УДК 519.67–519.24

А.А.Викентьев[1,2], Е.С.Кабанова[2]

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

Метрика между формулами n-значной логики и недостоверность в кластеризации высказываний

Объект исследования — высказывания, которые можно записать логическими многозначными формулами. В статье, используя теорию моделей, вводятся понятия более полного расстояния между формулами *n*-значной логики Лукасевича и меры недостоверности высказываний, кроме того, сформулированы теоремы о свойствах этих величин. С помощью введённых расстояния и меры недостоверности на основе известных алгоритмов кластеризации разработаны алгоритмы для кластеризации множеств высказываний и исследованы результаты для различных *n*.

Ключевые слова: многозначная логика, логика Лукасевича, расстояние между формулами, мера недостоверности, теория моделей, истинность на моделях, кластеризация множеств логических формул.

1. Введение

Задача определения меры близости между знаниями была поставлена ещё Г.С.Лбовым и Н.Г.Загоруйко. Под знаниями подразумеваются краткие обобщённые описания информации, содержащейся в данных [1–3]. В настоящей работе такими знаниями являются высказывания экспертов, которые можно записать в виде формул конечнозначной логики Лукасевича. Ранее для случая многозначной логики были введены расстояние и мера опровержимости, которые были необходимы для изучения экспертной информации [2, 4–7]. В работах [8–12], аналогично случаю классической логики [2, 4], нами были введены следующие величины: новое расстояние (более полно учитывающее промежуточные пары логических значений) между формулами пятизначной логики Лукасевича $Ł_5$ и мера недостоверности высказываний экспертов (равная (для высказывания) расстоянию между соответствующей высказыванию формулой $Ł_5$ до тождественно истинной формулой). Также были определены и доказаны свойства этих величин, учитывающие семантику сходства и различия информации в этих высказываниях. Для дальнейшего использования этих величин на практике необходимо обобщить их на случай n-значной логики $Ł_n$ (для $n \geq 2$) и доказать свойства, что является первой задачей данной работы. Свойства могут применяться в анализе баз знаний. Одно из возможных применений расстояния и меры недостоверности — использование их в кластеризации конечных множеств высказываний, которые можно записать в виде логических формул. Таким образом, вторая задача данной работы — показать применение этих величин в кластеризации множеств формул с учётом особенностей многозначной логики. Ранее кластеризация множеств логических формул (а тем более, многозначных) не использовалась. Также необходимо проанализировать результаты адаптированных к логическим формулам алгоритмов кластеризации высказываний при различных n.

2. Логика Лукасевича $Ł_n$

Определение 2.1. Пропозициональный язык L:

- x, y, z,… — формулы;
- $\neg$, $\rightarrow$ — пропозициональные связки (отрицание и импликация);
- (,) — вспомогательные символы.

Определение 2.2. Формула:

- x, y, z,… — формулы;
- если φ и ψ — формулы, то $\neg\varphi$ и $\varphi \rightarrow \psi$ — тоже формулы;
- никакие другие конечные последовательности из исходных символов, кроме построенных в силу пунктов 1-2, формулами не являются.

n-Значная матричная логика Лукасевича Ł$_n$ определяется логической матрицей $M_n =< V_n, \neg, \to, \{1\} >$, где $V_n = \left\{0, \frac{1}{n-1}, ..., \frac{n-2}{n-1}, 1\right\}$ — множество значений истинности; $\neg$, $\to$ — унарная операция отрицания и бинарная операция импликации соответственно, определённые на множестве V_n; $\{1\}$ — выделенное значение истины.

Логические операции на множестве V_n определяются следующим образом:

– $\neg x = 1 - x$ (отрицание) ;

– $x \to y = \min\{1, 1 - x + y\}$ (импликация).

Через эти операции выражаются другие:

– $x \vee y = (x \to y) \to y = \max\{x, y\}$ (дизъюнкция);

– $x \wedge y = \neg(\neg x \vee \neg y) = \min\{x, y\}$ (конъюнкция) [2].

3. Основные определения и обозначения

Σ — конечное множество формул Ł$_n$.

$S(\varphi)$ — множество переменных, используемых при написании формулы φ логики Ł$_n$ (носитель формулы $\phi \in \Sigma$).

$S(\Sigma) = \bigcup_{\phi \in \Sigma} S(\varphi)$ — множество переменных, участвующих в написании всех формул из Σ (носитель множества Σ).

Запись $\phi_{\frac{k}{n-1}}$ означает, что формула φ принимает на модели значение $\frac{k}{n-1}$, $k = 0, ..., n-1$.

Определение 3.1. Назовём моделью *M* кортеж из означенных переменных и значение формулы при данном означивании.

M не содержит одновременно $\varphi_{\frac{k}{n-1}}$ и $\varphi_{\frac{l}{n-1}}$ для любого $k \neq l$. Обозначим множество всех моделей как $P(S(\Sigma))$. Ясно, что $\left|P(S(\Sigma))\right| = n^{|S(\Sigma)|}$ [5].

Впервые использование теории моделей и модели для определения расстояния между логическими формулами предложил А. А.Викентьев [2, 4–7].

В работе [5] определены свойства моделей и другие связанные определения.

