S©MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru

Том 12, стр. 1006–1031 (2015) УДК 519.7
DOI 10.17377/semi.2015.12.087 MSC 68T37

ФОРМАЛИЗАЦИЯ «ЕСТЕСТВЕННОЙ» КЛАССИФИКАЦИИ И СИСТЕМАТИКИ ЧЕРЕЗ НЕПОДВИЖНЫЕ ТОЧКИ ПРЕДСКАЗАНИЙ

Е.Е ВИТЯЕВ, В.В МАРТЫНОВИЧ

Abstract. Nowadays there exist many approaches to classification and clustering; for instance one can mention those based on compactness and various metrics on feature spaces, based on etalons, on distributions composition partitioning, etc. In contrast to these approaches, the task of "natural" classification is to discover a classification as a law of nature that satisfy some requirements promoted by naturalists. The sense of this law is in the compression of information by extracting the structure of natural objects. We propose a formalization of this law based on fix–points of probabilistic laws of special type. We prove that the probabilistic laws we define solve the problem of statistical ambiguity and thus they enable us to predict without contradictions and to provide consistent fix–points. These fix–points form a "natural" classification. Finally we present the results of a computer experiment on building and recognition of classes of transcription factors binding sites.

Keywords: natural classification, clustering, fix–points, formal notion, building of notions, notions.

1. Введение

В рамках «классификационного движения» Забродиным В. Ю. были систематизированы критерии "естественности" классификации, которые выдвигались различными естествоиспытателями [5]. Эти критерии позволяют понять

Vityaev, E.E., Martinovich, V.V., Formalization of «natural» classification and systematics as fix-points ofpredictions.

Работа поддержана РФФИ (грант 15-07-03410-а).

Поступила 21 июля 2015 г., опубликована 24 декабря 2015 г.

в каком смысле естествоиспытатели понимали «естественную» классификацию как закон.

(1) Забродин В. Ю. [5]: “Естественной” является та, и только та классификация, которая выражает закон природы”.
(2) Критерий Е. С. Смирнова [9, с. 413]: “Таксономическая проблема заключается в “индикации”: от бесконечно большого числа признаков нам нужно перейти к ограниченному их количеству, которое заменило бы все остальные признаки”.
(3) Рутковский Л.: ”Чем в большем числе существенных признаков сходны сравниваемые предметы, тем вероятнее их одинаковость и в других отношениях”.
(4) Критерий В. Уэвель: “Чем больше общих утверждений об объектах дает возможность сделать классификация, тем она естественней”.
(5) Критерий А. А. Любищева [5]: “Наиболее совершенной системой является такая, где все признаки объекта определяются его положением в системе. Чем ближе система стоит к этому идеалу, тем она менее искусственна, и естественной следует называть такую, где количество свойств объекта, поставленных в функциональную связь с его положением в системе, является максимальным (в идеале это все его свойства)”.
(6) В работе Шрейдера С. А. [10]: “В многообразии объектов, образующих “естественную” классификацию, можно обнаружить два типа закономерностей:
    (a) соотношения, связывающие “короткое” описание архетипа, достаточное для диагностирования принадлежности объекта к данному классу, с “полным” описанием. В сущности, это законы, позволяющие на основании принадлежности объекта некоторому естественному классу прогнозировать все его свойства;
    (b) правила, показывающие, как деформируются свойства объектов при переходе к смежным классам. Именно они гарантируют возможность переноса знаний с одного объекта на все принадлежащие данному классу и, несколько сложнее, на объекты смежных классов”.
(7) В наших работах [1-3] был выдвинут следующий принцип построения “естественных” классификаций: “Разбиение на классы должно производиться так, чтобы объекты одного класса подчинялись одним и тем же закономерностям, объекты разных классов подчинялись разным группам закономерностей. Объекты одного класса, кроме того, должны обладать некоторой целостностью. Целостность определим как взаимную согласованность закономерностей каждой группы по предсказанию различных свойств объектов”.

2. Определение естественной классификации и систематики

Приведем формальные определения «естественной» классификации и систематики, позволяющие объяснить, приведенные критерии “естественности” классификации.

