УДК 51

А.А.Викентьев[1,2], М.С.Авилов[2]

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

Новые полные метрики и меры нетривиальности для формул многозначных логик в автоматической кластеризации формул из логической базы знаний. II

В статье рассмотрены следующие задачи: обобщены ранее введённые расстояния и меры нетривиальности на n-значный случай любой логики, сняв ограничения на параметры; используя новые расстояния, разработаны алгоритмы кластеризации множеств формул в n-значной логике; проведены кластеризации подмножеств формул из различных баз знаний и предложены способы сравнения результатов различных адаптированных алгоритмов кластеризации.

Ключевые слова: расстояния и меры, кластеризация множеств, логика, алгоритмы.

3 Обобщение расстояния и меры нетривиальности формул

Заменим логические коэффициенты в расстоянии и мере нетривиальности формул L_n на произвольные (с ограничениями, необходимыми для сохранения полезных свойств величин).

3.1 Обобщённое расстояние

Определение 3.1.1. Расстоянием между формулами φ и ψ n-значной логики L_n при $S(\varphi) \cup S(\psi) \subseteq\subseteq S(\Sigma)$ на множестве $P(S(\Sigma))$ будем называть величину

$$\rho(\varphi,\psi) = \frac{1}{n^{|S(\Sigma)|}} \cdot \sum_{k=0}^{n-1}\sum_{l=0}^{n-1} \lambda_{kl} \cdot M(\frac{k}{n-1}, \frac{l}{n-1})$$

$$0 = \lambda_{00} \leqslant \lambda_{01} \leqslant ... \leqslant \lambda_{n-1n-1} = 1; \quad n \geqslant 2,\ n \in N.$$

Далее докажем свойства введённого расстояния.

Теорема 3.1.1. Для любых формул φ, ψ, χ из Σ расстояние ρ, введенное в [3], удовлетворяет следующим свойствам:

1) $0 \leqslant \rho(\varphi,\psi) \leqslant 1$;

2) $\rho(\varphi,\psi) = 0 \Leftrightarrow \varphi \equiv \psi$;

3) $\rho(\varphi,\psi) = \rho(\psi,\varphi)$;

4) $\rho(\varphi,\psi) \leqslant \rho(\varphi,\chi) + \rho(\chi,\psi)$;

5) $\varphi \equiv \varphi_1, \psi \equiv \psi_1 \Rightarrow \rho(\varphi,\psi) = \rho(\varphi_1,\psi_1)$.

Доказательство.

1. В формуле для вычисления расстояния участвуют все модели с коэффициентами от 0 до 1. $\rho(\varphi,\psi)=0$, когда $\varphi\equiv\psi$; $\rho(\varphi,\psi)=1$, когда $\varphi\equiv\not\psi$; φ, ψ принимают на моделях только значения 0 и 1. Значит, $0\leqslant\rho(\varphi,\psi)\leqslant 1$.

2. Необходимость следует из доказательства предыдущего свойства. Достаточность следует из того, что, по определению эквивалентности, если $\varphi\equiv\psi$, то их значения на всех моделях совпадают. Тогда при $k=l$ все $M(\frac{k}{n-1},\frac{l}{n-1})$ входят в формулу $\rho(\varphi,\psi)$ с коэффициентом 0, следовательно, $\rho(\varphi,\psi)=0$.

3. Симметричные пары $M(\frac{k}{n-1},\frac{l}{n-1})\neq M(\frac{l}{n-1},\frac{k}{n-1})$ умножаются на один и тот же коэффициент $\Rightarrow\rho(\varphi,\psi)=\rho(\psi,\varphi)$.

4. Следует из работы [1], общих теоретико-модельных свойств и ограничений, указанных ранее.

5. Следует из определения эквивалентности двух формул [2].

Теорема доказана.

Заметим, что 2-4 – свойства метрики $\Rightarrow$, получено метрическое пространство на классах эквивалентных формул L_n.