Определение 3.2. Назовём формулы φ и ψ эквивалентными ($\varphi \equiv \psi$), если они имеют одно и то же множество моделей для каждого значения истинности [2, 4, 5].

В дальнейшем для краткости будем пользоваться следующими обозначениями:

$M(\varphi_{\frac{k}{n-1}})$ — количество моделей, на которых формула φ принимает значение $\frac{k}{n-1}$; $M(\frac{k}{n-1}, \frac{l}{n-1})$ — количество моделей, на которых φ принимает значение $\frac{k}{n-1}$, а ψ — $\frac{l}{n-1}$.

4. Расстояние между формулами Ł$_n$

В данном разделе результаты для случая пятизначной логики [8, 9, 11, 12] естественно обобщены на *n*-значный случай. Для определения расстояния мы учитываем разницу между значениями двух формул на каждой модели.

Естественно предположить, что чем меньше модуль разности между значениями φ и ψ, тем формулы более близки в данной модели. Следовательно, умножим количество моделей с одинаковыми модулями разности на коэффициент, учитывающий близость значений формул. В качестве таких коэффициентов возьмём *n* истинностных значений для Ł$_n$:

$$\tilde{\rho}(\varphi, \psi) = 0 \cdot \left(M(0,0) + M(\frac{1}{n-1}, \frac{1}{n-1}) + M(\frac{2}{n-1}, \frac{2}{n-1}) + ... + M(\frac{n-2}{n-1}, \frac{n-2}{n-1}) + M(1,1) \right) +$$

$$+ \frac{1}{n-1} \cdot \left(M(0, \frac{1}{n-1}) + M(\frac{1}{n-1}, 0) + M(\frac{1}{n-1}, \frac{2}{n-1}) + ... + M(\frac{n-2}{n-1}, 1) + M(1, \frac{n-2}{n-1}) \right) + ...$$

$$...+\frac{n-2}{n-1}\cdot\left(M(0,\frac{n-2}{n-1})+M(\frac{n-2}{n-1},0)+M(\frac{1}{n-1},1)+M(1,\frac{1}{n-1})\right)+$$
$$+1\cdot\left(M(0,1)+M(1,0)\right)=\sum_{k=0}^{n-1}\sum_{l=0}^{n-1}\frac{|k-l|}{n-1}\cdot M(\frac{k}{n-1},\frac{l}{n-1}).$$

Величину, стоящую возле коэффициента 0, планируем учитывать для определения степени близости в спорных случаях (при одинаковых значениях расстояния и меры недостоверности при кластеризации). А для введения расстояния надо только нормировать величину $\tilde{\rho}$.

Определение 4.1. Расстоянием между формулами φ и ψ n-значной логики Ł$_n$ при $S(\varphi)\cup S(\psi)\subseteq S(\Sigma)$ на множестве $P(S(\Sigma))$ назовём величину

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

Следующая теорема показывает, что величина, определённая равенством (1), действительно является расстоянием.

Теорема 1. «Расстояние» между двумя формулами Ł$_n$, определённое равенством (1), для любых $\varphi,\psi,\chi\in\Sigma$ удовлетворяет следующим свойствам:

1) $0\le\rho(\varphi,\psi)\le 1$;
2) $\rho(\varphi,\psi)=0\Leftrightarrow\varphi\equiv\psi$;
3) $\rho(\varphi,\psi)=\rho(\psi,\varphi)$;
4) $\rho(\varphi,\psi)\le\rho(\varphi,\chi)+\rho(\chi,\psi)$;
5) $\varphi\equiv\varphi_1,\psi\equiv\psi_1\Rightarrow\rho(\varphi,\psi)=\rho(\varphi_1,\psi_1)$;
6) $\rho(\varphi,\psi)=\rho(\neg\varphi,\neg\psi)$;
7) $\rho((\varphi\wedge\psi),(\varphi\vee\psi))=\rho(\varphi,\psi)$;
8) $\rho(\varphi,\psi)=\rho(\varphi\to\psi,\psi\to\varphi)$.

Доказательство рутинно и аналогично доказательству этой теоремы для случая пятизначной логики [6, 7, 12].

Пример. В случае пятизначной логики $\rho(x\wedge y,x\vee y)=0.4$, $\rho(x\wedge y\wedge z\wedge w,x\to w)=0.2576$.

Замечание 1. Свойства 2)-4) — это свойства метрики. Таким образом, мы получили метрическое пространство на классах эквивалентных высказываний.

Замечание 2. Без ограничения общности можно считать, что формулы φ и ψ — это формулы от одного числа переменных, некоторые переменные в которых принимают константные значения.

Обобщение расстояния между формулами Ł$_n$. Расстояние, заданное формулой (1), — это расстояние для случая, когда все значения всех переменных заранее не известны. Теперь рассмотрим случай, когда известны конкретные истинностные значения некоторых переменных (например, $x_1 = 0$, или x_1 точно не равно 1 и $\frac{n-2}{n-1}$). В работах [8, 9] определено такое расстояние для случая пятизначной логики Лукасевича. Обобщим его на *n*-значный случай.

