УДК 510. 67–519.24

А.А.Викентьев

Новосибирский государственный университет; Институт математики им. С.Л.Соболева СО РАН, Новосибирск, Россия (E-mail: vikent@math.nsc.ru)

Кластеризация экспертных высказываний с использованием моделей теорий

В статье представлены высказывания экспертов в виде логических формул. Ранее с использованием теоретико-модельных понятий были введены расстояния между формулами и меры опровержимости (недостоверности, информативности) высказываний и доказаны их свойства, учитывающие семантику сходства и различия информации, содержащейся в высказываниях. Данные величины использованы для кластеризации баз знаний. Рассмотрены примеры кластеризации конечной группы высказываний различными методами.

Ключевые слова: двузначная логика, экспертные высказывания, теории, расстояние между формулами, мера опровержимости, кластеризация, иерархический алгоритм, базы знаний, теория моделей.

Светлой памяти моего Учителя и Математика
(к 70-летиюТуленды Гарифовича Мустафина)

В базах знаний систем искусственного интеллекта накапливается большое количество знаний, которые извлекаются из данных автоматически или получаются инженерами по знаниям от экспертов. Попытки разобраться в содержании этого хранилища информации приводят к необходимости навести в нем некоторый порядок. В первую очередь возникает необходимость разобраться в структуре этого информационного массива, выделить в нем некоторые подмножества в чем-то похожих друг на друга знаний, найти типичных представителей каждого такого подмножества. Для введения расстояния между высказываниями экспертов, записанных в виде логических формул исчисления высказываний, использовался классический теоретико-модельный подход, предложенный Г.Кейслером и Ч.Чэном [1, 2]. При анализе знаний, заданных в виде высказываний экспертов, для различия содержащейся в них информации и группирования их по схожести, возникает необходимость введения расстояния между высказываниями экспертов и меры опровержимости (информативности) высказываний экспертов. Этой проблемой занимались Н.Г.Загоруйко, Г.С.Лбов, В.Б.Бериков, их коллеги и ученики [3–7]. В работе приводятся конкретные алгоритмы кластеризации различных множеств высказываний различными методами кластеризации. Результаты можно распространить с помощью специальных модельных расстояний на множества многозначных и вероятностных высказываний.

Кластерный анализ (Data clustering) — задача разбиения заданной выборки объектов на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Перейдем к формальной постановке задачи кластеризации.

Пусть X — множество объектов, Y — множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $\rho(x,x')$. Имеется конечная обучающая выборка объектов $X_m = \{x_1,...,x_m\}$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X_m$ приписывается номер кластера y_i. В нашем случае объекты — формулы.

Алгоритм кластеризации — это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного критерия качества кластеризации.

Можно выделить следующие основные этапы кластерного анализа [5]:

1. Формирование системы переменных.
2. Определение способа вычисления расстояния между объектами или группами объектов.
3. Группировка объектов.
4. Представление результатов.
5. Определение качества полученной группировки.

Применим различные алгоритмы кластеризации к конечной группе высказываний. Рассмотрим множество формул с $S(\sum) = \{x, y, z, w\}$:

$$\varphi_1 = x \to y;\ \varphi_2 = \neg(x \to y);\ \varphi_3 = (x \vee z) \to y;\ \varphi_4 = \neg((x \wedge y) \vee z) \to w;\ \varphi_5 = y \to (x \wedge z);$$
$$\varphi_6 = (\neg y \vee (x \to z)) \to w;\ \varphi_7 = ((x \to y) \to z) \to w;\ \varphi_8 = (w \to z) \wedge (y \to x);\ \varphi_9 = x \wedge y;\ \varphi_{10} = x \vee y.$$

Построим для них матрицу расстояний [4]:

