doi:https://doi.org/10.15407/emodel.40.03.063

УДК 004.932

Г.А. Кравцов[1], канд. техн. наук, **В.И. Кошель**[1], аспирант,
А.В. Долгоруков[1], аспирант, **В.В. Цуркан**[2], канд. техн. наук
[1] Ин-т проблем моделирования в энергетике им. Г.Е. Пухова НАН Украины
(Украина, 03164, Киев, ул. Генерала Наумова, 15,
e-mail: hryhoriy.kravtsov@gmail.com, vlad.koshell@gmail.com,
alexander.v.dolgorukov@gmail.com),
[2] Национальный технический университет Украины
«Киевский политехнический ин-т имени Игоря Сикорского»
(Украина, 03056, Киев, пр-т Победы, 37,
e-mail: v.v.tsurkan@gmail.com)

Обучаемая модель вычислений на классификациях

Рассмотрено классическое понятие меры в соответствии с условиями симметричности, рефлексивности и неравенства треугольника. Выдвинуты требования к мере для ее использования в теории вычислений на классификациях. Установлена ограниченность применения функций расстояния, коэффициента корреляции, косинусной меры подобия. Определена и проанализирована применимость используемых на практике мер сходства. Это позволило аргументировать необходимость введения новой меры. Дано формальное определение обучаемой модели вычислений на классификациях.

К л ю ч е в ы е с л о в а: мера сходства, мера отличия, функция расстояния, модель вычислений, обучаемая модель, континиум эквивалентных мер.

В работе [1], где рассматривается модифицированная мера Жаккара (Jaccard) как мера сходства (отличия) классов в классификации, не аргументирована необходимость введения новой меры и не поясняется, чем предложенная модифицированная мера Жаккара отличается от классического (исходного) варианта. Более того, предложенную в [1] модель нельзя назвать «гибкой», что значительно ограничивает сферу ее применения, так как гибкость модели определяет ее способность к обучению, что крайне важно в эру искусственного интеллекта.

Дадим формальное определение обучаемой модели вычислений на классификациях, опираясь на наиболее полный анализ существующих мер и их применимости на классификациях. Но прежде выдвинем требования,

которым должна соответствовать мера, приемлемая для использования в теории вычислений на классификациях. Будем рассматривать понятие меры в его классической интерпретации, а именно мера должна соответствовать трем условиям: симметричности, рефлексивности и неравенства треугольника [2]. Теоретические основы аксиоматического введения мер сходства, различия, совместимости и зависимости применительно к компонентам биоразнообразия даны в работе [3].

О сути критерия «сходство» в работе [4] сказано следующее: «То, что некоторые вещи обнаруживают между собой сходство или различие, является весьма важным моментом для процесса классификации. Несмотря на кажущуюся простоту, понятия сходства и особенно процедуры, используемые при измерении сходства, не так просты. В самом деле, понятие сходства тесно связано с такими основополагающими эпистемологическими проблемами, как: «Каким образом мы можем образовывать полезные абстрактные понятия, позволяющие внести порядок в то, что мы знаем?». Конечно, чтобы ответить на этот вопрос, нужно уметь рассортировывать вещи по классам, что требует умения объединять вещи, воспринимающиеся как схожие. Проблема сходства состоит, однако, не в простом распознавании сходных или несходных вещей, а в том, какое место эти понятия занимают в научных исследованиях. Наука для плодотворного развития должна базироваться на объективных, воспроизводимых процедурах; таким образом, разработка статистических процедур для измерения более «объективного» сходства вещей является естественным следствием необходимости в воспроизводимых и надежных классификациях. Количественное оценивание сходства отталкивается от понятия метрики. При этом подходе к сходству события представляются точками координатного пространства, причем замеченные сходства и различия между точками находятся в соответствии с метрическими расстояниями между ними, где размерность пространства определяется числом переменных, использованных для описания событий».

Напомним, что классификация представляет собой ориентированное дерево, которое является системой мериологических или таксономических делений [1]. Классы классификации, на которые делится произвольно выбранный класс, называются уточняющими (дочерними). Класс, который делится на уточняющие классы, по отношению к уточняющим классам, а также к уточняющим классам уточняющих классов, является обобщающим или родительским. Класс, который является обобщающим (родительским) для любого класса из классификации кроме самого себя, является суперклассом и представляет всю классификацию в целом.