Пусть переменные $x_1,...,x_p$, $x_i\in S(\phi)\cup S(\psi)$, $i=1,...,p$, $p=|S(\varphi)\cup S(\psi)|$ соответственно принимают $m_1,...,m_p$, $m_i\le n$ истинностные значения. Тогда формула для нахождения расстояния между формулами φ и ψ выглядит следующим образом:

$$\rho'(\varphi,\psi)=\frac{1}{m_1\cdot...\cdot m_p}\cdot\sum_{k=0}^{n-1}\sum_{l=0}^{n-1}\frac{|k-l|}{n-1}\cdot M(\frac{k}{n-1},\frac{l}{n-1}). \quad (2)$$

В данном случае при расчёте расстояния рассматриваются не все модели, а подмножество из $m_1\cdot...\cdot m_p$ моделей. Сам расчёт производится по тому же принципу. Формула (1) является частным случаем (2), если $m_1=...=m_p=n$. Для ρ' справедлива теорема 1 с заменой ρ на ρ'. Доказательство аналогично. Если все $m_1\cdot...\cdot m_p$ модели нумеровать, то формулу (2) можно переписать в виде

$$\rho'(\varphi,\psi)=\frac{1}{m_1\cdot...\cdot m_p}\cdot\sum_{i=1}^{m_1\cdot...\cdot m_p}|M_i(\varphi)-M_i(\psi)|,$$

где $M_i(\varphi)$ — значение формулы φ на модели M_i, $i = 1,...,m_1 \cdot ... \cdot m_p$.

Пример. Пусть $\varphi = (x \to y) \vee z$, $\psi = (x \wedge y) \to z$ — формулы трёхзначной логики Лукасевича. И пусть переменные, входящие в эти формулы, принимают следующие значения: $x \in \{\frac{1}{2}, 1\}$, $y \in \{1\}$, $z \in \{0, \frac{1}{2}, 1\}$. Тогда $\rho'(\varphi, \psi) = 0,3333$ (при этом $\rho(\varphi, \psi) = 0,2037$).

5. Мера недостоверности высказываний

В классической логике под информативностью высказывания подразумевается относительное число моделей, на которых данное высказывание эксперта ложно, или, что то же самое, нормированное расстояние от высказывания до тождественно истинной формулы. Чем больше моделей, на которых высказывание не истинно, тем оно менее достоверно и нетривиально.

Обобщая случай пятизначной логики [8, 9, 11, 12], зададим меру недостоверности для случая конечнозначной логики Ł$_n$.

Определение 5.1. Мера недостоверности I(φ) для формул n-значной логики Ł$_n$, при $S(\varphi) \subseteq S(\Sigma)$ на множестве $P(S(\Sigma))$ задаётся равенством

$$I(\varphi) = \rho(\varphi, 1) = \sum_{i=0}^{n-2} \frac{n-1-i}{n-1} \cdot \frac{M\left(\varphi_{\frac{i}{n-1}}\right)}{n^{|S(\Sigma)|}}. \qquad (3)$$

Теорема 2. Мера недостоверности, определённая равенством (3), для любых формул $\varphi, \psi, \chi \in \Sigma$ удовлетворяет следующим свойствам:

1) $0 \le I(\varphi) \le 1$;
2) $I(\varphi) + I(\neg\varphi) = 1$;
3) $I(\varphi \wedge \psi) \ge \max\{I(\varphi), I(\psi)\}$;
4) $I(\varphi \vee \psi) \le \min\{I(\varphi), I(\psi)\}$;
5) $I(\varphi \vee \psi) + I(\varphi \wedge \psi) \ge I(\varphi) + I(\psi)$;
6) $I(\varphi \wedge \psi) = \rho(\varphi, \psi) + I(\varphi \vee \psi)$;
7) $\rho(\varphi, \psi) \le I(\varphi) + I(\psi)$;
8) $I(\varphi) \ge \rho(\varphi \to \psi, \psi)$;
9) $I(\varphi \to \psi) \le \rho(\varphi, \psi)$;
10) $I(\varphi \vee \psi) \le \rho(\varphi \to \psi, \varphi \wedge \psi)$.

Доказательство комбинаторное и практически не отличается от аналогичной теоремы в известных случаях [6, 7, 12].

6. Кластеризация множеств высказываний

Кластеризация — это разбиение исходного множества объектов на подмножества (кластеры), при котором каждый объект может быть отнесен к одному или нескольким заранее неизвестным классам. Внутри каждого кластера должны оказаться схожие объекты, а объекты разных кластеров должны как можно больше отличаться.

Так как для таких множеств известны только расстояния между формулами и расстояния от каждой формулы до тождественно истинной, поэтому были выбраны два общеизвестных алгоритма кластеризации, (в которых расстояние имеет важное значение) адаптированы в данной работе для кластеризации конечных множеств логических формул.

Иерархический алгоритм. Пусть есть множество объектов I. Кластеризация происходит либо путём агломерации (объединения более мелких кластеров в более крупные), либо путём разделения крупных кластеров на более мелкие. В результате получается следующая структура: совокупность H вложенных подмножеств S (кластеров), удовлетворяющих свойству: при любых S_1 и S_2 из H их пересечение $S_1 \cap S_2$ либо пусто, либо совпадает с одним из них [10]. Графически такая структура представляется в виде дендрограммы (см рис.).

