УДК 004.932

Г.А. Кравцов, канд. техн. наук,
Ин-т проблем моделирования в энергетике
им. Г.Е. Пухова НАН Украины
(Украина, 03164, Киев, Генерала Наумова, 15,
e-mail: hryhoriy.kravtsov@gmail.com)

Мера отличия классификаций

Методы определения изоморфизма графов неприменимы к определению изоморфизма классификаций, так как не учитывают неделимости некоторых классов классификации. В то же время, структурное подобие не является отражением семантики классификации, что весьма важно при определении меры отличия двух классификаций. Предложено понятие полной корректной классификации и введена дуальная мера отличия, отражающая известную проблему формы и содержания в философии.

Методи визначення ізоморфізму графів не можуть бути застосовані до класифікацій, тому що вони не враховують неподільність деяких класів класифікації. В той же час, структурна схожість не є відображенням семантики класифікації, що досить важливо при визначенні міри відмінності двох класифікацій. Запропоновано поняття повної коректної класифікації та введено дуальну міру відмінності, що є відображенням відомої філософської проблеми форми і змісту.

К л ю ч е в ы е с л о в а: классификация, полнота, корректность, мера, дуальность.

Постановка задачи. Актуальной задачей семантического поиска [1] является автоматическое нахождение близких по значению концептуальных понятий [2]. Если представить концептуальное понятие как некоторую классификацию, то указанная задача сводится к нахождению меры между двумя классификациями.

Будем использовать термины из работ [3—7]:

классификация — ориентированное дерево, отражающее систему мереологических или таксономических делений (МТД);

классифицирование — процесс построения классификации;

определение классовой принадлежности — отнесение объекта к некоторому классу построенной ранее классификации с точностью, не ниже заданной;

классифицированный объект — объект, относительно которого выполнена задача определения классовой принадлежности.

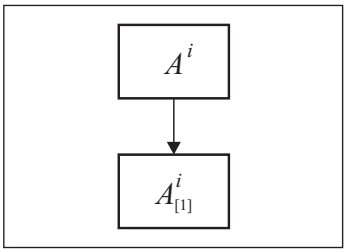


Рис. 1. Ориентированное дерево, не являющееся классификацией

Каждая классификация является ориентированным деревом [8], но не каждое ориентированное дерево является классификацией. Например, ориентированное дерево, представленное на рис. 1, не является классификацией.

Решение поставленной задачи представляется нетривиальным, так как требуется найти семантическую близость исходя из известной проблемы формы и содержания в философии [9]. Однако, для того чтобы гарантировать корректность нахождения некоторой меры между классификациями, следует потребовать корректности самих классификаций как исходных данных. Принципы корректного и эффективного классифицирования описаны Коротковым [10]. Несоблюдение их ведет к ряду ошибок при построении классификаций [3], а именно: неполное деление, деление с излишними членами, сбивчивое деление, скачек в делении.

Таким образом, задача нахождения меры близости или меры отличия между двумя классификациями является актуальной. При этом важна не только корректность классификаций, но и их однозначность, которая гарантируется порядком применения критериев деления. Однако, если классификации построены с применением одного и того же множества критериев разделения, то они являются синонимичными. В то же время, если предположить, что двум экспертам предложено выполнить классифицирование одного и того же концептуального понятия, нет уверенности, что оба эксперта для решения задачи сформируют равные множества критериев деления.

Формализм корректно построенной классификации. Классификация — это ориентированное дерево [8], в узлах которого находятся математические классы, семантически определяющие систему МТД. Это утверждение основано на описании математического класса с использованием понятия конгломерата классов [11].

В работе [12] выделены плоские и пространственные классификации. Следует заметить, что плоская классификация есть планарное ориентированное дерево с семантикой деления. Согласно [12] A_I^i означает некоторый класс в плоскости деления классификации i, имеющий путь уточнения I, который однозначно определяет путь в графе от самого общего класса классификации до некоторого уточнения A_I^i.

Под корректно построенной плоской классификацией будем понимать такую систему МТД, при которой для двух классов классификации, A_I^i и A_Y^i,

таких, что для относительных расстояний $R(A_I^i, A_I^i \cdot A_Y^i)$ и $R(A_Y^i, A_I^i \cdot A_Y^i)$, где относительное расстояние означает число переходов между классами классификации [12], такими, что

$$R(A_I^i, A_I^i \cdot A_Y^i) = R(A_Y^i, A_I^i \cdot A_Y^i) = 1, \quad (1)$$

где $A_I^i \cdot A_Y^i$ — ассоциативная бинарная операция обобщения классов классификации, а $\overline{Q}(A_I^i, A_Y^i) = 2/3$ — теоретическое значение меры отличия двух классов, уточняющих один и тот же класс,

$$\overline{Q}(A_I^i, A_Y^i) = 1 - \frac{R(A, A_I^i \cdot A_Y^i) + 1}{R(A, A_I^i \cdot A_Y^i) + R(A_I^i, A_I^i \cdot A_Y^i) + R(A_Y^i, A_I^i \cdot A_Y^i) + 1},$$

выполняется следующая система равенств:

$$\overline{Q}(A_I^i, A_I^i \cdot A_Y^i) = \overline{Q}(A_Y^i, A_I^i \cdot A_Y^i),$$

$$\overline{Q}(A_I^i, A_Y^i) = 2/3. \quad (2)$$

Система уравнений (2) и требование (1) (согласно которому классы A_I^i и A_Y^i имеют один и тот же ранг $K+1$) следует понимать так: если у произвольного класса $A_I^i \cdot A_Y^i$ ранга K существует менее двух уточняющих классов A_I^i и A_Y^i ранга $K + 1$, то мера отличия между классами одной классификации [12] на плоскости деления (измерении) $\overline{Q}(A_I^i, A_Y^i)$ есть величина постоянная, равная 2/3, независимо от выбора A_I^i и A_Y^i, если $I \neq Y$.