Переходом называется ребро в ориентированном дереве классификации, соединяющее обобщающий и уточняющие классы. Переходы, со-

гласно определению классификации, являются ориентированными и направлены в сторону уточнения (деления), т.е. от обобщающего класса к уточняющему. В переходе обобщающий класс называется основанием перехода, а уточняющий класс — назначением перехода. Путь в классификации — это совокупность переходов от суперкласса до некоторого произвольного класса классификации. Число переходов от суперкласса до некоторого произвольного класса есть ранг произвольного класса.

Аксиома. Любой путь в классификации является кратчайшим.

Аксиома верна, так как из суперкласса до произвольного класса существует единственная цепочка переходов и не существует альтернатив — она кратчайшая.

Дополнительно потребуем следующее.

Требование 1. Мера отличия двух уточняющих классов некоторого класса классификации с бесконечной длиной пути стремится к нулю.

Требование 2. Мера отличия между двумя различными уточняющими классами произвольно выбранного класса есть величина постоянная.

Наиболее используемыми мерами являются следующие [2]:

евклидово расстояние —

$$d_2(X_i, X_j) = \left[\sum_{k=1}^{p}(x_{k,i} - x_{k,j})^2\right]^{1/2};$$

l_1-норма —

$$d_1(X_i, X_j) = \left[\sum_{k=1}^{p}| x_{k,i} - x_{k,j}|\right];$$

сюпремиум-норма —

$$d_\infty(X_i, X_j) = \sup\{| x_{k,i} - x_{k,j}|\},\ k = 1,2,..., p;$$

l_p-норма —

$$d_p(X_i, X_j) = \left[\sum_{k=1}^{p}(x_{k,i} - x_{k,j})^p\right]^{1/p};$$

механобиса —

$$D^2(X_i, X_j) = (X_i - X_j)^T W^{-1}(X_i, X_j);$$

Джефриса-Матуситы —

$$M = \left[\sum_{k=1}^{p}\sqrt{x_{k,i}} - \sqrt{x_{k,j}}\right]^{1/2};$$

коэффициент дивергенции —

$$CD = \left\{ \frac{1}{p} \sum_{k=1}^{p} \left[\frac{x_{k,i} - x_{k,j}}{x_{k,i} + x_{k,j}} \right]^2 \right\}^{1/2}.$$

Однако указанные функции расстояния в общем виде не являются мерами сходства, для которых должны быть выполнены следующие условия [2]:

$$0 \le s(X_i, X_j) < 1,\ X_i \ne X_j,\ s(X_i, X_j) = 1,\ s(X_i, X_j) = s(X_j, X_i). \quad (1)$$

В статистике постоянно пользуются мерой линейного сходства, называемой коэффициентом корреляции $r_{i,j}$, который вычисляется по формуле [2]

$$r_{i,j} = \left[\sum_{k=1}^{p} x_{k,i} x_{k,j} \right] \Big/ \left[\sum_{k=1}^{p} x_{k,i}^2 \sum_{k=1}^{p} x_{k,j}^2 \right]^{1/2}.$$

В [2] утверждается, что коэффициент корреляции часто используется ошибочно. При этом подчеркивается, что если X_i и X_j рассматривать как координаты двух точек в пространстве E_p, которые являются концами двух векторов с началом в начале координат, то $r_{i,j} = \cos(\Theta)$, где Θ — угол между этими векторами. Очевидно, что коэффициент корреляции не является мерой сходства, так как $-1 \le r_{i,j} \le 1$. Если не важна сонаправленность векторов, то избежать этого неудобного момента можно с помощью функции модуля, что часто применяется на практике.

Однако в задачах вычислений на классификациях ни коэффициент корреляции, ни косинусная мера подобия не находят применения. Объясняется это тем, что в классификациях работают с некоторыми абстрактными метками классов (переменными), игнорируя какую-либо информацию, кроме путей до выбранных классов от суперкласса.

В [4] предложено рассматривать меры сходства на множествах, основанные на «совстречаемости» в условиях, когда элемент есть некий псевдоним или обозначение объекта и сам по себе не отражает его сути (переменная). Исследованные авторами [4] меры построены с помощью матрицы ассоциативности, в которой 1 указывает на наличие переменной, а 0 — на ее отсутствие:

	1	0
1	*a*	*b*
0	*c*	*d*

Из предложенных в работе [4] более тридцати различных мер лишь небольшое их число подверглось широкой проверке.

Однако существует и другая система переменных, часто используемая при рассмотрении мер сходства, а именно система пробных площадок:

число видов на первой пробной площадке — a;

число видов на второй пробной площадке — b;

число видов, общих для первой и второй площадок — c.