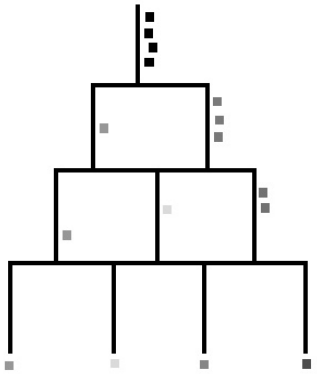

Рисунок. Дендрограмма для кластеризации
множества из четырёх объектов

Иерархический алгоритм для кластеризации множества формул Ł$_n$. Сначала зададим конечное множество формул n-значной логики Лукасевича. Далее перед началом работы алгоритма зададим величину *delta* — максимальную разницу между мерами недостоверности элементов одного кластера. Это является критерием остановки. Построим матрицу расстояний для заданного конечного множества формул (для построения используем расстояние).

Итерация:

Шаг 1. Ищем формулы, между которыми наименьшее расстояние, и объединяем их в один кластер. Если таких формул >2, то:

Случай 1. $\rho(\varphi,\psi) = \rho(\varphi,\chi) = \rho_{\min}$. Тогда объединяем φ,ψ,χ в один кластер.

Случай 2. $\rho(\varphi_1,\varphi_2) = \rho(\varphi_3,\varphi_4) = \rho_{\min}$. Тогда объединяем φ_1,φ_2 в один кластер, а φ_3,φ_4 — в другой.

Шаг 2. Далее объединяем кластеры по одному из методов. В данном случае по методу ближайшего соседа пересчитываем матрицу по следующему правилу:

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

Итерации продолжаются, пока не выполнится критерий остановки (т. е. пока величина *delta* не достигнет заданного значения).

Пример. Рассмотрим множество из восьми формул пятизначной логики Лукасевича:

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

$$\varphi_6 = (\neg y \vee (x \to z)) \to w;\ \varphi_7 = ((x \to y) \to z) \to w;\ \varphi_8 = (w \to z) \wedge (y \to x).$$

Их меры недостоверности соответственно равны:

$$I(\varphi_1) = 0{,}2000;\ I(\varphi_2) = 0{,}8000;\ I(\varphi_3) = 0{,}3000;\ I(\varphi_4) = 0{,}3584;$$
$$I(\varphi_5) = 0{,}3000;\ I(\varphi_6) = 0{,}4092;\ I(\varphi_7) = 0{,}2716;\ I(\varphi_8) = 0{,}3416.$$

Строим матрицу расстояний (табл. 1), используя расстояние (1).

Таблица 1

Матрица расстояний

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0	0,7600	0,1000	0,3416	0,4560	0,3876	0,2500	0,4248
2		0	0,6840	0,5472	0,5000	0,5004	0,6420	0,5032
3			0	0,3248	0,5120	0,3660	0,2460	0,4712
4				0	0,4032	0,0508	0,1300	0,4424
5					0	0,4212	0,4276	0,1416
6						0	0,1688	0,4628
7							0	0,4756
8								0

Наименьшее расстояние = 0,0508, между формулами φ_4 и φ_6. Объединяем их в кластер φ_{46} и далее действуем по алгоритму выше.

Итерация 1: $\min\limits_{i\neq j}\rho(\varphi_i,\varphi_j)=0,0508=\rho(\varphi_4,\varphi_6)$. Кластеры: $\varphi_1,\varphi_2,\varphi_3,\varphi_{46},\varphi_5,\varphi_7,\varphi_8$, delta=0,0508.

Итерация 2: $\min\limits_{i\neq j}\rho(\varphi_i,\varphi_j)=0,1000=\rho(\varphi_1,\varphi_3)$. Кластеры: $\varphi_{13},\varphi_2,\varphi_{46},\varphi_5,\varphi_7,\varphi_8$, delta =0,1000.

Итерация 3: $\min\limits_{i\neq j}\rho(\varphi_i,\varphi_j)=0,1300=\rho(\varphi_7,\varphi_{46})$. Кластеры: $\varphi_{13},\varphi_2,\varphi_{467},\varphi_5,\varphi_8$. delta =0,1376.

Итерация 4: $\min\limits_{i\neq j}\rho(\varphi_i,\varphi_j)=0,1416=\rho(\varphi_5,\varphi_8)$. Кластеры: $\varphi_{13},\varphi_2,\varphi_{467},\varphi_{58}$, delta=0,1376.

Итерация 5: $\min\limits_{i\neq j}\rho(\varphi_i,\varphi_j)=0,2460=\rho(\varphi_{13},\varphi_{467})$. Кластеры: $\varphi_2,\varphi_{58},\varphi_{13467}$, delta =0,2092.

Итерация 6: $\min\limits_{i\neq j}\rho(\varphi_i,\varphi_j)=0,4032=\rho(\varphi_{58},\varphi_{13467})$. Кластеры: $\varphi_2,\varphi_{1345678}$, delta=0,2092.

Итерация 7: $\rho(\varphi_2,\varphi_{1345678})=0,5000$. Кластер $\varphi_{12345678}$, delta =0,6000.

Если перед началом работы алгоритма задаём *delta*, равную, например, 0,1500, то алгоритм останавливается после четвёртой итерации и выдаёт результат:

Кластер 1: φ_1, φ_3. Кластер 2: φ_2. Кластер 3: φ_4, φ_6, φ_7. Кластер 4: φ_5, φ_8.