Отсюда вытекает следующее требование: у любого класса, являющегося вершиной ориентированного дерева, — классификации, должно быть не менее двух уточняющих классов, ранг которых на единицу больше ранга произвольно выбранного класса, или не должно быть ни одного. Именно поэтому не каждое произвольно ориентированное дерево является классификацией.

Интуитивно понятно, что при делении класса на уточняющие классы можно допустить ошибку: $E(A_I^i, A_Y^i) = 2/3 - \overline{O}_o(A_Y^i, A_I^i)$, где $\overline{O}_o(A_Y^i, A_I^i)$ — наблюдаемая мера отличия между классами одной классификации на плоскости деления (измерении), а $I \neq Y$. Случай, когда $I = Y$, является исключением, так как ошибка $E(A_I^i, A_Y^i)$ есть линейная функция от относительного расстояния $R(A_I^i, A_Y^i)$, которое при $I = Y$ равно нулю [12]. Поэтому ошибка может возникнуть только при наблюдении меры между двумя разными классами. В таком случае определим ошибку при наблю-

дении меры отличия между двумя разными классами как разность между теоретической мерой отличия, равной 2/3, и наблюдаемой мерой в виде

$$\overline{Q}\,(A_I^i, A_I^i \cdot A_Y^i) = \overline{Q}\,(A_Y^i, A_I^i \cdot A_Y^i) = 1,$$
$$E(A_I^i, A_Y^i) = 2/3 - \overline{O}_{\text{o}}(A_Y^i, A_I^i), \tag{3}$$

где $\overline{O}_{\text{o}}(A_Y^i, A_I^i)$ — наблюдаемое значение меры отличия между двумя классами A_I^i и A_Y^i; $E(A_I^i, A_Y^i)$ — ошибка наблюдения меры отличия, такая, что $-0{,}5 \le E(A_Y^i, A_I^i) \le 2/3$, если $I \ne Y$, и $E(A_Y^i, A_I^i) = 0$, если $I = Y$.

Очевидно, что если некоторый класс A^i делится на N классов $A_{[j]}^i$, где $j = \overline{1, N}$, то существует N^2 неуникальных значений $E(A_{[j]}^i, A_{[k]}^i)$, где $k = \overline{1, N}$. Поскольку $\overline{O}_{\text{o}}(A_{[j]}^i, A_{[k]}^i) = \overline{O}_{\text{o}}(A_{[k]}^i, A_{[j]}^i)$ [12], то согласно (3) $E(A_{[j]}^i, A_{[k]}^i) = E(A_{[k]}^i, A_{[j]}^i)$. Теоретически число уникальных значений $E(A_{[j]}^i, A_{[k]}^i)$ равно $(N^2 - N)/2$. Среднее абсолютное значение ошибки наблюдения меры отличия составит

$$\overline{E}(A^i) = \frac{1}{N^2 - N} \sum_{j=1}^{N} \sum_{k=1}^{N} |E(A_{[j]}^i, A_{[k]}^i)|, \tag{4}$$

где $|E(A_{[j]}^i, A_{[k]}^i)|$ — абсолютное значение ошибки. Очевидно, что если мера отличия $\overline{O}_{\text{o}}(A_Y^i, A_I^i) = 2/3$ для любых двух классов, удовлетворяющих (3) и таких, что $I \ne Y$, то $\overline{E}(A^i) = 0$. По сути, величина (4) представляет собой интегральный показатель согласованности (усредненную ошибку) деления класса ранга K на кассы ранга $K + 1$ и удовлетворяет условию $0 \le \overline{E}(A_I^i) \le 2/3$.

Рассмотрим в качестве примера некоторую плоскую классификацию (рис. 2, *а*). Заменим обозначения классов интегральным показателем согласованности деления $\overline{E}(A^i)$ класса ранга K на классы ранга $K + 1$ (рис. 2, *б*). Очевидно, что согласно (4) интегральный показатель согласованности деления $\overline{E}(A^i)$ в листьях деревьев равен нулю. Количественный показатель согласованности (consistency) плоской классификации определим как

$$C(A^i) = \sum_I \overline{E}\,(A_I^i)^{R(A_I^i,\, A^i)+1)^M}, \tag{5}$$

где I — множество известных путей уточнения классификации A^i; M — среднее число уточняющих классов в узлах ориентированного дерева, не являющихся листьями (параметр полноты). Выражение (5) семантически есть сумма действительных значений ошибок $\overline{E}\,(A^i)$ на интервале

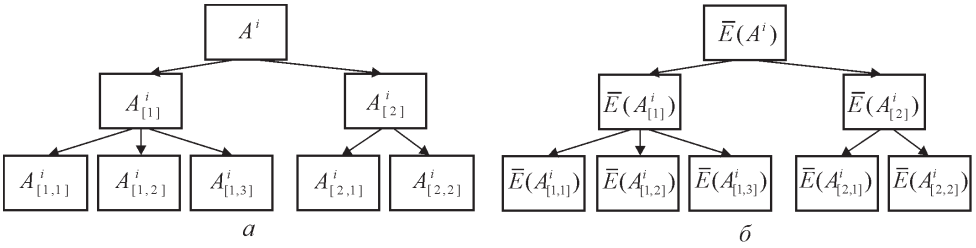


Рис. 2. Плоская классификация (*а*) и распределение интегрального показателя согласованности (*б*)

$0 \le \overline{E}(A_I^i) \le 2/3$, возведенных в степень, отражающую удаленность класса от корня дерева классификации, при условии, что ошибка для корневого класса имеет значение 1. Это не уменьшает абсолютного значения ошибки при понимании факта, что любая большая единицы положительная степень вещественного числа в указанном диапазоне меньше исходного значения ошибки.