Рассмотрим проверенные и используемые на практике меры сходства: простой коэффициент совстречаемости, коэффициент Жаккара и коэффициент Гауэра.

Простой коэффициент совстречаемости в терминах матрицы ассоциативности имеет вид

$$S = \frac{(a+d)}{(a+b+c+d)},$$

где S — сходство между двумя объектами, изменяющееся от нуля до единицы. Ссылаясь на более ранние труды, авторы работы [4] констатируют, что простой коэффициент совстречаемости нелегко преобразовать в метрику, хотя было направлено много усилий на то, чтобы установить приблизительные доверительные пределы. К положительным свойствам простого коэффициента следует отнести способность учитывать одновременное отсутствие признака у обоих объектов, как показано в матрице ассоциативности вариантом d.

Коэффициент Жаккара [4, 5], известный как коэффициент флористической общности, определяемый (в терминах матрицы ассоциативности) в виде $J = \frac{a}{(a+b+c)}$ не учитывает одновременного отсутствия признака при вычислении сходства (вариант d матрицы ассоциативности не рассматривается). Подобно простому коэффициенту совстречаемости коэффициент Жаккара изменяется от нуля до единицы. В терминах системы пробных площадок коэффициент Жаккара представляем в виде

$$K_J = \frac{c}{(a+b-c)}.$$

Как показано в [1], коэффициент Жаккара не является строгой математической мерой. Однако выражение $\bar{J} = 1 - J$ есть строгая математическая мера, для которой выполняются симметричности, рефлексивности и неравенства треугольника.

Строгое доказательство выполнения требований симметричности, рефлексивности и неравенства треугольника для $\bar{J}$ в системе аксиом Цермело—Френкеля (ZF), а также в системе аксиом Цермело—Френкеля с аксиомой выбора (ZFC), приведено в работе [6]. В [1] приведено дока-

зательство выполнения для $\bar{J}$ требований симметричности, рефлексивности и неравенства треугольника в системе аксиом NBG (von Neumann — Bernays — Gödel). Полученный в [1] результат закономерен, так как система NBG равнозначна системе ZF, поскольку любая теорема о множествах (в которой не упоминаются классы), доказуемая в одной системе, доказуема и в другой.

Рассмотрим классическое множественное представление меры Жаккара

$$K_{1,-1} = \frac{n(A \cap B)}{n(A \cup B)},$$

где $n(A)$ — мощность множества A. Пусть даны два класса классификации, y и x, в классификации A с суперклассом a. Теоретически в классификации возможны следующие пути от суперкласса до классов y и x: $a \to b \to c \to d \to \to y$, $a \to b \to c \to e \to x$ и $a \to d \to b \to c \to y$, где a, b, c, d, e, x, y — классы классификации A.

Рассмотрим матрицу попарного сходства $K_{1,-1}$:

	$a \to b \to c \to d \to y$	$a \to b \to c \to e \to x$	$a \to d \to b \to c \to y$
$a \to b \to c \to d \to y$	$\frac{n\{a,b,c,d,y\}}{n\{a,b,c,d,y\}}$	$\frac{n\{a,b,c\}}{n\{a,b,c,d,y,e.x\}}$	$\frac{n\{a,b,c,d,y\}}{n\{a,b,c,d,y\}}$
$a \to b \to c \to e \to x$	$\frac{n\{a,b,c\}}{n\{a,b,c,d,y,e,x\}}$	$\frac{n\{a,b,c,e,x\}}{n\{a,b,c,e,x\}}$	$\frac{n\{a,b,c\}}{n\{a,b,c,d,y,e,x\}}$
$a \to d \to b \to c \to y$	$\frac{n\{a,b,c,d,y\}}{n\{a,b,c,d,y\}}$	$\frac{n\{a,b,c\}}{n\{a,b,c,d,y,e,x\}}$	$\frac{n\{a,b,c,d,y\}}{n\{a,b,c,d,y\}}$

или

	$a \to b \to c \to d \to y$	$a \to b \to c \to e \to x$	$a \to d \to b \to c \to y$
$a \to b \to c \to d \to y$	1	3/7	1
$a \to b \to c \to e \to x$	3/7	1	3/7
$a \to d \to b \to c \to y$	1	3/7	1