Кроме того, в случае, когда истинностные значения некоторых переменных заранее известны, формулу для нахождения расстояния можно записать иначе. *Замечание.* Пусть переменные $x_1,...,x_p, x_i\in S(\varphi)\cup S(\psi), i=1,...,p, p=\mid S(\varphi)\cup S(\psi)\mid$ соответственно принимают $m_1,...,m_p, m_i\leqslant\leqslant n$ истинностных значений. Тогда обобщённое расстояние $\rho(\varphi,\psi)$ принимает вид

$$\rho'(\varphi,\psi)=\frac{1}{m_1\cdot...\cdot m_p}\cdot\sum_{k=0}^{n-1}\sum_{l=0}^{n-1}\lambda_{kl}\cdot M(\frac{k}{n-1},\frac{l}{n-1}).$$

В данном случае при вычислении расстояния рассматриваются не все модели, а подмножество из $m_1\cdot...\cdot m_p$ моделей. Для ρ' также справедлива теорема 3.1.1, доказательство проводится по той же схеме.

3.2 Мера нетривиальности

Определение 3.2.1. *Мерой нетривиальности формулы φ n-значной логики L_n при $S(\varphi)\subseteq S(\Sigma)$ на множестве $P(S(\Sigma))$ будем называть величину*

$$I(\varphi)=\rho(\varphi,1)=\sum_{i=0}^{n-2}\alpha_i\cdot\frac{M(\varphi_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}} \tag{2}$$

$$0\leqslant\alpha_i\leqslant 1;$$

$$\alpha_k\geqslant\alpha_i\,\forall k\leqslant i;$$

$$\alpha_i+\alpha_{n-1-i}=1\,\forall i=0,...,\frac{n-1}{2}.$$

Докажем свойства введённой меры нетривиальности.

Теорема 3.2.1 *Для любых формул φ,ψ из Σ мера нетривиальности I удовлетворяет следующим свойствам*:

1) $0\leqslant I(\varphi)\leqslant 1$;
2) $I(\varphi)+I(\neg\varphi)=1$;

3) $I(\varphi \wedge \psi) \geqslant max\{I(\varphi), I(\psi)\}$;
4) $I(\varphi \vee \psi) \leqslant min\{I(\varphi), I(\psi)\}$;
5) $I(\varphi \wedge \psi) + I(\varphi \vee \psi) \geqslant I(\varphi) + I(\psi)$.

Доказательство.

1. $I(\varphi) = \rho(\varphi, 1) \Rightarrow 0 \leqslant I(\varphi) \leqslant 1$.

2. $I(\varphi) + I(\neg\varphi) = \alpha_0 \cdot \dfrac{M(\varphi_0)}{n^{|S(\Sigma)|}} + \alpha_{n-1} \cdot \dfrac{M(\varphi_1)}{n^{|S(\Sigma)|}} + \sum\limits_{i=1}^{n-2} (\alpha_i + \alpha_{n-1-i}) \cdot \dfrac{M(\varphi_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}} =$

$= \dfrac{\mid P(S(\Sigma)) \mid}{n^{|S(\Sigma)|}} = 1.$

3. $I(\varphi \wedge \psi) = \sum\limits_{i=0}^{n-2} \alpha_i \cdot \dfrac{M((\varphi \wedge \psi)_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}} = \sum\limits_{i=0}^{n-2} \alpha_i \left(\sum\limits_{k=i}^{n-1} \left(\dfrac{M(\varphi_{\frac{k}{n-1}} \wedge \psi_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}} + \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} \right) \right) -$

$-\alpha_i \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}}.$

$I(\varphi) = \sum\limits_{i=0}^{n-2} \alpha_i \sum\limits_{k=0}^{n-1} \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} = \sum\limits_{i=0}^{n-2} \alpha_i \left(\sum\limits_{k=i}^{n-1} \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} + \sum\limits_{k=0}^{i} \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} \right) -$

$-\alpha_i \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}}.$

$I(\varphi \wedge \psi) - I(\varphi) = \sum\limits_{i=0}^{n-2} \sum\limits_{k=0}^{i} (\alpha_k - \alpha_i) \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} + \sum\limits_{i=0}^{n-2} \sum\limits_{k=i}^{n-1} \alpha_i \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} \geqslant 0.$

Значит, $I(\varphi \wedge \psi) \geqslant I(\varphi)$. Аналогично получаем неравенство для ψ: $I(\varphi \wedge \psi) \geqslant I(\psi)$. Следовательно, $I(\varphi \wedge \psi) \geqslant max\{I(\varphi), I(\psi)\}$.