Заменив на рис. 2, *а*, обозначения классов ранга K числами, равными количеству уточняющих классов ранга $K + 1$, получим представление классификации в виде количества уточняющих классов (рис. 3). Очевидно, что на рис. 3 нулями обозначены листья дерева. Для такой классификации параметр полноты следующий:

$$M = \frac{2+3+2}{3} = 2(3). \qquad (6)$$

Как видим, для полных ориентированных деревьев M есть арность (для полного бинарного ориентированного дерева $M = 2$).

Семантически (5) есть интегральная оценка внутренней согласованности плоской классификации A^i, применяемой при оценке относительной компетентности в обсуждаемом вопросе членов группы для принятия групповых решений [13], когда обсуждаемым вопросом является некоторая классификация.

Введем понятие вертикально полной классификации, означающее такую классификацию, представимую ориентированным деревом, у которой классы, представляющие собой листья ориентированного дерева, семантически неделимы по определению (т.е. класс не может быть разделен на подклассы с семантикой).

Горизонтально полной классификацией назовем классификацию, у которой каждый обобщающий класс поделен на уточняющие классы так, что уточняющие классы покрывают все возможные варианты деления, и между двумя различными уточняющими классами мера отличия удовлетворяет системе (2).

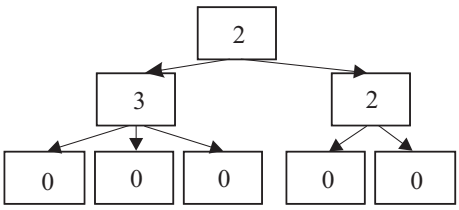


Рис. 3. Представление классификации (см. рис. 2, *а*) количеством уточняющих классов

Полной корректной классификацией назовем такую, которая является горизонтально полной и вертикально полной одновременно.

Из определения полной корректной классификации следует, что каждый уточняющий класс, имеющий уточняющие классы, является полной корректной классификацией. Это позволяет определить метод нисходящей декомпозиции: любая полная корректная классификация может быть разложена на составляющие исключительно перечислением полных корректных классификаций уточняющих классов.

Рассмотрим классификацию *A* (см. рис. 2, *а*) и классификации *B* и *C* (рис. 4). Представим классификацию *A* в виде декомпозиции. Согласно данному выше определению получим две полные классификации (рис. 5, *а*). В результате декомпозиции полных корректных классификаций *B* и *C* получим полные классификации *B* и *C* (рис. 5, *б*, *в*).

Описанный способ декомпозиции имеет принципиальное значение при определении структурного отличия двух классификаций.

Изоморфность классификаций. Структурную меру отличия определим как изоморфизм классификаций. Напомним, что в теории графов изоморфизмом графов $G = \langle V_G, E_G \rangle$ и $H = \langle V_H, E_H \rangle$ называется биекция между множествами вершин графов $f : V_G \to V_H$ такая, что любые две вершины, *u* и *v*, графа *G* смежны тогда и только тогда, когда вершины $f(u)$ и $f(v)$ смежны в графе *H*. Здесь графы считаем неориентированными и не имеющими весов вершин и ребер. В случае, если понятие изоморфизма применяется к ориентированным или взвешенным графам, накладываются дополнительные ограничения на сохранение ориентации дуг и значений весов.

Изоморфизму графов посвящено множество работ. Например, в [14] предложен подход, позволяющий проводить поиск изоморфных пересечений двух графов за полиномиальное время (не более чем $O(n^4)$), основываясь на одном из двух известных подходов — поиске максимального изоморфного пересечения и минимального количества редактирований (Minimal Edit Distance, MED). Несмотря на различие этих подходов согласно результатам работы [15] меры подобия, полученные этими способами, зависят друг от друга линейно. Так, в методах MED вводятся операторы редактирования графа, такие как удаление, добавление, замещение вершин и др. Существует набор операторов, которые преобразуют один

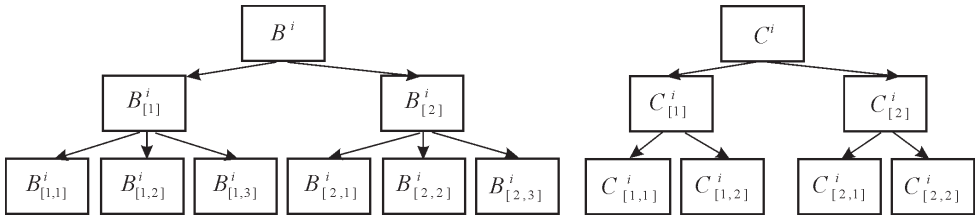


Рис. 4. Классификации *B* и *C*

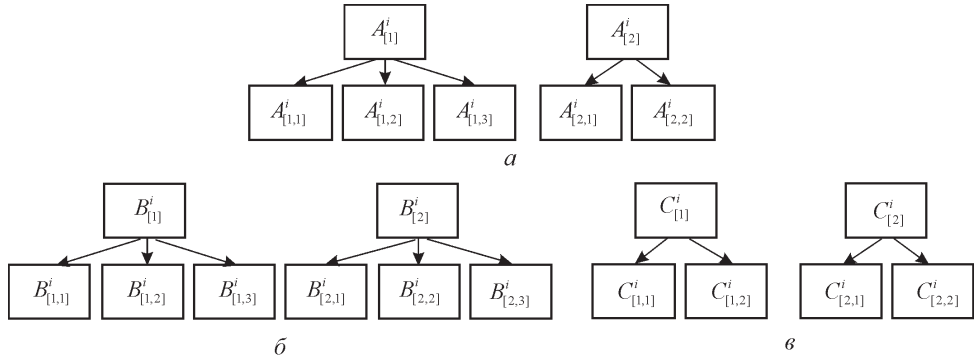


Рис. 5. Декомпозиции классификаций *A*, *B* и *C*

граф в другой. Минимальный из таких наборов определяет меру подобия между графами.