Отсюда видно, что $a \to b \to c \to d \to y$ и $a \to d \to b \to c \to y$ абсолютно сходны при оценке схожести согласно классическому множественному представлению меры Жаккара $K_{1,-1}$. Решается эта задача просто — необходимо в мере Жаккара использовать не множество классов из классификации, а множество переходов от суперкласса до конкретного класса, а мощность множества переходов увеличить на единицу. Покажем, как это будет выглядеть для $a \to b \to c \to d \to y$ и $a \to d \to b \to c \to y$:

$$K_{1,-1} = \frac{n(A \cap B)+1}{n(A \cup B)+1} =$$

$$= \frac{n\{b \to c\}+1}{n\{a \to b, b \to c, c \to d, d \to y, a \to d, d \to b, c \to y\}+1} = \frac{1}{4}, \qquad (2)$$

где $A = \{a \to b, b \to c, c \to d, d \to y\}$, $B = \{a \to d, d \to b, b \to c, c \to y\}$ — множества переходов. Мера Жакарра в виде (2) получила название модифицированной меры Жаккара в работе [1], где описаны ее свойства применительно к классификациям.

Коэффициент Гауэра — единственный в своем роде, так как при оценке сходства допускает одновременно использование переменных, измеренных по разным шкалам [4]:

$$s_{i,j} = \frac{\sum_{k=1}^{p} S_{i,j,k}}{\sum_{k=1}^{p} W_{i,j,k}},$$

где $W_{i,j,k}$ — весовая переменная, принимающая значение 1, если сравнение объектов по признаку k следует учитывать, и значение 0 — в противном случае; $S_{i,j,k}$ — «вклад» в сходство объектов, зависящий от того, учитывается ли признак k при сравнении объектов i и j. В случае бинарных признаков $W_{i,j,k} = 0$, если признак k отсутствует у одного или обоих сопоставляемых объектов. Для так называемых «негативных переменных» [4] $W_{i,j,k} = 0$. Понятно, что если все данные — двоичные, то коэффициент Гауэра сводится к коэффициенту Жаккара.

Коэффициент Серенсена — бинарная мера сходства [7], эквивалентная мере Жаккара (связаны одной монотонно возрастающей зависимостью). Соответственно системе пробных площадок коэффициент Серенсена имеет вид $K_S = 2c/(a+b)$. Согласно теории множеств коэффициент Серенсена имеет вид

$$K_{0,-1} = \frac{2n(A \cap B)}{n(A)+n(B)}.$$

Для рассмотренных ранее путей $a \to b \to c \to d \to y$ и $a \to d \to b \to c \to y$ коэффициент Серенсена аналогичен мере Жаккара, что закономерно. Модифицированный коэффициент Серенсена для множеств переходов имеет следующий вид:

$$K_{0,-1} = \frac{2(n(A \cap B)+1)}{n(A)+n(B)+2}. \qquad (3)$$

Рассмотрим модифицированные коэффициенты Жаккара и Серенсена на следующем примере. Исследуем ту же тройку путей: $a \to b \to c \to d \to y$, $a \to b \to c \to e \to x$ и $a \to d \to b \to c \to y$, где a, b, c, d, e, x, y — классы классификации A. Построим матрицы попарных сравнений схожести для $K_{1,-1}$ по формуле (2) и $K_{0,-1}$ по формуле (3).

Попарное сравнение по формуле (2):

	$a \to b \to c \to d \to y$	$a \to b \to c \to e \to x$	$a \to d \to b \to c \to y$
$a \to b \to c \to d \to y$	1	3/7	1/4
$a \to b \to c \to e \to x$	3/7	1	1/4
$a \to d \to b \to c \to y$	1/4	1/4	1

Попарное сравнение по формуле (3):

	$a \to b \to c \to d \to y$	$a \to b \to c \to e \to x$	$a \to d \to b \to c \to y$
$a \to b \to c \to d \to y$	1	3/5	3/10
$a \to b \to c \to e \to x$	3/5	1	1/5
$a \to d \to b \to c \to y$	3/10	1/5	1

Легко предположить, что модифицированный коэффициент Серенсена более «оптимистичен» при расчете схожести, чем модифицированный коэффициент Жаккара: $K_{1,-1} \le K_{0,-1}$. Покажем справедливость сделанного предположения:

$$\frac{K_{1,-1}}{K_{0,-1}} = \frac{n(A \cap B)(n(A)+n(B))}{n(A \cup B)\,2n(A \cap B)} = \frac{(n(A)+n(B)}{2n(A \cup B)} \le 1.$$