Алгоритм к-средних (k-means). Пусть имеется множество объектов I. Сначала каким-либо образом выбираются K начальных точек (центров). Затем осуществляется последовательность итераций, каждая из которых состоит их двух шагов:

1. Обновление кластеров. При заданных K центрах C_k, $k=(1,2,..,K)$, каждый объект $i\in I$ приписывается к ближайшему из центров C_k. Таким образом, образуются кластеры S_k, $k=(1,2,..,K)$.
2. Обновление центров. Для каждого кластера S_k вычисляется его центр тяжести (внутрикластерное среднее), который объявляется новым центром C_k'.

Процесс останавливается, когда кластеры на шаге t совпадут с кластерами на шаге $t-1$ [10].

Алгоритм к-средних для кластеризации множества формул Ł$_n$. Рассмотрим конечное множество логических формул Ł$_n$. Центрами будут являться некоторые K формул из данного множества. Сначала определяемся с количеством кластеров, затем подбираем центры кластеров, анализируя матрицу расстояний. Для простоты будем считать следующее:

- центры должны быть примерно равноудалены друг от друга;
- расстояния между кластерами должны быть максимально возможными, с учётом предыдущего пункта.

Итерация:

Шаг 1. Приписываем каждую формулу из множества к ближайшему центру.

Шаг 2. Центр масс — это столбец значений логики Ł$_n$. Для определения этого столбца учитывается специфика многозначных логических формул.

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 $S_a\notin V_n$, то в столбец значений записывается ближайшее снизу (или ближайшее сверху, это определяется до начала работы алгоритма) значение из V_n (чтобы оставаться в том же множестве моделей, т. е. в той же логике Ł$_n$).

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

Пример. Рассмотрим множество из восьми формул из предыдущего примера. Допустим, нам нужно получить три кластера. Анализируя матрицу расстояний, выбираем центрами формулы

$$\varphi_2,\varphi_4,\varphi_5.\ (\rho(\varphi_2,\varphi_4)=0,5472,\ \rho(\varphi_2,\varphi_5)=0,5000,\ \rho(\varphi_4,\varphi_5)=0,4032).$$

Распределяем оставшиеся формулы по центрам. Получаются кластеры:

$$\varphi_2;\ \varphi_1,\ \varphi_3,\ \varphi_4,\ \varphi_6,\ \varphi_7;\ \varphi_5,\ \varphi_8.$$

Ищем центры масс. Рассмотрим наглядно, как это происходит.

Таблица 2

Определение центра масс кластера

x	y	z	w	φ_5	φ_8	C_{58}
0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$	C_1
$\frac{1}{4}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{3}{4}$	C_2
…						

$$C_1 = \left(\frac{1}{2}+\frac{1}{2}\right)/2 = \frac{1}{2} \in \left\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right\}, C_2 = \left(\frac{1}{2}+\frac{3}{4}\right)/2 = \frac{5}{8} \notin \left\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right\}.$$

Допустим, что в качестве значения мы определились брать ближайшее сверху значение. Тогда $C_2 = \frac{3}{4}$. Остальное — аналогично. Таким образом, мы вычисляем центры тяжести кластеров. Снова распределяем формулы по обновлённым центрам. Получаются следующие кластеры:

$$\varphi_2;\ \varphi_1,\ \varphi_3,\ \varphi_4,\ \varphi_6,\ \varphi_7;\ \varphi_5,\ \varphi_8.$$

Кластеры не изменились. Следовательно, алгоритм останавливается и выдаёт получившиеся кластеры в качестве результата. Заметим, что при таком начальном выборе центров получившиеся кластеры совпадают с кластерами на пятой итерации иерархического алгоритма.

7. Примеры

Был создан банк из 250 различных логических формул, откуда случайным образом выбирались подмножества формул. С помощью адаптированных алгоритмов, описанных в предыдущем разделе, было кластеризовано более 30 таких подмножеств (от 8 до 30 формул в каждом подмножестве) при различных *n*, где *n* — это значность логики $Ł_n$. Для данных вычислений расстояние (1) и адаптированные алгоритмы кластеризации были программно реализованы. Сложность вычисления расстояния — экспоненциальная.

Ниже представлены типичные примеры множеств формул, в процессе кластеризации которых образуются несколько неодноэлементных кластеров. Для краткости далее символы. и + в формулах обозначают конъюкцию и дизъюнкцию соответственно.

1. ¬((x.y.z)→w).
2. ¬((x.y.z)→(x+w)).
3. z→(w→(x+y)).
4. w→(x→(y+z)).
5. (x.y)+(y.z)+(z.w).
6. (x+y)+(y+z) →(z+w).
7. ((x.y)→(y+z))→(x.y).
8. ((y.z)→(x+y))→(x+y).
9. ((x+y)→(y+z))→(x.y).
10. z→(w→(x→y)).

Иерархический алгоритм.