В методах, основанных на поиске максимального изоморфного пересечения, для двух графов проводится поиск пар подграфов, принадлежащих этим двум графам, которые являются изоморфными друг другу и определяют изоморфное пересечение. Изоморфное пересечение, имеющее максимальное число вершин, определяет меру подобия между графами. Вообще говоря, задача поиска максимального изоморфного пересечения является *NP*-полной и не решается за полиномиальное время. Однако в настоящее время разработано множество методов, позволяющих путем введения некоторых ограничений и дополнительных условий решать данную задачу за полиномиальное время [16]. Предложенные в [14, 16] и в аналогичных работах подходы не могут быть использованы для классификаций, так как в них при определении любых двух изоморфных подграфов игнорируется требование неделимости листьев ориентированного дерева классификации.

Принято использовать кодирование графов для рассмотрения их свойств [17], в частности смыслового подобия, в метаэвристике [18], для компактного хранения и поиска в базах данных. Код Прюфера [17], поз-

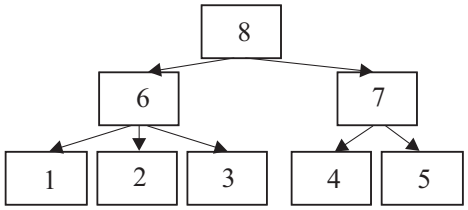


Рис. 6. Вариант разметки классификации, представленной на рис. 2

воляющий записать дерево n-го порядка (т.е. с пронумерованными вершинами) последовательностью $n-2$ его вершин, оказывается бесполезным для кодирования ориентированного дерева, так как не позволяет отразить ориентацию ребер при произвольной нумерации вершин. Напомним, что кодирование Прюфера переводит помеченные деревья порядка n в последовательность чисел от единицы до n по такому алгоритму: пока количество вершин больше двух, выбирается лист v с минимальным номером и в код Прюфера добавляется номер вершины, смежной с v, а вершина v и инцидентное ей ребро удаляются из дерева. Однако, если предположить, что существует некоторый строгий алгоритм нумерации вершин классификации, то код Прюфера может оказаться полезным.

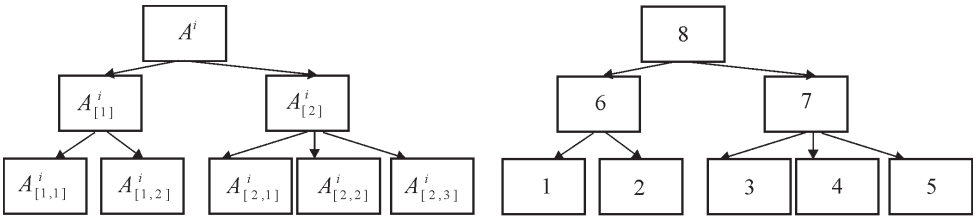
Положим, что выбор принципов разметки не ограничен и рассмотрим следующие принципы.

Восходящий ранговый принцип разметки. Любая вершина классификации ранга K имеет больший номер разметки, чем любая вершина ранга $K+1$.

Согласно восходящему ранговому принципу разметки одним из возможных вариантов разметки классификации, изображенной на рис. 2, *а*, может быть вариант, представленный на рис. 6. Несложно показать, что код Прюфера для ориентированного дерева (см. рис. 5, *а*) будет выглядеть так: 6, 6, 6, 7, 7, 8. Как видим, при восходящем ранговом принципе разметки код Прюфера указывает на корень дерева 8. Более короткая запись может быть представлена в виде $6^3\,7^2\,8^1$, где корень дерева есть 8.

Очевидно, что у 7 существует два уточняющих класса, а у 6 — три. Но у 8 в коде значится только один уточняющий класс, что не соответствует действительности. Данный факт — следствие из алгоритма формирования кода Прюфера. Чтобы ответить на вопрос, сколько всего уточняющих классов первого ранга у классификации, достаточно показатель степени при корне увеличить на единицу. Для удобства будем использовать запись $6^3\,7^2\,8^2$, которую назовем модифицированным кодом Прюфера.

Рассмотрим следующую классификацию, изоморфную классификации, представленной на рис. 2, *а*. Представим вариант разметки классификации (рис. 7, *а*) согласно восходящему принципу (рис. 7, *б*). Запишем код Прюфера для классификации, изображенной на рис. 7, *б*, 6, 6, 7, 7, 7, 8 и представим его в виде модифицированного кода: $6^2\,7^3\,8^2$.



а *б*

Рис. 7. Изоморфная классификация (относительно классификации, показанной на рис. 2) (*а*) и ее разметка (*б*)

Итак, для двух изоморфных классификаций (см. рис. 2, *а*, и рис. 6) получены модифицированные коды соответственно $6^3\ 7^2\ 8^2$ и $6^2\ 7^3\ 8^2$. Очевидно, что решающую роль играет число уточняющих классов классификаций, что можно рассматривать как критический параметр алгоритма нумерации вершин классификации. Из рис. 3 видно, что предложенный принцип разметки обеспечивает разметку, представленную на рис. 6.

Восходящий рангово-арный принцип разметки. Любая вершина классификации ранга K имеет больший номер разметки, чем любая вершина ранга $K+1$, при условии, что для любых двух вершин ранга K вершина с бóльшим числом уточняющих классов имеет меньший номер разметки и ее уточняющие классы имеют меньшие номера разметки, чем уточняющие классы класса с меньшим числом уточняющих классов. Очевидно, что восходящая рангово-арная разметка есть вариант восходящей ранговой разметки. Для классификации, приведенной на рис. 2, *а*, восходящая рангово-арная разметка показана на рис. 7, *б*.