Таблица 1

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8	φ_9	φ_{10}
φ_1	0	1	0,125	0,3125	0,625	0,4375	0,3125	0,375	0,5	0,5
φ_2	1	0	0,875	0,6875	0,375	0,5625	0,6875	0,625	0,5	0,5
φ_3	0,125	0,875	0	0,4375	0,75	0,4375	0,3125	0,5	0,375	0,375
φ_4	0,3125	0,6875	0,4375	0	0,4375	0,25	0,375	0,375	0,5625	0,3125
φ_5	0,625	0,375	0,75	0,4375	0	0,5625	0,5625	0,25	0,625	0,625
φ_6	0,4375	0,5625	0,4375	0,25	0,5625	0	0,125	0,5625	0,4375	0,4375
φ_7	0,3125	0,6875	0,3125	0,375	0,5625	0,125	0	0,5625	0,5625	0,4375
φ_8	0,375	0,625	0,5	0,375	0,25	0,5625	0,5625	0	0,625	0,625
φ_9	0,5	0,5	0,375	0,5625	0,625	0,4375	0,5625	0,625	0	0,5
φ_{10}	0,5	0,5	0,375	0,3125	0,625	0,4375	0,4375	0,625	0,5	0

Меры информативности для указанных формул равны:

Таблица 2

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8	φ_9	φ_{10}
μ	0,25	0,75	0,375	0,1875	0,375	0,4375	0,3125	0,375	0,75	0,25

Иерархический подход

Описание алгоритма. Применим иерархический алгоритм кластеризации к группе n высказываний. Предположим, что у нас есть n кластеров. Построим матрицу расстояний для группы из n высказываний, потом выделим наименьшее расстояние между формулами φ_i и φ_j и объединим формулы φ_i и φ_j в один кластер. Затем пересчитаем матрицу расстояний для уже $n-1$ высказывания и будем повторять действия до тех пор, пока все высказывания не объединятся в один кластер. Кластеры будем объединять по методу ближайшего соседа, то есть, $\rho(\varphi_k, \varphi_{ij}) = \min\{\rho(\varphi_k, \varphi_i), \rho(\varphi_k, \varphi_j)\}$.

1.1.1 Случай с неизвестным числом кластеров

Объединение кластеров:

Шаг 1: $\min\limits_{i \neq j} \rho(\varphi_i, \varphi_j) = 0{,}125 = \rho(\varphi_1, \varphi_3) = \rho(\varphi_6, \varphi_7)$. Кластеры: $\varphi_{13}, \varphi_2, \varphi_4, \varphi_5, \varphi_{67}, \varphi_8, \varphi_9, \varphi_{10}$.

Шаг 2: $\min\limits_{i \neq j} \rho(\varphi_i, \varphi_j) = 0{,}25 = \rho(\varphi_4, \varphi_{67}) = \rho(\varphi_5, \varphi_8)$. Кластеры: $\varphi_{13}, \varphi_2, \varphi_{467}, \varphi_{58}, \varphi_9, \varphi_{10}$.

Шаг 3: $\min\limits_{i \neq j} \rho(\varphi_i, \varphi_j) = 0{,}3125 = \rho(\varphi_{467}, \varphi_{13}) = \rho(\varphi_{467}, \varphi_0)$. Кластеры: $\varphi_{134670}, \varphi_2, \varphi_{58}, \varphi_9$.

Шаг 4: $\min\limits_{i\neq j}\rho(\varphi_i,\varphi_j)=0,375=\rho(\varphi_2,\varphi_{58})$. Кластеры: $\varphi_{134670},\varphi_{258},\varphi_9$.

Шаг 5: $\min\limits_{i\neq j}\rho(\varphi_i,\varphi_j)=0,5=\rho(\varphi_{134670},\varphi_9)=\rho(\varphi_{258},\varphi_9)$. Кластер: $\varphi_{1234567890}$.

В случае, если оптимальное число кластеров заранее неизвестно, в качестве критерия остановки алгоритма можно взять меру информативности высказываний. Например, если перед началом кластеризации задать максимальную допустимую разницу между мерами информативности элементов одного кластера, то алгоритм будет продолжаться до достижения этого значения.