4. $I(\varphi \vee \psi) = \sum\limits_{i=0}^{n-2} \alpha_i \cdot \dfrac{M((\varphi \wedge \psi)_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}} = \sum\limits_{i=0}^{n-2} \alpha_i \left(\sum\limits_{k=i}^{n-1} \left(\dfrac{M(\varphi_{\frac{k}{n-1}} \wedge \psi_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}} + \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} \right) \right) -$

$-\alpha_i \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}}.$

$I(\varphi) = \sum\limits_{i=0}^{n-2} \alpha_i \sum\limits_{k=0}^{n-1} \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} = \sum\limits_{i=0}^{n-2} \alpha_i \left(\sum\limits_{k=i}^{n-1} \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} + \sum\limits_{k=0}^{i} \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} \right) -$

$-\alpha_i \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}}.$

$I(\varphi) - I(\varphi \vee \psi) = \sum\limits_{i=0}^{n-2} \sum\limits_{k=0}^{i} \alpha_i \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}} - \sum\limits_{i=0}^{n-2} \sum\limits_{k=0}^{i} \alpha_i \dfrac{M(\varphi_{\frac{k}{n-1}} \wedge \psi_{\frac{i}{n-1}})}{n^{|S(\Sigma)|}} \geqslant \sum\limits_{i=0}^{n-2} \sum\limits_{k=0}^{i} \alpha_i \dfrac{M(\varphi_{\frac{i}{n-1}} \wedge \psi_{\frac{k}{n-1}})}{n^{|S(\Sigma)|}}.$

Значит, $I(\varphi \vee \psi) \leqslant I(\varphi)$. Аналогично получаем неравенство для ψ: $I(\varphi \vee \psi) \leqslant I(\psi)$. Следовательно, $I(\varphi \vee \psi) \leqslant min\{I(\varphi), I(\psi)\}$.

5. Из тождеств, использовавшихся при доказательстве пунктов 3 и 4, следует, что

$$I(\varphi \wedge \psi) + I(\varphi \vee \psi) \geqslant I(\varphi) + I(\psi).$$

Теорема доказана. ⊠

4 Алгоритмы кластеризации множеств формул в L_n

Кластерный анализ данных имеет большое значение в работе с базами знаний, высказываниями экспертов и в статистическом моделировании [3]. В связи с этим поставлена задача класте-

ризации множеств высказываний в n-значной логике на основе обобщённого расстояния и меры нетривиальности, введенных выше.

Для множеств высказываний известны только расстояния между формулами и меры нетривиальности. Поэтому для кластеризации были выбраны два общеизвестных алгоритма, в которых существенную роль играет расстояние – иерархический алгоритм и алгоритм $k = means$. Сложность вычисления расстояния – экспоненциальная. В разработанном программном комплексе эти алгоритмы были адаптированы для работы с конечными множествами формул в L_n [4].

4.1 Иерархический алгоритм

Пусть I – множество объектов. Кластеризация происходит путём объединения мелких кластеров в более крупные (агломерации).

В результате получается совокупность H вложенных подмножеств S (кластеров), удовлетворяющих свойству: при любых $S_1, S_2 \in H$ их пересечение $S_1 \cap S_2$ либо пусто, либо совпадает с одним из них.

Применительно к множествам формул в L_n иерархический алгоритм работает следующим образом:

Рассматриваем конечное множество логических формул L_n.

Используя введённое расстояние, строим матрицу расстояний для этого набора формул. Каждая итерация алгоритма будет состоять из двух шагов:

1. Ищем формулы, между которыми наименьшее расстояние и объединяем их в 1 кластер.

2. Объединяем кластеры по методу ближайшего соседа. Для этого пересчитываем матрицу расстояний по следующему правилу:

$$\rho(\varphi_k, \varphi_{ij}) = min\{\rho(\varphi_k, \varphi_i), \rho(\varphi_k, \varphi_j)\}.$$

Итерации продолжаются, пока максимальная разница между мерами нетривиальности элементов одного кластера не достигнет заранее заданной величины Δ.

4.2 Алгоритм k-средних

Пусть I – множество объектов, K – множество начальных точек (центров). Каждая итерация алгоритма $k = means$ состоит из двух шагов:

1. Обновление кластеров – при заданных K центрах $C_k, k = 1, ..., K$, каждый объект $i \in I$ приписывается к ближайшему из центров C_k. Таким образом, формируются кластеры $S_k, k = 1, ..., K$.