Выполним рангово-арную разметку для изоморфной классификации, приведенной на рис. 7, *а*. Запишем код Прюфера для классификации с разметкой 6, 6, 6, 7, 7, 8 и представим его в виде модифицированного кода: $6^3\ 7^2\ 8^2$ (рис. 8). Полученный код совпадает с кодом при восходящей рангово-арной разметке изоморфной классификации на рис. 2, *а*. Отсюда следует: если модифицированные коды Прюфера ориентированных деревьев, размеченных согласно восходящему рангово-арному принципу, совпадают, то ориентированные деревья изоморфны.

Нисходящий ранговый принцип разметки. Любая вершина классификации ранга K имеет меньший номер разметки, чем любая вершина ранга $K+1$. Представим вариант разметки классификации, представленный на рис. 2, *а*, согласно нисходящему ранговому принципу (рис. 9, *а*).

Запишем код Прюфера для классификации с разметкой 2, 2, 3, 3, 3, 1 и представим его в виде модифицированного кода $2^2\ 3^3\ 1^2$. Следует обратить внимание, что 1 — корень ориентированного дерева. Если по коду $6^2\ 7^3\ 8^2$ видно, что в классификации, которая представляет собой корень, восемь

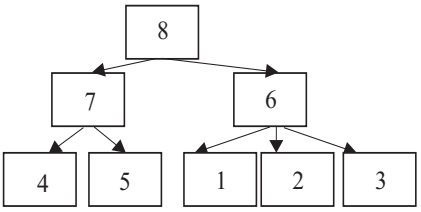


Рис. 8. Разметка изоморфной классификации 6, 6, 6, 7, 7, 8

классов по 8, то как понять, что код $2^2\,3^3\,1^2$ также представляет классификацию из восьми классов? Это просто: если к сумме степеней модифицированного кода добавить единицу, то получим число классов классификации (вершин ориентированного дерева). В табл. 1 приведены некоторые свойства модифицированного кода Прюфера для классификации, представленной на рис. 2, *а*.

Поскольку сформулированный нисходящий ранговый принцип разметки не позволяет судить об изоморфности классификаций, уточним его.

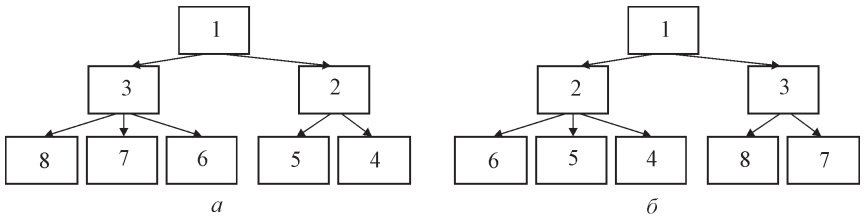
Нисходящий рангово-арный принцип разметки. Любая вершина классификации ранга K имеет меньший номер разметки, чем любая вершина ранга $K+1$, при условии, что для любых двух вершин ранга K вершина с бóльшим числом уточняющих классов имеет меньший номер разметки и ее уточняющие классы имеют меньшие номера разметки, чем уточняющие классы класса с меньшим числом уточняющих классов.

Представим вариант разметки классификации (см. рис. 2, *а*) согласно нисходящему рангово-арному принципу (рис. 9, *б*). Запишем код Прюфера для классификации с разметкой 3, 3, 2, 2, 2, 1 и представим его в виде модифицированного кода $3^2\,2^3\,1^2$. Для одной и той же классификации получим следующие коды: восходящий — $6^2\,7^3\,8^2$, нисходящий — $3^2\,2^3\,1^2$.

Очевидно, что порядок следования степеней модифицированного кода Прюфера не зависит от выбора направления рангово-арного принципа разметки (восходящий или нисходящий), т.е. для подтверждения изоморфности двух классификаций достаточно получить степени модифицированного кода Прюфера. Записи 2, 3, 2, которую назовем рангово-арным кодом, достаточно, чтобы восстановить структуру классификации.

Из табл. 2 видно, что ориентированное дерево, как математическую формализацию классификации, можно закодировать меньшим количеством символов, чем это определено теоремой Кэли [17], согласно которой на n вершинах, пронумерованных числами от 1 до n, существует ровно n^{n-2} различных деревьев, и эти деревья могут быть закодированы последовательностью из $n-2$ чисел. При использовании рангово-арных принципов разметки длина кодирующей последовательности равна числу классов, которые не являются листьями ориентированного дерева.

Нисходящий рангово-арный принцип разметки имеет следующие преимущества перед восходящим рангово-арным принципом. Если в классификацию добавляется или удаляется вершина, при использовании восходящего рангово-арного принципа необходимо переразметить все остальные вершины. При использовании нисходящего рангово-арного принципа



а *б*

Рис. 9. Разметка согласно нисходящим ранговому (*а*) и рангово-арному (*б*) принципам

Таблица 1. **Свойства модифицированного кода Прюфера**

Свойство классификации	Значение	Алгоритм получения значения из кода	
		восходящего $6^2\,7^3\,8^2$	нисходящего $2^2\,3^3\,1^2$
Число классов	8	Сумма степеней кода, увеличенная на единицу: $2 + 3 + 2 + 1 = 8$	
Число листьев	5	Число классов минус число уникальных оснований в коде: $8 - 3 = 5$	
Максимальный ранг класса в классификации	2	Ранг корня классификации, который записан как самое правое число, равен нулю [12]. Степень корня показывает, сколько элементов кода имеет ранг, на единицу больший, чем ранг корня, т.е. 1. Таких чисел два. Последним числом, имеющим ранг 1, есть последнее число в модифицированном коде. Поскольку в коде не отражены уточняющие классы, максимальный ранг равен 2	
Параметр полноты M	2.(3)	$\frac{2+3+2}{3} = 2.(3)$	