На шаге 1 максимальная разница между мерами информативности одного кластера: $\max|\mu(\varphi_i)-\mu(\varphi_j)|=0,125$. На шаге 2: 0,25. Шаг 3: 0,25. Шаг 4: 0,375. Шаг 5: 0,5.

Таким образом, если задать максимальное допустимое значение $|\mu(\varphi_i)-\mu(\varphi_j)|=0,375$, алгоритм кластеризации остановится, и результатами будут кластеры $\varphi_{134670},\varphi_{258},\varphi_9$.

Если мы зададим более строгий критерий, например, $|\mu(\varphi_i)-\mu(\varphi_j)|=0,25$, алгоритм остановится уже после 2 шага, разбив исходное множество на 6 кластеров: $\varphi_{13},\varphi_2,\varphi_{467},\varphi_{58},\varphi_9,\varphi_{10}$.

1.1.2. Случай с заранее заданным числом кластеров

Можно заметить, что в данном наборе экспертных высказываний содержатся два противоположных высказывания — импликация $\varphi_1=x\to y$ и ее отрицание $\varphi_2=\neg(x\to y)$.

Логично предположить, что противоположные формулы будут принадлежать разным кластерам. Найдем разбиение указанного множества на два кластера: $K_1\,|\,\varphi_1\in K_1$ и $K_2\,|\,\varphi_2\in K_2$. Искать будем так: для каждого элемента найдем ближайшего соседа из тех высказываний, которые уже приписаны к какому-нибудь кластеру, и припишем его к тому же кластеру.

Изначально известно, что $\varphi_1\in K_1$ и $\varphi_2\in K_2$.

Шаг 1: $\min\limits_{i<3}\rho(\varphi_i,\varphi_3)=0,125=\rho(\varphi_1,\varphi_3)\Rightarrow\varphi_3\in K_1$.

Шаг 2: $\min\limits_{i<4}\rho(\varphi_i,\varphi_4)=0,3125=\rho(\varphi_1,\varphi_4)\Rightarrow\varphi_4\in K_1$.

Шаг 3: $\min\limits_{i<5}\rho(\varphi_i,\varphi_5)=0,375=\rho(\varphi_2,\varphi_5)\Rightarrow\varphi_5\in K_2$.

Шаг 4: $\min\limits_{i<6}\rho(\varphi_i,\varphi_6)=0,25=\rho(\varphi_4,\varphi_6)\Rightarrow\varphi_6\in K_1$.

Шаг 5: $\min\limits_{i<7}\rho(\varphi_i,\varphi_7)=0,125=\rho(\varphi_6,\varphi_7)\Rightarrow\varphi_7\in K_1$.

Шаг 6: $\min\limits_{i<8}\rho(\varphi_i,\varphi_8)=0,25=\rho(\varphi_5,\varphi_8)\Rightarrow\varphi_8\in K_2$.

Шаг 7: $\min\limits_{i<9}\rho(\varphi_i,\varphi_9)=0,375=\rho(\varphi_3,\varphi_9)\Rightarrow\varphi_9\in K_1$.

Шаг 8: $\min\limits_{i<10}\rho(\varphi_i,\varphi_{10})=0,3125=\rho(\varphi_4,\varphi_{10})\Rightarrow\varphi_{10}\in K_1$.

Таким образом, мы получим классы $K_1=\{\varphi_1,\varphi_3,\varphi_4,\varphi_6,\varphi_7,\varphi_9,\varphi_{10}\}$ и $K_2=\{\varphi_2,\varphi_5,\varphi_8\}$.

Полученные результаты применимы для расстояний с учетом степени разнесенности моделей, а также распространяются на многозначные и вероятностные высказывания с помощью специальных теоретико-модельных расстояний между соответствующими объектами.

Работа поддержана грантом РФФИ; номера проектов 10–01–00113а, 11–07–00345а.

References

1 *Karpenko A.S.* Lukasiewicz logic and prime numbers. — Moscow: Nauka, 2000. — 319 p.

2 *Ershov Yu.L., Palyutin E.A.* Mathematical logic. — 3nd ed. — Moscow: Fizmatlit, 2011. — 358 p.