Показаны итерации алгоритма при различных *n*. Значение *delta* заранее не задано.

n=2
1: {1}; {2}; {3}; {4}; {5}; {6}; {7}; {8}; {9}; {10} delta = 0,00000
2: {1}; {2}; {3}; {4}; {5}; {6}; {7, 8}; {9}; {10} delta = 0,00000
3: {1}; {2}; {3, 4, 10}; {5}; {6}; {7, 8}; {9} delta = 0,00000
4: {1, 2, 3, 4, 7, 8, 10}; {5}; {6}; {9} delta = 0,37500
5: {1, 2, 3, 4, 5, 6, 7, 8, 10}; {9} delta = 0,43750
6: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} delta = 0,56250
n=3
1: {1}; {2}; {3}; {4}; {5}; {6}; {7}; {8}; {9}; {10} delta = 0,00000
2: {1}; {2}; {3}; {4}; {5}; {6}; {7, 8}; {9}; {10} delta = 0,01852

3: {1}; {2}; {3, 4, 10}; {5}; {6}; {7, 8}; {9} delta = 0,01852
4: {1, 2}; {3, 4, 10}; {5}; {6}; {7, 8}; {9} delta = 0,16667
5: {1, 2}; {3, 4, 10}; {5, 6}; {7, 8}; {9} delta = 0,16667
6: {1, 2}; {3, 4, 7, 8, 10}; {5, 6}; {9} delta = 0,22222
7: {1, 2, 3, 4, 7, 8, 10}; {5, 6}; {9} delta = 0,35802
8: {1, 2, 3, 4, 5, 6, 7, 8, 10}; {9} delta = 0,46296
9: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} delta = 0,57407

n=4

1: {1}; {2}; {3}; {4}; {5}; {6}; {7}; {8}; {9}; {10} delta = 0,00000
2: {1}; {2}; {3}; {4, 10}; {5}; {6}; {7}; {8}; {9} delta = 0,00781
3: {1}; {2}; {3, 4, 10}; {5}; {6}; {7}; {8}; {9} delta = 0,00781
4: {1}; {2}; {3, 4, 10}; {5}; {6}; {7, 8}; {9} delta = 0,01562
5: {1, 2}; {3, 4, 10}; {5}; {6}; {7, 8}; {9} delta = 0,15104
6: {1, 2}; {3, 4, 10}; {5, 6}; {7, 8}; {9} delta = 0,15104
7: {1, 2, 7, 8}; {3, 4, 10}; {5, 6}; {9} delta = 0,15104
8: {1, 2, 3, 4, 7, 8, 10}; {5, 6}; {9} delta = 0,34505
9: {1, 2, 3, 4, 5, 6, 7, 8, 10}; {9} delta = 0,47266
10: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} delta = 0,57682

n=5

1: {1}; {2}; {3}; {4}; {5}; {6}; {7}; {8}; {9}; {10} delta = 0,00000
2: {1}; {2}; {3, 4, 10}; {5}; {6}; {7}; {8}; {9} delta = 0,00840
3: {1}; {2}; {3, 4, 10}; {5}; {6}; {7, 8}; {9} delta = 0,01600
4: {1, 2}; {3, 4, 10}; {5}; {6}; {7, 8}; {9} delta = 0,14600
5: {1, 2}; {3, 4, 10}; {5, 6}; {7, 8}; {9} delta = 0,14600
6: {1, 2, 7, 8}; {3, 4, 10}; {5, 6}; {9} delta = 0,14600
7: {1, 2, 5, 6, 7, 8}; {3, 4, 10}; {9} delta = 0,28760
8: {1, 2, 3, 4, 5, 6, 7, 8, 10}; {9} delta = 0,47760
9: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} delta = 0,57760

n=6

1: {1}; {2}; {3}; {4}; {5}; {6}; {7}; {8}; {9}; {10} delta = 0,00000
2: {1}; {2}; {3}; {4, 10}; {5}; {6}; {7}; {8}; {9} delta = 0,00864
3: {1}; {2}; {3, 4, 10}; {5}; {6}; {7}; {8}; {9} delta = 0,00864
4: {1}; {2}; {3, 4, 10}; {5}; {6}; {7, 8}; {9} delta = 0,01481
5: {1, 2}; {3, 4, 10}; {5}; {6}; {7, 8}; {9} delta = 0,14028
6: {1, 2}; {3, 4, 10}; {5, 6}; {7, 8}; {9} delta = 0,14028
7: {1, 2, 7, 8}; {3, 4, 10}; {5, 6}; {9} delta = 0,14028
8: {1, 2, 3, 4, 7, 8, 10}; {5, 6}; {9} delta = 0,32948
9: {1, 2, 3, 4, 5, 6, 7, 8, 10}; {9} delta = 0,48056
10: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} delta = 0,57778

При *n*=6 и далее состав кластеров на каждой итерации совпадает. Меняется только *delta*.

Алгоритм *к*-средних. Центры — формулы 1, 3, 9. Для определения центра масс рассматриваем как ближайшее снизу, так и ближайшее сверху значение из множества логических значений. *n*=2.
Ближайшее снизу: итераций 2; кластеры: {1, 2, 5, 6}, {3, 4, 7, 8, 10}, {9}; delta = 0,25000.
Ближайшее сверху: итераций 2; кластеры: {1, 2, 5, 6}, {3, 4, 7, 8, 10}, {9}; delta = 0,25000.
При *n*=3 и далее состав кластеров остаётся таким же, как при *n*=2. Меняется только *delta*.

1. ¬(z.(y→x)).
2. (x→y)→(y→x).
3. ¬(y+z).
4. (x→z)→(z→y)→(y→x).
5. y→(x+z).
6. y→(x.z).
7. (x→y).(x+y+z).
8. (y+z)→(x+y+z+w).
9. ¬((x→y)→z).
10. ¬(z→(w→(x+y))).
11. ¬(y→(z→(w→x))).