Мера Кульчинского редко используется на практике [8], однако встречается в теоретических работах по таксономии [9]. Мера Кульчинского имеет вид

$$K_{0,1} = \frac{n(A \cap B)}{2}\left[\frac{1}{n(A)} + \frac{1}{n(B)}\right].$$

В литературе встречается и другое название — «второй коэффициент Кульчинского». Первый коэффициент Кульчинского имеет вид

$$K_{0,1} = \frac{n(A \cap B)}{n(A)+n(B)-2n(A \cap B)}.$$

Модифицированный коэффициент Кульчинского (в семантике множества переходов в классификации) имеет вид

$$K_{0,1} = \frac{n(A \cap B)+1}{2}\left[\frac{1}{n(A)+1}+\frac{1}{n(B)+1}\right].$$

Мера Шимкевича—Симпсона — бинарная мера сходства, независимо предложенная Шимкевичем [10] и Симпсоном [11], представляемая в виде

$$K_{0,+1} = \frac{n(A \cap B)}{\min(n(A), n(B))} = \max\left[\frac{n(A \cap B)}{n(A)}, \frac{n(A \cap B)}{n(B)}\right] =$$

$$= \frac{2n(A \cap B)}{n(A)+n(B)-|n(A)-n(B)|}, \qquad (4)$$

где $n(A)$ — мощность множества A; $n(B)$ — мощность множества B.

Меру Шимкевича—Симпсона, применимую к объектам, представленным множествами переходов от суперкласса в NBG, можно представить в виде

$$K_{0,+1} = \frac{n(A \cap B)+1}{\min(n(A), n(B))+1},$$

где множества A и B есть множества переходов.

Мера Браун—Бланке — бинарная мера сходства, формально весьма подобная мере Шимкевича—Симпсона, отличается лишь тем, что минимальная мощность двух множеств в знаменателе формулы (4) заменена максимальной мощностью [12]:

$$K_{0,-1} = \frac{n(A \cap B)}{\max(n(A), n(B))} = \min\left[\frac{n(A \cap B)}{n(A)}, \frac{n(A \cap B)}{n(B)}\right] =$$

$$= \frac{2n(A \cap B)}{n(A)+n(B)+|n(A)-n(B)|}.$$

Для множеств переходов в рамках классификации два класса классификации сходны на множестве путей переходов согласно мере Браун—Бланке:

$$K_{0,-1} = \frac{n(A \cap B)+1}{\max(n(A), n(B))+1}.$$

Коэффициент Отиаи — бинарная мера сходства [13], близкая к косинусной мере, задаваемая выражением $K = \frac{n(A \cap B)}{\sqrt{n(A)\, n(B)}}$, где $n(A)$ и $n(B)$ — мощности множеств A и B. Коэффициент Отиаи принято еще называть мерой Отиаи, коэффициентом Оцуки—Отиаи, коэффициентом Отиаи—Баркмана, геометрическим коэффициентом. В случае представления объектов множеством переходов коэффициент Отиаи имеет вид

$$K_{0,2} = \frac{n(A \cap B)+1}{\sqrt{(n(A)+1)(n(B)+1)}}.$$

В работе [14] рассматривается континиум эквивалентных мер

$$C(A,B) = \frac{2n(A \cap B)}{(1+\alpha)[(n(A)+n(B)-2\alpha n(A \cap B)]}, \qquad (5)$$

где $-1 < \alpha < \infty$. При $\alpha = 0$ выражение (5) сводится к мере Чекановского—Серенсена, а при $\alpha = 1$ оно численно сводится к мере Жаккара. К континиуму эквивалентных мер не могут быть отнесены мера Кульчинского, мера Шимкевича—Симпсона, мера Браун—Бланке и коэффициент Отиаи.

Представим континиум эквивалентных мер (5) в семантике множеств переходов в виде

$$K_{\infty} = \frac{2[n(A \cap B)+1]}{(1+\alpha)[(n(A)+n(B)+2-2\alpha\,[n(A \cap B)+1]]}$$

и составим таблицу мер. Введем следующую супер-позицию мер:

$$K = \beta_0 K_{\infty}(\alpha) + \beta_1 K_{0,1} + \beta_2 K_{0,+1} + \beta_3 K_{0,-1} + \beta_4 K_{0,2}, \qquad (6)$$

где $\sum_{i=0}^{4} \beta_i = 1$ и $0 \le \beta_i \le 1$. Супер-позиция мер K является линейной комбинацией мер, в которой β_i — весовой коэффициент. Как показано в [1, 5—14], все меры подобия в (6) могут быть заменены мерами отличия, которые являются строгими мерами:

$$\overline{K} = \beta_0(1-K_{\infty}(\alpha)) + \beta_1(1-K_{0,1}) + \beta_2(1-K_{0,+1}) + \beta_3(1-K_{0,-1}) + \beta_4(1-K_{0,2}) =$$

$$= 1 - [\beta_0 K_{\infty}(\alpha) + \beta_1 K_{0,1} + \beta_2 K_{0,+1} + \beta_3 K_{0,-1} + \beta_4 K_{0,2}], \qquad (7)$$

где $\sum_{i=1}^{4} \beta_i = 1$; $0 \le \beta_i \le 1$; $-1 < \alpha < \infty$.

