения задачи

Описываемые далее алгоритмы состоят из двух этапов:

1) построение численных характеристик степени сходства (матрицы сходства *R)*;

2) классификация (преобразование матрицы сходства).

1.1. Первый этап (общий для всех алгоритмов). Построение численных характеристик сходства между объектами

Набор численных характеристик степеней сходства между разными объектами предметной деятельности (элементы матрицы сходства $R = \{ r_{ij} \}, i, j = 1, \ldots N, i \neq j$) определяется экспертным путем.

Сходство между любыми двумя объектами может быть:

– ассоциативным (сходство объектов может быть охарактеризовано качественно и (или) количественно);

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

– композитным (один объект является составной частью второго объекта);

– обобщающим (один объект можно рассматривать как частный случай второго объекта).

Для получения численной оценки сходства между объектами зададим следующие правила экспертного оценивания:

– если эксперт считает, что сходство между объектами ν_i и ν_j композитное, полагаем $r_{ij} = 1$;

– если эксперт считает, что сходство между объктами ν_i и ν_j агрегатно-ассоциативное или обобщающее, полагаем $r_{ij} = 0{,}75$;

– если эксперт считает, что сходство между объектами ν_i и ν_j ассоциативное, он должен оценить его числом, лежащим в диапазоне $0 < r_{ij} < 0{,}75$;

– если эксперт считает, что сходство между объектами ν_i и ν_j отсутствует, полагаем $r_{ij} = 0$.

После экспертного оценивания по указанным правилам множество оценок сходства между объектами предметной деятельности $\nu_{1, \ldots,} \nu_N$ становится числовым множеством, которое задается матрицей сходства $R = \{ r_{ij} \}, \ i, j = 1, \ldots N, i \neq j$. Именно матрица сходства и есть тот эмпирический материал, который обрабатывается с помощью описываемых далее алгоритмов.

1.2. Вариационный подход к решению задачи классификации

Пусть имеются N объектов $\nu_{1, \ldots,} \nu_N$ и соответствующая им матрица сходства $R = \{ r_{ij} \}, \ i, j = 1, \ldots N, i \neq j$, все элементы которой неотрицательные числа.

Пользуясь матрицей *R,* требуется разбить объекты на *N'* подмножеств (классов) $q_1, \ldots, q_{N'}$, где число *N'* считается заданным заранее.

Зададим критерий качества классификации в виде:

$$F = \sum_{k=1}^{N'} \frac{m_k}{N} \left[\frac{1}{m_k (m_k - 1)} \sum_{i, j \subset q_k, i \neq j} r_{ij} \right] = \\ = \frac{1}{N} \sum_{k=1}^{N'} \frac{1}{m_k - 1} \sum_{i, j \subset q_k, i \neq j} r_{ij}, \quad (1)$$

где m_k – число объектов в соответствующем классе. Условие $i \neq j$ введено в функционал (1) для того, чтобы степень сходства объекта с самим собой (если таковая имеется) не влияла на результат классификации. Для определенности, при $m_k = 1$, будем полагать $\frac{1}{m_k - 1} \sum_{i, j \subset q_k, i \neq j} r_{ij} = 0$.

Сумма $\sum_{i, j \subset q_k, i \neq j} r_{ij}$ - это сумма всех степеней сходства между разными объектами, попавшими в один класс q_k. Величина $m_k (m_k - 1)$ – общее число таких величин, а число m_k / N – доля объектов, попавших в класс q_k, т. е. число, характеризующее «размер» этого класса. Поэтому функционал *F* имеет смысл суммы взвешенных средних степеней сходства внутри каждого класса, причем коэффициенты взвешивания пропорциональны размерам класса. В связи с этим максимизация функционала *F* приводит к более «плотным» (с большей средней степенью сходства между объектами) 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:

$$F_1 = \sum_{k=1}^{N'} \frac{m_k}{N} \sum_{i,j \subset q_k,\, i \neq j} r_{ij}\,, \qquad (2)$$

$$F_2 = \sum_{k=1}^{N'} \frac{1}{m_k(m_k-1)} \sum_{i,j \subset q_k,\, i \neq j} r_{ij}\,. \qquad (3)$$

Функционал (2) на первый взгляд кажется достаточно разумным, так как его максимизация означает такое разделение объектов на классы, при котором сумма всех степеней сходства внутри классов будет максимальной. Вместе с тем, величина (2) при прочих равных условиях будет тем большей, чем больше элементов матрицы R попадет в блоки. Поэтому, если бы все элементы матрицы R были равны между собой, то при $N = 2$ максимум функционала (2) достигался бы при таком разбиении элементов на два (непустых) класса, при котором в один класс попадет $N - 1$ объект, а в другой — только один объект.

Функционал (3) имеет смысл суммы средних степеней сходства внутри классов, т.е. смысл, весьма близкий к смыслу функционала F. Вместе с тем, экспериментальные расчеты показывают, что максимизация функционала (3) в сложных случаях приводит к плохим результатам, хотя до сих пор нет убедительного объяснения этого факта. При достаточно большом числе классов максимуму (3) 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 $\frac{1}{m_k(m_k-1)} \sum_{i,j \subset q_k,\, i \neq j} r_{ij}$ функционала (3) 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 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.