3 *Lbov G.S., Startseva N.G.* Logical decision functions and aspects of statistical stability of the solutions. — Novosibirsk: Izd-vo In-ta matematiki, 1999. — 212 p.

4 *Vikentyev A.A., Lbov G.S.* Setting the metric and informativeness on statements of experts // Pattern Recognition and Image Analysis. — 1997. — Vol. 7. — № 2. — P. 175–183.

5 *Vikentyev A.A.* Measure falsifiability statements of experts, distances in multi-valued logic and the process of adaptation // XIV international Conference «Knowledge-Dialogue-Solution» KDS 2008. — Bulgaria: Varna, 2008.

6 *Berikov V.B., Lbov G.S.* Construction of decision functions and sustainability issues. — Novosibirsk: Izd-vo In-ta matematiki, 2006. — 218 p.

7 *Zagoruyko N.G.* Applied methods analysis and knowledge. — Novosibirsk: IM SB RAS, 1999. — 270 p.

А.А.Викентьев

Модельдер теориясын пайдаланып, сарапшылық тұжырымдарды кластерлеу

Мақалада сарапшылардың пікірлері логикалық формулалар түрінде көрсетілген. Ертерек теоретикалық-модельдік ұғымдар қолданып, формулалар және тұжырымдарды терістеу (дұрыс еместік, ақпараттау) шамасы арасындағы арақашықтық енгізілді және тұжырымдардағы ақпараттың ұқсастығы мен айырмашылығы семантикасын ескеретін қасиеттері дәлелденген. Берілген шамалар білім қорларын кластерлеу үшін қолданылған. Тұжырымдардың ақырлы топтарын әр түрлі әдістермен кластерлеудің мысалдары қарастырылған.

A.A.Vikentyev

Clusterization of expert statements using models for theories

In this paper we represent experts' statements by means of logical formulas. Earlier, there were introduced distances between formulas and measures of refutation (of unreliability, informativeness) for propositions that were based on model-theoretic concepts; and their properties were established that take into account the semantics of similarity and difference between the information represented in the propositions. These notions are used for clusterization of data bases. We consider several examples of clusterization for a finite group of propositions.

УДК 519.67–519.24

А.А.Викентьев, Е.С.Кабанова

Новосибирский государственный университет; Институт математики им. С.Л.Соболева СО РАН, Новосибирск, Россия (E-mail: vikent@math.nsc.ru)

Расстояние между формулами пятизначной логики Лукасевича и мера недостоверности высказываний экспертов

В статье высказывания экспертов представлены в виде формул пятизначной логики Лукасевича. Аналогично случаю классической логики, используя теорию моделей, были введены понятия расстояния между формулами и меры недостоверности высказываний. Определены и доказаны свойства введённых понятий, учитывающие семантику сходства и различия информации, содержащейся в высказываниях. Данные величины могут быть использованы для кластеризации многозначных баз знаний. Рассмотрен пример кластеризации группы высказываний иерархическим методом. Мера недостоверности в данном случае выступает в качестве критерия остановки алгоритма.

Ключевые слова: многозначная логика, логика Лукасевича, расстояние между формулами, мера недостоверности, кластеризация, иерархический алгоритм, базы знаний, экспертные высказывания, теория моделей.

Посвящается Учителю по теории моделей — Туленды Гарифовичу Мустафину

Введение

На сегодняшний день актуальной является проблема анализа многозначной экспертной информации, представленной в виде высказываний экспертов, которые можно записать в виде логических формул исчисления высказываний. В данной работе экспертные высказывания представлены в виде формул пятизначной логики Лукасевича [1]. Ясно, что различные высказывания и соответствующие им формулы несут разное количество информации. Поэтому возникает вопрос о сравнении экспертных высказываний по информативности и, как следствие, их ранжировании. Ясно, что информативность всего высказывания должна зависеть от информативности элементарных компонент и степени различия содержащейся в них информации. Следовательно, необходимо ввести «расстояние» между