Сравнивая меры Жаккара и Серенсена, можно сделать вывод об «оптимистичности» или «пессимистичности» одной меры по отношению к другой. Предложенная мера (6) в зависимости от α в $K_\infty(\alpha)$ и от весовых коэффициентов β_i может быть пессимистичной, нейтральной или оптимистичной, что позволяет сделать вывод о возможности ее «обучения».

В работе [1] предложена модель вычислений на многомерных классификациях с определенной степенью доверия:

$$\overline{O}_K(F)=\frac{1}{NK}\sum_{i=1}^{N}\overline{O}(F^i)p(F^i),\ p(F^{k+1})>(2L_{F^{k+1}}+1)\sum_{i=1}^{N}p(F^i),$$

где $F^1,...,F^i,...,F^N$, $i=\overline{1,N}$, — совокупность плоских классификаций (плоскостей деления), формирующих многомерную классификацию $F=F^1\times\times F^2\times...\times F^N$, $i=\overline{1,N}$; $p(F^i)$ — вес или коэффициент влияния плоскости деления F^i на меру отличия $\overline{O}_K(F)$; L_{F^k} — максимальный ранг плоской классификации F^k.

Обучаемую модель вычислений на классификациях можно получить в результате замены меры отличия на плоской классификации $\overline{O}_K(F^i)$ мерой отличия (7) в виде $\overline{O}_K(F^i,\alpha^i,{}^i)$, где $B^i=\{\beta_1^i,\beta_2^i,\beta_3^i,\beta_4^i\}$ и $\sum_{j=1}^{4}\beta_j^i=1$:

$$\overline{O}_K(F)=\frac{1}{NK}\sum_{i=1}^{k}\overline{O}_K(F^i,\alpha^i,{}^i)\,p(F^i),\ p(F^{k+1})>(2L_{F^{k+1}}+1)\sum_{i=1}^{k}p(F^i). \quad (8)$$

Сводная таблица мер

Мера	Обозначение	Формула
Континиум эквивалентных мер	$K_\infty(\alpha)$	$\frac{2[n(A\cap B)+1]}{(1+\alpha)[(n(A)+n(B)+2-2\alpha[n(A\cap B)+1]]}$, где $-1<\alpha<\infty$
Коэффициент Кульчинского	$K_{0,1}$	$\frac{n(A\cap B)+1}{2}\left[\frac{1}{n(A)+1}+\frac{1}{n(B)+1}\right]$
Мера Шимкевича—Симпсона	$K_{0,+1}$	$\frac{n(A\cap B)+1}{\min(n(A),n(B))+1}$
Мера Браун—Бланке	$K_{0,-1}$	$\frac{n(A\cap B)+1}{\max(n(A),n(B))+1}$
Коэффициент Отиаи	$K_{0,2}$	$\frac{n(A\cap B)+1}{\sqrt{(n(A)+1)(n(B)+1)}}$

Модель (8) является адаптируемой (обучаемой) параметрами α^i и B^i для каждой плоскости деления F^i. Поэтому суть обучения модели (8) сводится к нахождению значений матрицы размером $5 \times N$, в которой α^i и $B^i = \{\beta_1^i, \beta_2^i, \beta_3^i, \beta_4^i\}$ формируют строку из пяти значений для каждой плоскости делений из N.

Выводы

В предложенной обучаемой модели вычислений на пространственных классификациях мера отличия представлена конфигурируемой линейной комбинацией применимых на классификациях мер отличий. Разработанная модель позволяет строить адаптивные системы поддержки принятия решения, в частности в сфере кибербезопасности. К основным ее преимуществам следует отнести способность оперировать семантическими единицами, формируя предложения по оптимальному управлению с объяснением сути рекомендации.