1.3. Эвристический подход к решению задачи классификации

Эффективность эвристических алгоритмов решения задачи классификации зависит от ее сложности. Если элементы действительно группируются в «плотные» классы, а сходство между любыми объектами из разных классов существенно меньше, чем между объектами 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» степень сходства у объектов из одного класса больше, чем у объектов из разных классов. И чем сложнее задача, т. е. чем сильнее она отличается от «идеальной», тем труд-

нее выделить классы и тем сложнее для этого должен быть алгоритм. Рассмотрим один из наиболее распространенных эвристических алгоритмов — иерархический алгоритм классификации [3].

Пусть два подмножества q_p и q_s объектов включают в себя, соответственно, m_p и m_s объектов. Будем измерять степень сходства между этими двумя подмножествами величиной

$$K(q_p,q_s)=\frac{1}{m_p m_s}\sum_{i\subset q_p}\sum_{j\subset q_s} r_{ij}. \qquad (4)$$

Каждый шаг алгоритма заключается в объединении в один класс двух наиболее «схожих» (в смысле (4)) 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 (4). Пусть, например, на некотором шаге классы q_p и q_s объединяются 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. Новые величины рассчитываются по формуле (5):

$$K(q_u,q_t)=\frac{1}{(m_p+m_s)m_t}\sum_{i\subset q_p\cup q_s}\sum_{j\subset q_t} r_{ij}= \\ =\frac{m_p K(q_p,q_t)+m_s K(q_s,q_t)}{m_p+m_s}. \qquad (5)$$

В качестве начального для работы этого алгоритма можно взять разбиение на N' классов, каждый из которых содержит по одному элементу. В свою очередь, итоговую классификацию, получаемую в результате работы иерархического алгоритма, можно использовать в качестве начального разбиения для работы описанного выше вариационного алгоритма.

Заключение

Задачу классификации объектов нечисловой природы можно свести к задаче преобразования полученной экспертным путем матрицы сходства к диагональному виду. Описаны два алгоритма решения этой задачи. В первом алгоритме реализуется вариационный подход, когда вводится в рассмотрение функционал качества классификации, зависящий от разбиения элементов матрица сходства R на классы, причем «хороший» в интуитивном понимании способ классификации соответствует экстремальному значению функционала. Реализация этого алгоритма наталкивается на ряд трудностей, особенно если число классифицируемых элементов достаточно велико (порядка нескольких сотен). Второй алгоритм не имеет строгой формализации постановки, в нём реализована эвристическая процедура классификации элементов матрица R.

Литература

1. Ю.А. Дорофеюк, А.А. Дорофеюк, И.В. Покровская, А.Г .Спиро Методы интеллектуального анализа данных при исследовании сложных систем управления / Труды ИСА РАН, Том 66, вып. 4, 2016. – С. 36-46.
2. Никитин В.В. Информационно-методические обеспечение формирования перечня направлений и специальностей в области информационно-коммуникационных технологий. М.: МАКС Пресс, 2006. – 272 с.
3. Дорофеюк А.А. Методология экспертно-классификационного анализа в задачах управления и обработки сложноорганизованных данных (история и перспективы развития) / Проблемы управления 2009. № 3.1. – С. 19-28.

Дорофеюк Александр Александрович. Главный научный сотрудник ИСА ФИЦ ИУ РАН, главный научный сотрудник ИПУ РАН. Окончил МФТИ в 1965 году. Доктор технических наук, профессор. Количество печатных работ: 239, в том числе 15 монографий. Область научных интересов: математическая статистика, функциональный анализ, интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений, системный анализ. E-mail: daa2@mail.ru.

Покровская Ирина Вячеславовна. Старший научный сотрудник ИПУ РАН. Окончила МГУ в 1976 году. Кандидат технических наук. Количество печатных работ: 64. Область научных интересов: интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений. E-mail: ivp750@mail.ru.

Чернявский Александр Леонидович. Старший научный сотрудник ИПУ РАН. Окончил МФТИ в 1964 году. Кандидат технических наук. Количество печатных работ: 47. Область научных интересов: интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений. E-mail: achern@ipu.ru

Data structuring for nonnumeric objects

A.A. Dorofeyuk, I.V. Pokrovskaya, A.L. Chernyavsky

For clustering of the nonnumeric objects the similarity matrix of the expert judgments is used. The clustering is performed by diagonalization of similarity matrix. The variational and the heuristic diagonalization algorithms are proposed.

Keywords: nonnumeric objects, similarity matrix, cluster-analysis.

References

1. J.A. Dorofeyuk, A.A. Dorofeyuk, I.V. Pokrovskaya, A.G. Spiro. The methods of intellectual data analysis for investigation of complicate management systems / Proc. ISA RAS, vol. 66, 4, 2016, pp. 36-46 (in Russian).
2. V.V. Nikitin. Information and methodological support of the directions and specialties list formation in the information and communication technologies field. / -M.: MACS Press, 2006. – 272 pp. (in Russian).
3. A.A. Dorofeyuk. The expert-classification methods for the complex data analysis and management (history and prospect) / Problemy upravleniya 2009. №3.1 – pp. 19-28. (in Russian).

Alexander A. Dorofeyuk. Institute for System Analysis of the Federal Research Center «Information and Control», RAS, Moscow, Chief Researcher; V.A.Trapeznikov's Institute of Control Sciences of RAS, Moscow, Chief Researcher. Doctor of Sciences (Computer Sciences), Professor.

Irina V. Pokrovskaya V.A.Trapeznikov's Institute of Control Sciences of RAS, Moscow, Senior Researcher. PhD (Computer Sciences), Associate Professor.

Alexander L. Chernyavsky. V.A.Trapeznikov's Institute of Control Sciences of RAS, Moscow, Senior Researcher. PhD (Computer Sciences), Associate Professor.