2. Обновление центров – для каждого кластера S_k вычисляется его центр тяжести (внутрикластерное среднее), который назначается новым центром C'_k.

Процесс останавливается, когда кластеры на шаге t совпадут с кластерами на шаге $t - 1$.

Применительно ко множествам формул в L_n алгоритм k-средних работает следующим образом:

Рассматриваем конечное множество логических формул L_n.

Затем выбираем некоторые K формулы из данного множества, которые будут являться центрами, и определяемся с количеством кластеров. При выборе центров будем руководствоваться следующими соображениями: центры должны быть примерно равноудалены друг от друга, и расстояние между кластерами должно быть наибольшим. Каждая итерация алгоритма будет состоять из двух шагов:

1. Приписываем каждую формулу из множества к ближайшему центру.

2. Центр масс кластера – это столбец значений логики L_n. Для определения этого столбца вычисляется среднее арифметическое S_a значений элементов одного кластера на каждой модели.

Если S_a принадлежит множеству логических значений $V_n = \left\{0, \frac{1}{n-1}, ..., \frac{n-2}{n-1}, 1\right\}$, то оно записывается в столбец значений.

Если S_a не принадлежит множеству логических значений V_n, то в столбец значений записывается ближайшее (снизу или сверху) значение из V_n.

Итерации алгоритма продолжаются, пока кластеры не останутся такими же, как на предыдущей итерации.

Пример: покажем детальнее, как определяется центр масс. Пусть во время работы алгоритма k-средних m формул $\varphi_1, ..., \varphi_m$ n-значной логики L_n, в написании которых участвуют l переменных $x_1, ..., x_l$, составили один кластер $C_{1...m}$. Тогда вычисление центра масс C_M кластера $C_{1...m}$ можно проиллюстрировать следующей таблицей:

x_1	...	x_l	φ_1	...	φ_m	C_M
0	...	0	p_{11}	...	p_{1m}	C_1
$\frac{1}{n-1}$		0	p_{21}		p_{2m}	C_2
...						

$$C_1 = \frac{1}{m} \cdot \sum_{i=1}^{m} p_{1i} = p_1 \in V_n = \left\{0, \frac{1}{n-1}, ..., p_1, ..., \frac{n-2}{n-1}, 1\right\}.$$

Все p_{qi} являются истинностными значениями из V_n, но их нормированная сумма может не принадлежать V_n. Допустим, что это произошло в случае C_2:

$$C_2 = \frac{1}{m} \cdot \sum_{i=1}^{m} p_{2i} = p_2 \notin V_n = \left\{0, \frac{1}{n-1}, ..., \frac{n-2}{n-1}, 1\right\}.$$

Пусть ближайшее к p_2 значение V_n равняется p_2^0. Тогда полагаем C_2 равным p_2^0. Остальные центры вычисляются аналогично. В итоге получаем C_M – столбец значений логики L_n, который является центром масс кластера $C_{1...m}$.

5 Численные эксперименты

Посмотрим, как описанные выше алгоритмы будут работать для множеств формул n-значной логики.

Пусть $n = 5$. Рассмотрим следующее множество *test* 1 формул пятизначной логики L_5:

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

Применим к множеству формул *test* 1 иерархический алгоритм кластеризации. Выберем $\lambda_0 == 0$, $\lambda_1 = \frac{1}{4}, \lambda_2 = \frac{2}{4}, \lambda_3 = \frac{3}{4}, \lambda_4 = 1$ (стандартный случай). Сначала строится матрица расстояний.

Наименьшим расстоянием является $\rho_{46} = 0,0510$, первым кластером будет φ_{46}. Спустя несколько итераций результатом работы иерархического алгоритма будет следующая таблица:

	1	2	3	4	5	6	7	8
1	0	0,7587	0,1000	0,3349	0,4555	0,3861	0,2502	0,4250
2		0	0,6674	0,5468	0,5000	0,4996	0,6423	0,5029
3			0	0,3251	0,5098	0,3660	0,2477	0,4503
4				0	0,4131	0,0510	0,1295	0,4442
5					0	0,4197	0,4370	0,1488
6						0	0,1672	0,4637
7							0	0,4726
8								0