12.¬(x.(y+z)).
13.¬(y→(x+z)).
14.(x→(y.z))+(x→z).
15.¬((x→z)→y).

Иерархический алгоритм.

Пусть задана *delta*=0,5; значность логики *n*=7. Тогда алгоритм останавливается на десятой итерации и выдаёт кластеры:
{1, 12}; {2, 4, 5, 6, 9, 14, 15}; {3, 8}; {7}; {10, 11}; {13}; delta = 0,22449.
На следующей итерации *delta* равна уже 0,62925.

Алгоритм *к*-средних.

Исходя из результатов кластеризации с помощью иерархического алгоритма, подбираем центры будущих кластеров: 1, 5, 7, 10.

n=2

Ближайшее снизу: итераций 2; кластеры: {1, 12, 13}, {2, 4, 5, 6, 9, 14, 15}, {7}, {3, 8, 10, 11}.

Ближайшее сверху: итераций 2; кластеры: 1, 12, 13}, {2, 4, 5, 6, 9, 14, 15}, {7}, {3, 8, 10, 11}.

n=3

Ближайшее снизу: итераций 2; кластеры: {1, 8, 12, 13}, {2, 4, 5, 6, 9, 14, 15}, {7}, {3, 10, 11}.

Ближайшее сверху: итераций 2; кластеры: {1, 8, 12, 13}, {2, 4, 5, 6, 9, 14, 15}, {7}, {3, 10, 11}.

n=4

Ближайшее снизу: итераций 2; кластеры: {1, 3, 8, 12, 13}, {2, 4, 5, 6, 9, 14, 15}, {7}, {10, 11}.

Ближайшее сверху: итераций 2; кластеры: {1, 3, 8, 12, 13}, {2, 4, 5, 6, 9, 14, 15}, {7}, {10, 11}.

При *n*=5 и далее состав кластеров и количество итераций такое же, как при *n*=4.

8. Наблюдения и выводы для различных n, $n \geq 2$

Исходя из рассмотренных примеров, делаем следующие выводы:

1. Для $n = 2,\ldots,6$ наблюдается разница в составе кластеров. Начиная с $n = 7$, кластеры и последовательность итераций не меняются (7 — это максимальное такое значение n для рассмотренных примеров. Для одних множеств состав кластеров не меняется после $n = 4$, для других — после $n = 5$ и т.д.).

Таким образом, возникает гипотеза о нецелесообразности использования логики большой значности в реальных задачах для формул от малого числа переменных. Частично это подтверждается самой конструкцией введённого в данной работе расстояния.

2. Для алгоритма *к*-средних при вычислении центров масс наблюдаются одни и те же результаты как при замене среднего арифметического ближайшим сверху значением из V_n, так и ближайшим снизу.

9. Заключение

Расстояние между логическими формулами и мера недостоверности высказываний обобщены на случай *n*-значной логики Лукасевича, в полной мере учитывая многозначность. Доказаны свойства этих величин, схожие со свойствами расстояния и меры как в случае классической, так и в случае пятизначной логики. Помимо этого, определены и доказаны новые свойства, включающие в себя операцию импликации, что является очень важным для анализа высказываний, так как реальные высказывания часто имеют вид «если …, то…» или «… следует…».

Также определён общий случай расстояния между логическими формулами, когда некоторые значения переменных заранее известны, что также является актуальным для реальных задач, когда некоторая информация уже задана. Остался вопрос, можно ли придумать другие расстояния с учетом всех логических пар значений для формул? Оказалось, что это возможно, что отражено в работе первого автора с В.Фефеловой [13].

Для кластеризации множеств многозначных высказываний адаптированы два алгоритма кластеризации — иерархический и к-средних (k-means). В обоих случаях используется расстояние между формулами и учитывается специфика формул конечнозначной логики Лукасевича. Результаты работы алгоритмов были исследованы на примерах при различных n.

В дальнейшем планируется провести анализ результатов кластеризации множеств, состоящих из большего, чем тридцать, количества формул, и формализацию этих результатов. Также предполагается применение свойств расстояния и меры недостоверности при анализе множеств высказываний экспертов.

Список литературы

1 *Ершов Ю. Л., Палютин Е.А.* Математическая логика. — 2-е изд. — М.: Наука, 1987. — 336 с.

2 *Лбов Г. С., Старцева Н.Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Изд-во Ин-та математики, 1999. — 212 с.

3 *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Изд-во Ин-та математики, 1999. — 270 с.

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

5 *Викентьев А.А.* Мера опровержимости высказываний экспертов, расстояния в многозначной логике и процессы адаптации // XIV International Conference «Knowledge-Dialogue-Solution» KDS, 2008, Varna, Bulgaria, 2008. — С. 179–188.

6 *Викентьев А. А., Викентьев Р.А.* Расстояния и меры недостоверности на высказываниях *n*-значной логики // Вестн. НГУ. Сер. Математика, механика, информатика. — 2011. — Т. 11. — Вып. 2. — С. 51–64.