Таблица 2. **Алгоритм восстановления дерева по рангово-арному принципу разметки для кода 2, 3, 2**

Шаг восстановления	Рангово-арный принцип разметки	
	Восходящий	Нисходящий
Количество классов в классификации	$2 + 3 + 2 + 1 = 8$	$2 + 3 + 2 + 1 = 8$
Модифицированный код Прюфера	$6^2\,7^3\,8^2$	$3^2\,2^3\,1^2$
Короткий код Прюфера	$6^2\,7^3\,8^1$	$3^2\,2^3\,1^1$
Полный код Прюфера	6, 6, 7, 7, 7, 8	3, 3, 2, 2, 2, 1
Графическое представление дерева с разметкой	Рис. 7, *б*	Рис. 9
Исходная классификация (ориентированное дерево)	Рис. 2, *а*	Рис. 2, *а*

это не потребуется, так как разметка корня дерева в этом случае никогда не меняется. Структура данных, представляющая дерево, передается указателем на корень, а рекурсивная разметка, начинающаяся с корня, является приоритетной. Поэтому использование нисходящего рангово-арного принципа разметки становится целесообразным.

Рангово-арная структура и проверка корректности рангово-арного кода. Рангово-арная структура данных отражает структуру классификации, размеченной согласно рангово-арному принципу и представленной модифицированным кодом Прюфера. Алгоритм проверки построения кода 2, 3, 2 следующий (код читается справа налево).

1. Записываем первое значение степени, соответствующее нулевой арности:

Ранг	Показатель степени	Число классов		
		Ожидаемое	Реальное	Ошибка
0	2	1	1	0

2. Сумма показателей степеней свидетельствует о том, что следует ожидать два показателя степени, релевантных первому рангу:

Ранг	Показатель степени	Число классов		
		Ожидаемое	Реальное	Ошибка
0	2	1	1	0
1	2, 3	2	2	0

3. Поскольку выбраны все показатели степеней из кода 2, 3, 2, построение рангово-арной структуры данных завершено.

Следует обратить внимание на то, что число ожидаемых классов нулевого ранга всегда равно единице по определению, так как указывает на корень дерева. Каждое последующее ожидаемое число содержит сумму показателей степеней предыдущей строки. На этом основан подход к проверке корректности рангово-арного кода.

Рассмотрим код 2, 3, 3 в виде рангово-арной структуры:

Ранг	Показатель степени	Число классов		
		Ожидаемое	Реальное	Ошибка
0	3	1	1	0
1	2, 3	3	2	1

Предложенная структура данных позволяет определить, является ли некоторая последовательность чисел рангово-арным кодом.

Представим A, B и C в виде кортежа модифицированных кодов Прюфера в комбинации с рангово-арной разметкой (табл. 3). Воспользовавшись идеей меры Жаккара [19], определим структурную меру подобия кортежей $\bar{I}(X, Z)$, равную величине единица минус отношение числа одинаковых кодов в кортежах к сумме одинаковых (неуникальных) и отличных (уникальных) кодов в кортежах, где X, Z — некоторые полные корректные классификации.

Представленная на рис. 10 мера отличия для классификаций D и F свидетельствует о том, что классы, делимые на разное число уточняющих классов, являются абсолютно различными, а система деления таких классов — различными классификациями. Классификации, представленные кодами (2, 3, 4, 3) и (2, 3, 2) с учетом декомпозиции $\bar{I}((2, 3, 4, 3), (2, 3, 2)) =$

Таблица 3. **Кортежи модифицированных кодов Прюфера**

Классификация	Декомпозиция классификации	Изображение
A	(2,3,2), (3), (2)	Рис. 1 и 3
B	(3,3,2), (3), (3)	Рис. 2, *а* и 4, *а*
C	(2,2,2), (2), (2)	Рис. 2, *б* и 4, *б*

Таблица 4. **Декомпозиция модифицированных кодов Прюфера**

Модифицированный код Прюфера	Декомпозиция классификации
(2,3,4,3)	(2,3,4,3), (2), (3), (4)
(2,3,2)	(2,3,2), (2), (3)

Таблица 5. **Алгоритм определения вложения одной классификации в другую**

Шаг	Классификация	
	A (рис. 2, *а*)	*D* (рис. 10)
Модифицированный код Прюфера	2,3,2	3
Кортеж модифицированных кодов Прюфера, определяющий декомпозицию классификации	(2, 3, 2), (3), (2)	(3)
Определение кортежа минимальной длины	Длина кортежа: 3 (кортеж максимальной длины)	Длина кортежа: 1 (кортеж минимальной длины)
Самый длинный код кортежа	(2, 3, 2)	(3)
Содержит ли кортеж максимальной длины наиболее длинный код из кортежа минимальной длины	Да: (3)	Да: (3)

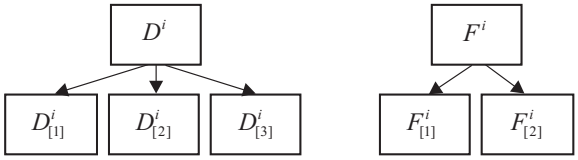


Рис. 10. Мера $\bar{I}(D,F)=1$ для классификаций D и F

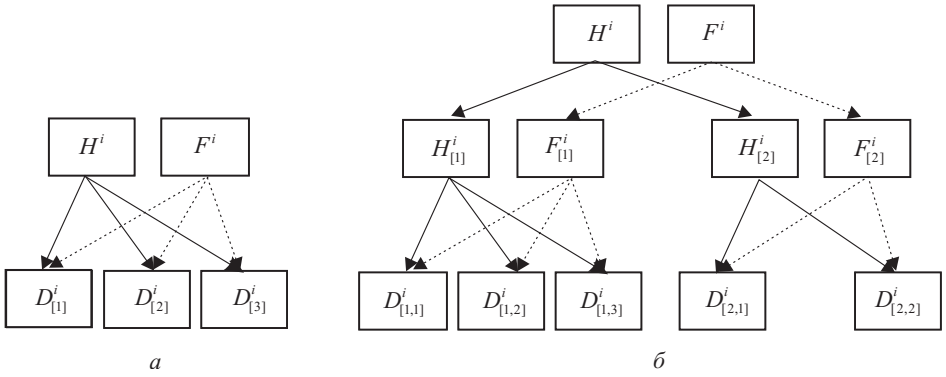


Рис. 11. Классификации H и F с одинаковой структурой и одинаковыми неделимыми классами