Итерация	Δ	Кластеры
1	0,0508	$\varphi_1, \varphi_2, \varphi_3, \varphi_{46}, \varphi_5, \varphi_7, \varphi_8$
2	0,1000	$\varphi_{13}, \varphi_2, \varphi_{46}, \varphi_5, \varphi_7, \varphi_8$
3	0,1376	$\varphi_{13}, \varphi_2, \varphi_{467}, \varphi_5, \varphi_8$
4	0,1376	$\varphi_{13}, \varphi_2, \varphi_{46}, \varphi_{58}$
5	0,2092	$\varphi_2, \varphi_{58}, \varphi_{13467}$
6	0,2092	$\varphi_2, \varphi_{1345678}$
7	0,6000	$\varphi_{1235678}$

За семь итераций все формулы слились в один кластер. Δ используется в качестве критерия остановки работы алгоритма – например, можно задать $\Delta = 1500$, тогда алгоритм остановится после четвёртой итерации и выдаст результат из четырёх кластеров: $\varphi_{13}, \varphi_2, \varphi_{46}, \varphi_{58}$.

Теперь применим к множеству формул *test* 1 алгоритм k-средних.

Допустим, нам необходимо получить 3 кластера. Выбираем стандартные веса.

На основе матрицы расстояний выбираются 3 центра: $\varphi_2, \varphi_4, \varphi_5$ (центры примерно равноудалены друг от друга, и сумма расстояний между ними наибольшая).

Распределяем оставшиеся формулы по центрам. Получаем кластеры:$\varphi_2, \varphi_{13467}, \varphi_{58}$. После этого ищем центры масс и снова распределяем формулы по обновленным центрам. Получаем кластеры: $\varphi_2, \varphi_{13467}, \varphi_{58}$. Кластеры не изменились. Значит, алгоритм останавливается и выдает получившиеся кластеры в качестве результата.

Результаты кластеризации множества *test* 1 двумя алгоритмами не совпадают с результатами, для получения которых использовались расстояние и мера нетривиальности с жёсткими весами [1]. Это показывает целесообразность использования различных расстояний в алгоритмах кластеризации.

Для дальнейших численных экспериментов с программой был создан банк из 100 логических формул, откуда случайным образом выбирались подмножества формул для кластеризации.

Затем задавалась размерность логики L_n, выбирался алгоритм кластеризации и вводились коэффициенты λ_{kl}. Результаты кластеризации выводились в виде таблиц с итерациями алгоритма.

Пусть $n = 9$. Рассмотрим следующее множество *test* 2 формул девятизначной логики L_9:

$\varphi_1 = \neg(z \vee y)$;

$\varphi_2 = (x \to y) \to w$;

$\varphi_3 = \neg((x \to y) \wedge z)$;

$\varphi_4 = (x \vee z) \wedge y$;

$\varphi_5 = z \to (x \vee y)$;

$\varphi_6 = (\neg y \vee (x \to z))$;

$\varphi_7 = w \wedge (x \to z)$;
$\varphi_8 = (y \vee z) \to (x \vee w)$.
$\varphi_9 = z \to (x \wedge w)$;
$\varphi_{10} = \neg(x \to y)$.
Применим к множеству *test* 2 иерархический алгоритм кластеризации.
$\lambda_0 = 0, \lambda_1 = \frac{1}{30}, \lambda_2 = \frac{1}{20}, \lambda_3 = \frac{1}{10}, \lambda_4 = \frac{1}{5}, \lambda_5 = \frac{3}{10}, \lambda_6 = \frac{2}{5}, \lambda_7 = \frac{3}{5}, \lambda_8 = 1.$

Итерация	Δ	Кластеры
0	0,0000	$\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5, \varphi_6, \varphi_7, \varphi_8, \varphi_9, \varphi_{10}$
1	0,0073	$\varphi_1, \varphi_{23}, \varphi_4, \varphi_5, \varphi_6, \varphi_7, \varphi_8, \varphi_9, \varphi_{10}$
2	0,0173	$\varphi_{23}, \varphi_4, \varphi_5, \varphi_{167}, \varphi_8, \varphi_9, \varphi_{10}$
3	0,0952	$\varphi_{23}, \varphi_4, \varphi_5, \varphi_{167}, \varphi_{89}, \varphi_{10}$
4	0,0952	$\varphi_{23}, \varphi_5, \varphi_{1467}, \varphi_{89}, \varphi_{10}$
5	0,1907	$\varphi_{23}, \varphi_5, \varphi_{1467}, \varphi_{8910}$
6	0,1907	$\varphi_{235}, \varphi_{1467}, \varphi_{8910}$
7	0,3433	$\varphi_{235}, \varphi_{14678910}$
8	0,5892	$\varphi_{12345678910}$