СПИСОК ИСПОЛЬЗОВАНОЙ ЛИТЕРАТУРЫ

1. *Кравцов Г.А.* Модель вычислений на классификациях // Электрон. моделирование, 2016, **38**, № 1, с. 73—87.
2. *Дюран Б., Оделл П.* Кластерный анализ. Пер. с англ. Е.З. Демиденко. Под ред. А.Я. Боярского. М.: Статистика, 1977, 128 с.
3. *Семкин Б.И., Горшков М.В.* Аксиоматическое введение мер сходства, различия, совместности и зависимости для компонентов биоразнообразия // Вест. Тихоокеанского государственного экономического университета, 2008, №4, с. 31—46.
4. *Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р. и др.* Факторный, дискриминантный и кластерный анализ. Пер. с англ. Под ред. И.С. Енюкова. М.: Финансы и статистика, 1989, 215 с.
5. *Jaccard P.* Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines // Bulletin de la Societe Vaudoise des Seinces Naturelles, 1901, Vol. 37 (140), p. 241—272. — DOI : 10.5169/seals-266440.
6. *Levandowsky M., Winter D.* Distance between Sets // Nature, 1971, Vol. 234, pp. 34—35. — DOI : 10.1038/234034a0.
7. *Sörensen T.* A method of establishing groups of equal amplitude in plant sociology based on similarity of species content // Biologiske Skrifter, 1948, Vol. 5, № 4, p. 1—34.
8. *Kulczyński S.* Zespoly róslin w Pieninach (Die Pflanzenassociationen der Pienenen) // Bulletin International de L'Acad´emie Polonaise des Sciences et des Letters, Classe des Sciences Mathematiques et Naturelles. Serie B, Suppl´ement II, 2, 1927, p. 57—203.
9. *Sokal R.R., Sneath P.H.A.* Principles of numerical taxonomy. New York : W.H. Freeman & Co., 1963, 359 p.
10. *Szymkiewicz D.* Une contribution statistique a la géographie floristique // Acta Soc. Bot. Polon, 1934, Vol. 34, № 3, p. 249—265.
11. *Simpson G.G.* Holarctic mammalian faunas and continental relationship during the Cenozoic // Bull. Geol. Sci. America, 1947, Vol. 58, № 2, p. 613—688.
12. *Braun-Blanquet J.* Pflanzensoziologie Grundzüge der Vegetationskunde. Berlin : Springer-Verlag Wien, 1951, 632 p. — DOI : 10.1007/978-3-7091-4078-9.

13. *Ochiai A.* Zoogeographical studies on the soleoid fishes found Japan and its neighboring regions-II // Bull. Jap. Soc. sci. Fish, 1957, Vol. 22, № 9, p. 526—530. — DOI: 10.2331/suisan.22.526.
14. *Семкин Б.И.* Эквивалентность мер близости и иерархическая классификация многомерных данных // Иерархические классификационные построения в географической экологии и систематике, 1979, с. 97—112.

Получена 02.05.18

REFERENCES

1. Kravtsov, H.A. (2016), "Measure of difference between classifications", *Elektronnoe modelirovanie*, Vol. 38, no. 1, pp. 73-87.
2. Diuran, B. and Odell, P. (1977), *Klasternyi analiz* [Cluster analysis], Translated by E.Z. Demidenko, Statistika, Moscow, USSR.
3. Semkin, B.I. and Gorshkov, M.V. (2008), "The axiomatic introduction of similarity measures, differences measures, compatibility and dependencies for components of the biological variety", *Vestnik Tikhookeanskogo gosudarstvennogo ekonomicheskogo universiteta*, no. 4, pp. 31-46.
4. Kim, J.-O., Miuller, Ch.U., Klekka, U.R., et al. (1989), *Faktornyi diskriminantnyi i klasternyi analiz* [Factorial, discriminant and cluster analysis], Translated from English, Finansy i statistika, Moscow, USSR.
5. Jaccard, P. (1901), Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines, *Bulletin de la Societe Vaudoise des Sciences Naturelles*, Vol. 37 (140), pp. 241-272, DOI : 10.5169/seals-266440.
6. Levandowsky, M. and Winter, D. (1971), Distance between Sets, *Nature*, Vol. 234, pp. 34-35, DOI : 10.1038/234034a0.
7. Sörensen, T. (1948), A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, *Biologiske Skrifter*, Vol. 5, no. 4, pp. 1-34.
8. Kulczyński, S. (1927), Zespoly róslin w Pieninach (Die Pflanzenassociationen der Pienenen), *Bulletin International de L'Acad´emie Polonaise des Sciences et des Letters, Classe des Sciences Mathematiques et Naturelles, Serie B, Suppl´ement II*, *2*, pp. 57-203.
9. Sokal, R.R. and Sneath, P.H.A. (1963), Principles of numerical taxonomy, W.H. Freeman & Co., New York, USA.
10. Szymkiewicz, D. (1934), Une contribution statistique a la géographie floristique, *Acta Soc. Bot. Polon*, Vol. 34, no. 3, pp. 249-265.
11. Simpson, G.G. (1947), Holarctic mammalian faunas and continental relationship during the Cenozoic, *Bull. Geol. Sci. America*, Vol. 58, no. 2, pp. 613-688.
12. Braun-Blanquet, J. (1951), Pflanzensoziologie Grundzüge der Vegetationskunde, Springer-Verlag Wien, Berlin, Germany, DOI : 10.1007/978-3-7091-4078-9.
13. Ochiai, A. (1957), Zoogeographical studies on the soleoid fishes found Japan and its neighboring regions-II, *Bull. Jap. Soc. sci. Fish*, Vol. 22, no. 9, pp. 526-530, DOI : 10.2331/suisan.22.526.
14. Semkin, B.I. (1979), "The equality of similarity measures and hierarchical classification of multidimensional data", *Hierarchical structures built over classifications in the geographical ecology and systematics*, pp. 97-112.