$=1-\frac{2}{5}=0{,}6$, свидетельствуют о том, что уточнение неделимых классов классификаций D и F позволяет значительно уменьшить меру отличия двух классификаций (табл. 4).

Согласно описанному поведению меры структурного отличия для классификаций, закодированных последовательностями $(k,k,...,k,m)$ и $(n,n,...,n,m)$, где $n,k\geq 2$, $k\neq n$ и $k,n,m\in N$, мера структурного отличия имеет вид $\bar{I}((k,k,...,k,m),(n,n,...,n,m))=1$. Тогда несложно определить, является ли одна классификация классом другой по кортежу модифицированных кодов Прюфера. Для этого достаточно выполнить шаги, указанные в табл. 5.

Следовательно, для того чтобы определить, содержит ли одна классификация другую, достаточно ответить на вопрос: содержит ли кортеж модифицированных кодов Прюфера максимальной длины наиболее длинный код из кортежа минимальной длины.

Синонимичные классификации и дуализм меры отличия. Сформулируем следующую гипотезу: если у двух классификаций совпадает структура и все классы, кроме класса нулевого ранга, то классы нулевого ранга этих классификаций семантически есть синонимы. Напомним, что синонимами в биологической таксономии являются два или более названия, относящиеся к одному и тому же биологическому таксону [4].

Если данная гипотеза верна для классификаций, представленных на рис. 11, *а*, то она верна и для классификаций, представленных на рис. 11, *б*, что можно объяснить, как представление множества элементов в виде множества единичных множеств элементов исходного множества: $\{a,b,c,d\}\rightarrow$ $\rightarrow\{\{a\},\{b\},\{c\},\{d\}\}$. Отсюда следует вывод: две классификации являются синонимичными, если у них одинаковые неделимые классы.

Это позволяет ввести еще одну меру на классификациях — меру несинонимичности, которая эквивалентна множественной мере Жаккара [19] для множеств неделимых классов двух классификаций:

$$\bar{S}(A,B)=1-\frac{n(L(A)\bigcap L(B))}{n(L(A)\bigcup L(B))},$$

где $n(L(A))$ — мощность множества неделимых классов классификации A; $L(A)$ — множество неделимых классов классификации A; $n(L(B))$ — мощность множества неделимых классов классификации B; $L(B)$ — множество неделимых классов классификации B.

Таким образом, для двух классификаций, A и B , могут быть определены две меры — мера структурного отличия $\bar{I}(A,B)$ и мера несинонимичности $\bar{S}(A,B)$, где мера $\bar{I}(A,B)$ отвечает за форму, а мера $\bar{S}(A,B)$ — за содержание. Дуальная мера отличия классификаций, не являясь строго математической мерой, есть двойка вида $M(A,B)=(\bar{I}(A,B),\ \bar{S}(A,B))$, отражающая известную проблему формы и содержания в философии [9].

Выводы

Решение актуальной задачи семантического поиска сводится к нахождению меры отличия между двумя классификациями, которая имеет дуальную природу и предложена как двойка мер структурного отличия и несинонимичности.

СПИСОК ЛИТЕРАТУРЫ

1. *Басипов А.А., Демич О.В.* Семантический поиск: проблемы и технологии // Вест. Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. — 2012. — № 1. — С. 104—111.
2. *Болдачев А.* Хабрахабр. Концептуальное описание индивидов. — [Электронный ресурс]. — Режим доступа: https://habrahabr.ru/post/276271/ . — Дата доступа: май, 2016.
3. *Ивлев Ю.В.* Логика // ТК «Велби». — М. : изд-во «Проспект», 2008. — 304 с.
4. *Шаталкин А.И.* Таксономия. Основания, принципы и правила. — М. : Товарищество научных изданий КМК, 2012. — 600 с.
5. *Ушаков Д.Н.* Толковый словарь русского языка. — М. : Альта-Принт, 2005. — 1216 с.
6. *Вьюгин В.* Математические основы теории машинного обучения и прогнозирования. — М. : МЦНМО, 2013. — 387 с.