Теперь зададим размерность логики $n = 10$ и снова применим иерархический алгоритм к множеству формул *test* 2.

$$\lambda_0 = 0, \lambda_1 = \frac{1}{30}, \lambda_2 = \frac{1}{20}, \lambda_3 = \frac{1}{10}, \lambda_4 = \frac{1}{5}, \lambda_5 = \frac{3}{10}, \lambda_6 = \frac{2}{5}, \lambda_7 = \frac{3}{5}, \lambda_8 = \frac{7}{10}, \lambda_9 = 1.$$

Итерация	Δ	Кластеры
0	0,0000	$\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5, \varphi_6, \varphi_7, \varphi_8, \varphi_9, \varphi_{10}$
1	0,0081	$\varphi_1, \varphi_{23}, \varphi_4, \varphi_5, \varphi_6, \varphi_7, \varphi_8, \varphi_9, \varphi_{10}$
2	0,0081	$\varphi_1, \varphi_{23}, \varphi_4, \varphi_5, \varphi_{67}, \varphi_8, \varphi_9, \varphi_{10}$
3	0,0173	$\varphi_{23}, \varphi_4, \varphi_5, \varphi_{167}, \varphi_8, \varphi_9, \varphi_{10}$
4	0,0952	$\varphi_{23}, \varphi_4, \varphi_5, \varphi_{167}, \varphi_{89}, \varphi_{10}$
5	0,0952	$\varphi_{23}, \varphi_5, \varphi_{1467}, \varphi_{89}, \varphi_{10}$
6	0,1907	$\varphi_{23}, \varphi_5, \varphi_{1467}, \varphi_{8910}$
7	0,1907	$\varphi_{235}, \varphi_{1467}, \varphi_{8910}$
8	0,3864	$\varphi_{235}, \varphi_{14678910}$
9	0,5121	$\varphi_{12345678910}$

Пусть $n = 7$. Рассмотрим следующее множество *test* 3 формул семизначной логики L_7:
$\varphi_1 = y \to (x \wedge z)$;
$\varphi_2 = \neg z \to w(x \wedge y)$;
$\varphi_3 = z \to (x \vee y)$;
$\varphi_4 = \neg((x \wedge y) \vee z) \to w$;
$\varphi_5 = \neg(x \wedge z) \to y$;
$\varphi_6 = (\neg y \vee (x \to z))$;
$\varphi_7 = z \to (x \vee y)$;
$\varphi_8 = \neg((z \wedge y) \to x$.
$\varphi_9 = \neg z \to x$;

$\varphi_{10} = \neg((x \wedge y) \vee z) \to w$;
$\varphi_{11} = y \to (x \wedge w) \to z$;
$\varphi_{12} = y \vee (x \to z)$;
$\varphi_{13} = z \wedge (x \to y)$;
$\varphi_{14} = (x \wedge z) \to w$;
$\varphi_{15} = (x \vee w) \to y$.

Применим к множеству *test* 3 алгоритм k-средних.

Допустим, нам необходимо получить 4 кластера. Выбираем стандартные веса. Подбираем центры будущих кластеров: $\varphi_2, \varphi_5, \varphi_7, \varphi_9$. Затем распределяем формулы по выбранным центрам, формируя первые кластеры, после чего вычисляем центры масс и обновляем получившиеся кластеры. Алгоритм останавливается на третьей итерации и выдаёт кластеры: φ_{23810}, φ_{145}, φ_{67}, $\varphi_{9111214}$. Для выбора наилучшей кластеризации можно использовать следующий принцип – диаметр кластера должен быть как можно меньше, а расстояние между кластерами – как можно больше. Возможно построить функционал от этих двух величин и использовать его для выбора наилучшей кластеризации. В качестве такого функционала возьмём отношение максимума из диаметров n получившихся кластеров D_i к сумме всех расстояний между получившимися кластерами ρ_i

$$F = \frac{max\{D_i\}}{\sum\limits_{i=1}^{n} \rho_i}.$$

Этот функционал можно вычислить для каждой использующейся кластеризации. Наилучшей будет та, для которой F минимален.