Received 02.05.18

Г.О. Кравцов, В.І. Кошель, А.В. Долгоруков, В.В. Цуркан

МОДЕЛЬ, ЩО НАВЧАЄТЬСЯ, ОБЧИСЛЕНЬ НА КЛАСИФІКАЦІЯХ

Розглянуто класичне поняття міри відповідно до умов симетричності, рефлективності та нерівності трикутника. Висунуто вимогу до міри для її використання в теорії обчислень на класифікаціях. Встановлено обмеженість використання функцій відстані, коефіцієнта кореляції, косинусної міри подібності. Визначено та проаналізовано застосовність мір подібності, що використовуються на практиці. Це дозволило аргументувати необхідність введення нової міри. Наведено формальне визначення моделі обчислень на класифікаціях, яка навчається.

К л ю ч о в і с л о в а: міра подібності, міра відмінності, функція відстані, модель обчислень, модель, що навчається, континіум еквівалентних мір.

H.A. Kravtsov, V.I. Koshel, A.V. Dolgorukov, V.V. Tsurkan

TRAINABLE MODEL OF THE CALCULUS OVER CLASSIFICATIONS

The authors investigate the classical concept of measure in accordance with symmetry conditions, reflexivity and triangle inequality. The requirements to the measure have been formulated for its further use in the theory of the calculus over classification. Some features of the distance function, correlation coefficient, cosine measure of similarity are signification restrictions for applying them for the theory. All measures used in practice have been studied. The results of research have shown that new measure should be introduced that has been proposed. The authors have given formal definition of the trainable model of calculus over classification.

K e y w o r d s: similarity measure, difference measure, distance function, calculation model, model trainability, continuum of equivalent measures.

КРАВЦОВ Григорий Алексеевич, канд. техн. наук, докторант Ин-та проблем моделирования в энергетике им. Г.Е. Пухова НАН Украины. В 2000 г. окончил Севастопольский военно-морской ин-т им. П.С. Нахимова. Область научных исследований — кибербезопасность смарт-грид, криптография, программирование, разработка распределенных гетерогенных вычислительных систем.

КОШЕЛЬ Владимир Иванович, аспирант Ин-та проблем моделирования в энергетике им. Г.Е. Пухова НАН Украины. В 2002 г. окончил Харьковский национальный университет им. В.Н. Каразина. Область научных исследований — искусственный интеллект, интеллектуальный анализ данных, искусственные нейронные сети, обработка естественного языка.

ДОЛГОРУКОВ Александр Владимирович, аспирант Ин-та проблем моделирования в энергетике им. Г.Е. Пухова НАН Украины. В 2000 г. окончил Национальный технический университет Украины «Киевский политехнический ин-т». Область научных исследований — построение, внедрение, сопровождение и модернизация гетерогенных информационных систем.

ЦУРКАН Василий Васильевич, канд. техн. наук, доцент кафедры Национального технического университета Украины «Киевский политехнический ин-т имени Игоря Сикорского», который окончил в 2005 г. Область научных исследований — информационная безопасность, кибербезопасность и защита критической информационной инфраструктуры, теория рисков, социальная инженерия.