7. *Голдблатт Р.* Топосы. Категорный анализ логики. — М. : Мир, 1983. — 488 с.
8. *Альфс Берзтисс* Структуры данных. — М. : Статистика, 1974. — 408 с.
9. *Карелина Е.В.* Теоретическая строгость как соответствие системы и метода в философии. — Красноярск: Сибирский федеральный университет, 2012. — 120 с.
10. *Коротков Э.М.* Исследование систем управления. — М. : ДеКА, 2004. — 336 с.
11. *Adamek J., Herrlich H., Strecker G.E.* Abstract and Concrete Categories. The Joy of Cats. Availabe: http://katmat.math.uni-bremen.de/acc/acc.pdf. [Aceess: February of 2016]
12. *Кравцов Г.А.* Модель вычислений на классификациях // Электрон. моделирование.— 2016. — **38**, № 1. — С. 73—87.
13. *Тоценко В.Г.* Методы и системы поддержки принятия решений. Алгоритмический аспект. — Киев: Наук. думка. — 2002. — 382 с.
14. *Агарков А.В.* Метод сравнения двух графов за полиномиальное время // Искусственный интеллект. — 2003. — № 4. — С. 172—184.
15. *Bunke H.* On a relation between graph edit distance and maximum common subgraph // Pattern Recognition Letters. — 1997. — Vol. 18. — P. 689—694.
16. *Messmer B.T., Bunke H.* Subgraph Isomorphism in Polynomial Time. — University of Bern, Institut fur Informatik und angewandte Mathematik. — [On-line] Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.43.870&rep=rep1&type=pdf. [Accessed: March, 2016]
17. *Касьянов В.Н., Евстигнеев В.А.* Графы в программировании: обработка, визуализация и применение — Санкт-Петербург: БХВ-Петербург. — 2003. — 1104 с.
18. *Люк Ш.* Основы метаэвристик [Электронный ресурс]. — Режим доступа: http://qai.narod.ru/GA/metaheuristics.html . — Дата доступа: май 2016.
19. *Семкин Б.И., Горшков М.В.* Аксиоматическое введение мер сходства, различия, совместимости и зависимости для компонентов биоразнообразия // Изв. Дальневосточного федерального университета. Экономика и управление. — 2008. — № 4. — С. 31—46.

H.A. Kravtsov

MEASURE OF DIFFERENCE BETWEEN CLASSIFICATIONS

Methods for determining graph isomorphism are not applicable to determining isomorphism of classifications because of ignoring the indivisibility of certain classes of classification. At the same time structural similarity does not reflect the classification semantics that is very important in determining the measure of difference between the two classifications. The author proposes the concept of total correct classification and introduces the dual measures the difference reflecting the known problem of form and content in philosophy.

K e y w o r d s: classification, completeness, correctness, measure, duality.

REFERENCES

1. Basipov, A.A. and Demich, O.V. (2012), “Semantic search: problems and technology”, *Vestnik Astrakhanskogo gosudarstvennogo universiteta, Seria: Upravlenie, vychislitelnaya tekhnika i informatika*, no. 1, pp. 104-111.
2. Boldachev, A. “Habrahabr: Conceptual description of individuals”, available at: https://habrahabr.ru/post/276271/ (assessed May, 2016).
3. Ivlev, Yu.V. (2008), *Logika* [Logic], TK Velbi, Prospekt, Moscow, Russia.
4. Shatalkin, A.I. (2012), *Taksonomiya. Osnovaniya, printsipy i pravila* [Taxonomy. Grounds, principles, and rules], Tovarishchestvo nauchnykh izdaniy KMK, Moscow, Russia.

5. Ushakov, D.N. (2005), *Tolkovyi slovar russkogo yazyka* [Explanatory dictionary of the Russian language], Alta-Print, Moscow, Russia.
6. Vyugin, V. (2013), *Matematicheskie osnovy teorii mashinnogo obucheniya i prognozirovaniya* [Mathematical grounds of the theory of computer teaching and prediction], MTsNMO, Moscow, Russia.
7. Goldblatt, R. (1983), *Toposy. Kategornyi analiz logiki* [Toposes. Category analysis of logic], Mir, Moscow Russia.
8. Alfs Berztiss (1974), *Struktura dannykh* [Data structure], Statistika, Moscow, Russia.
9. Karelina, E.V. (2012), *Teoreticheskaya strogost kak sootvetstvie sistemy i metoda v filosofii* [Theoretical strictness as correspondence of the system and method in philosophy], Siberian Federal University, Krasnoyarsk, Russia.
10. Korotkov, E.M. (2004), *Issledovanie system upravleniya* [Study of control systems], DeKA, Moscow, Russia.
11. Adamek, J., Herrlich, H. and Strecker, G.E. "Abstract and concrete categories. The Joy of Cats", available at: http://katmat.math.uni-bremen.de/acc/acc.pdf. (accessed February, 2016).
12. Kravtsov, H.A. (2016), "Model of computations on classifications", *Elektronnoe modelirovanie*, Vol. 38, no. 1, pp.73-87.
13. Totsenko, V.G. (2002), *Metody i sistemy podderzhki prinyatiya resheniy. Algoritmicheskiy aspect* [Models and systems for decision making support], Naukova dumka, Kiev, Ukraine.
14. Agarkov, A.V. (2003), "Method of comparison of two graphs for polynomial time", *Iskusstvennyi intellect*, no. 4, pp.172-184.
15. Bunke, H. (1997), "On a relation between graph edit distance and maximum common subgraph", *Pattern Recognition Letters,* Vol. 18, pp. 689-694.
16. Messmer, B.T. and Bunke, H. "Subgraph Isomorphism in Polynomial Time", University of Bern, Institut fur Informatik und angewandte Mathematik, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.43.870&rep=rep1&type=pdf. (accessed March, 2016).
17. Kasyanov, V.N. and Evstigneyev, V.A. (2003), *Grafy v programmirovanii: obrabotka, vizualizatsiya, primenenie* [Graphs in programming: processing, visualization and application], BHV-Petersburg, St.-Petersburg, Russia.
18. Lyuk, S. (2013), "Essentials of metaheuristics", available at: http://qai.narod.ru/GA/metaheuristics.html (accessed May, 2016).
19. Semkin, B.I. and Gorshkov, M.V. (2008), "Axiomatic introduction of the measures of similarity, difference, compatibility and dependence for biodiversity components", *Izvestiya Dalnevostochnogo federalnogo universiteta. Ekonomika i upravlenie*, no. 4, pp. 31-46.

Поступила 25.02.16;
после доработки 26.05.16

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 — математическое моделирование, кибербезопасность смарт-грид, криптография, разработка распределенных гетерогенных вычислительных систем.