7 *Викентьев А. А.* О возможных расстояниях и степенях недостоверности в многозначных высказываниях экспертов и приложение этих понятий в проблемах кластеризации и распознавания // Проблемы информатики. — 2011. — № 3 (11). — С. 33–45.

8 *Kabanova E.* Distance between formulas of the five-valued Lukasiewicz logic and the uncertainty measure of expert statements // 6th International Workshop «Weighted Automata: Theory and Applications» WATA, 2012, Dresden, Germany, 2012. — P. 62, 63.

9 *Кабанова Е.С.* Расстояние между формулами пятизначной логики Лукасевича и мера недостоверности высказываний экспертов // Материалы 50-й юбилейной МНСК «Студент и научно-технический прогресс». — Новосибирск: Изд-во НГУ, 2012.

10 *Кабанова Е.С.* Применение расстояния между формулами конечнозначной логики Лукасевича и меры недостоверности высказываний в кластеризации // Материалы 50-й юбилейной МНСК «Студент и научно-технический прогресс». — Новосибирск: Изд-во НГУ, 2012.

11 *Викентьев А. А., Кабанова Е.С.* Расстояние между формулами пятизначной логики Лукасевича и мера недостоверности высказываний экспертов в кластеризации // Материалы Междунар. науч. конф., посвящ. памяти и 70-летию проф. Т.Г.Мустафина. — Караганда, 2012. — С. 28, 29.

12 *Викентьев А. А., Кабанова Е.С.* Расстояние между формулами пятизначной логики Лукасевича и мера недостоверности высказываний экспертов // Вестн. Караганд. ун-та. — Сер. математика. — 2013. — № 1 (69). — С. 18–27.

13 *Викентьев А.А., Фефелова В.В.* Введение полных расстояний и мер недостоверности для формул логик Лукасевича для автоматической кластеризации множеств логических высказываний из базы знаний // Вестн. Караганд. ун-та. Сер. Математика. — 2015. — Вып. № 3(79). — С. 17–24.

А.А.Викентьев, Е.С.Кабанова

Кластерлеу тұжырымдарындағы қателік және *n*-мағыналы логикалық формулалар арасындағы метрика

Зерттеу нысаны — көп мағыналы логикалық формулалар түрінде жазуға болатын тұжырымдар. Мақалада модельдер теориясын қолданып, n-мағыналы Лукасевич логикасының формулалар және сенімсіз шамалар тұжырымдар арасындағы жалпы қашықтығы, шамалардың қасиеттері туралы теореманың құрылымы қарастырылған. Енгізілген қашықтық пен сенімсіз шамалар көмегімен, танымал алгоритмдарды кластерлеу негізінде жиындар тұжырымын кластерлеуге алгоритм құрылды, әр түрлі *n* үшін нәтижелер зерттелді.

A.A.Vikent'ev, E.S.Kabanova

Metric formulas between the *n*-valued logic and the unreliability of statements in clustering

The object of study — statements that can be written multiple-valued logic formulas. In the article, using the theory of models introduced over the total distance between the formulas n-valued logic of Lukasiewicz and measure the unreliability of statements formulated the theorem about the properties of these variables. With the distance and entered the unreliability of measures, based on the known algorithms for clustering algorithms for clustering sets of statements, we examine the results for different n.

References

1 Ershov Yu.L., Palyutin E.A.*Mathematical Logic, 2nd ed.*, Moscow: Nauka, 1987, 336 p.

2 Lbov G.S., Startseva N.G. *Logical decision functions and the issues of the statistical stability of the solutions,* Novosibirsk: Publ. Institute of Mathematics, 1999, 212 p.

3 Zagoruiko N.G. *Applied methods of data analysis and knowledge,* Novosibirsk: Publ. House of the Institute of Mathematics, 1999, 270 p.

4 Vikent'ev A.A., Lbov G.S. *Pattern Recognition and Image Analysis*, 1997, 7, 2, p. 175–183.

5 Vikent'ev A.A. *XIV International Conference «Knowledge-Dialogue-Solution» KDS, 2008,* Varna, Bulgaria, 2008, p. 179–188.

6 Vikent'ev A.A., Vikent'ev R.A. *Bull. of Novosibirsk State University,* ser. Mathematics, mechanics, computer science, 2011, 11, 2, p. 51–64.

7 Vikent'ev A.A. *Problems of Informatics,* 2011, 3 (11), p. 33–45.

8 Kabanova E. *6th International Workshop «Weighted Automata: Theory and Applications» WATA, 2012,* Dresden, Germany, 2012, p. 62, 63.

9 Kabanova E.S. *Proceedings of the 50th anniversary of ISSC «Student and technological progress»*, Novosibirsk: Publ. NSU, 2012.

10 Kabanova E.S. *Proceedings of the 50th anniversary of ISSC «Student and technological progress»*, Novosibirsk: Publ. NSU, 2012.

11 Vikent'ev A.A., Kabanova E.S. *Proceedings of the international scientific conference dedicated to memory and 70th anniversary of prof. T.Mustafin*, 2012, p. 28, 29.

12 Vikent'ev A.A., Kabanova E.S. *Bull. of Karaganda State University,* ser. Mathematics, 2013, 1 (69), p. 18–27.

13 Vikent'ev A.A., Fefelova V.V. *Bull. of Karaganda State University,* ser. Mathematics, 2015, 3 (79), p. 17–24.