Заключение

В данной работе выполнены следующие задачи. Расстояние и мера нетривиальности формул обобщены на n-значный случай с возможностью выбора весов. Доказаны теоремы, в которых исследованы свойства этих величин. Также рассмотрен случай известных истинностных значений некоторых переменных. На основе старого расстояния с использованием теоретико-модельных свойств введено счётное число новых расстояний (а значит, и мер нетривиальности), которые можно теперь использовать для анализа баз знаний, создания экспертных систем, а также для построения логических решающих функций, в распознавании образов.

На основе новых величин для кластеризации множеств экспертных высказываний адаптированы два алгоритма – иерархический и алгоритм k-средних. Для примера вычислений выбрана n-значная логика L_n, однако адаптированные нами алгоритмы работают и с другими многозначными логиками, вот только результаты могут оказаться другими, поскольку сами модели с их значениями будут распределены, вообще говоря, по-другому. Алгоритмы выдают результаты кластеризаций для различных размерностей логик и различных множеств формул.

Проведена кластеризация множеств формул многозначной логики. Число формул, размерность логики и коэффициенты расстояния выбирались различными. Рассмотрены случаи n=5, n=7, n=9, n=10. Проведено сравнение результатов кластеризации с результатами предыдущих работ для случая n=5, предложен критерий выбора наилучшего алгоритма кластеризации.

В дальнейшем планируется применение и других новых метрик и мер нетривиальностей для анализа множеств логических высказываний экспертов или из базы знаний, порождение ими "коллективных"расстояний с весами, с учетом их индексов качества кластеризации. Планируется также добавление в программу новых алгоритмов кластеризаций и истиностных таблиц нечетких логик.

Список литературы

1 *Викентьев А.А., Кабанова Е.С.* Расстояния между формулами пятизначной логики Лукасевича и мера недостоверности высказываний экспертов в кластеризации баз знаний // Вестн. Томск. гос. ун-та. — 2013. — № 2(23). — С. 121–129.

2 *Ершов Ю.Л., Палютин Е.А.* Математическая логика. — М.: Наука, 1987.

3 *Berikov V.B.* Grouping of objects in a space of heterogeneous variables with the use of taxonomic decision trees // Pattern Recognition And Image Analysis. — 2011. — Vol. 21. — No. 4. — P. 591–598.

4 *Авилов М.С.* Программный комплекс вычислений расстояний, мер нетривиальности и кластеризации множеств высказываний в n-значной логике // Студент и научно-технический прогресс: материалы 52-й Междунар.научн. студ. конф.-Новосибирск, 2015.

А.А.Викентьев, М.С.Авилов

Логикалық білім қорынан автоматтандырылған формулалар класстеризациясында көпмәнді логика формулалары үшін нольдік емес өлшемдер мен жаңа толық метрикалар II

Мақалада келесі есептер қарастырылған: параметрлерге шектеу алынып, кез келген n-мәнді логиканың жағдайына бұрын енгізілген нольдік емес өлшемдер мен аралықтар жалпыланған; жаңа аралықтарды қолдана отырып, n-мәнді логикадағы формулалар жиындарын кластеризациялау алгоритмдері құрылған; әр түрлі білім қорынан формулалардың ішкі жиындарына кластеризация жүргізіліп, кластеризацияның әр түрлі алгоритмдерінің нәтижелерін салыстыру тәсілдері ұсынылған.

A.A.Vikent'ev , M.S. Avilov

New comprehensive metrics and nontrivial steps to formulas multi-valued logic in the automatic clustering of logical formulas Knowledge Base. II

This article describes the following tasks: Summarized previously entered distances and non-triviality of the measures on the n-digit case of any logic, removing restrictions on the parameters; With the new range, developed algorithms for clustering sets of formulas in the n-valued logic; Conducted clustering subsets of formulas from different knowledge bases and provides methods of comparison adapted results of various clustering algorithms.

References

1 *Vikent'ev A.A., Kabanova E.S. Bull. of Tomsk State University*, 2013, 2(23), p. 121–129.

2 Ershov Yu.L., Palyutin E.A. Mathematical logic, Moscow: Nauka, 1987.

3 Berikov V.B. *Pattern Recognition And Image Analysis*, 2011, 21, 4, p. 591–598.

4 Avilov M.S. *Student and scientific-technical progress*: proceedings of the 52 nd International Scientific Student Conference, Novosibirsk, 2015.