Будем предполагать, что исследуемая предметная область задана ***эмпирической системой*** $M = \langle A, W\rangle$ — конечной моделью сигнатуры

$$\Im = \langle \mathrm{P}_1, \ldots, \mathrm{P}_k\rangle,$$

где $A = \{a_1, a_2, \ldots, a_m\}$ — конечное множество классифицируемых объектов; $W = \{P_1, \ldots, P_k\}$ – кортеж предикатов сигнатуры $\Im$, определенных на A вида $(x_i(a) = y_j^i)$, означающих, что признак x_i принимает на объекте a значение y_j^i; $x_1, x_2, ..., x_N$ – признаки, характеристики и величины объектов из A, принимающие значения $I_1, \ldots, I_N, y_j^i \in I_i$; для каждого признака x_i и его значения $y_j^i \in I_i$, существует предикат из W; $\mathrm{P}_1, \ldots, \mathrm{P}_k$ — символы предикатов из W.

Определение 1. *Определим* ***закономерную модель*** $\mathrm{M}_a = \langle \Omega_\mathrm{a}, \mathrm{Z}_\mathrm{a}\rangle$ *объекта* a*, где* Ω_a *— множество предикатов из* W*, выполнимых на объекте* a*,* Z_a *— множество закономерностей вида*

$$\mathrm{Z}_\mathrm{a} = \{(x_{i_1}(u) = y_{j1}^{i_1})\&(x_{i_2}(u) = y_{j2}^{i_2})\&...\&(x_{i_n}(u) = y_{jn}^{i_n}) \Rightarrow (x_{i_0}(u) = y_{j_0}^{i_0})\},$$

выполненных на объекте a*.*

Рассмотрим некоторый класс $\mathfrak{C}$ объектов.

Определение 2. *Определим* ***закономерную модель класса*** $\mathrm{M}_\mathfrak{C} = \langle \Omega_\mathfrak{C}, \mathrm{Z}_\mathfrak{C}\rangle$ *как пересечение всех закономерных моделей объектов класса* $\mathfrak{C}$*. Здесь* $\Omega_\mathfrak{c}$ *— множество значений признаков, которые одинаковы для всех объектов класса* $\mathfrak{C}$*,* $\mathrm{Z}_\mathfrak{C} = \bigcap\limits_{a\in\mathfrak{C}} \mathrm{Z}_\mathfrak{a}$.

Проанализируем критерий Е. С. Смирнова [9]. Разнообразие классов всегда несопоставимо меньше разнообразия комбинаций значений признаков и, следовательно, между значениями признаков должно существовать большое число закономерных связей. Если, например, число классов 64, а признаки бинарные, то независимыми среди них могут быть только 6 признаков, т. к. $2^6 = 64$. При классификации животных, растений, почв и т.д. естествоиспытатели могут использовать огромное, потенциально бесконечное, множество признаков и характеристик. Но среди них только 6 признаков могут быть независимыми, а остальные признаки связаны между собой закономерностями так, что из 6-и признаков предсказываются значения всех остальных признаков. Найти признаки, из которых предсказываются все остальные признаки, и составляет проблему «индикации» [9]. Такими значениями признаков в закономерной модели класса $M_\mathfrak{C}$ являются *порождающие совокупности значений признаков.*

Определение 3. *Набор значений признаков* $\langle y_{j_1}^{i_1}, y_{j_2}^{i_2}, \ldots, y_{j_t}^{i_t}\rangle$ *является* ***порождающим*** *в закономерной модели класса* $\mathrm{M}_\mathfrak{C} = \langle \Omega_\mathfrak{C}, \mathrm{Z}_\mathfrak{C}\rangle$*, если по этим значениям признаков и закономерностям из* $\mathrm{Z}_\mathfrak{C}$ *можно вывести все остальные значения признаков из* $\Omega_\mathfrak{C}$*.*

Набор значений порождающих признаков определяется неоднозначно.

Рассмотрим задачу построения *систематики.* Рассмотрим в качестве примера таблицу 1. В ней множество объектов $A = \{a_1, a_2, \ldots, a_9\}$ разбито на 4 класса, описываемых 30-ю признаками $\{x_1, x_2, \ldots, , x_{30}\}$. Предположим, что классы $\mathfrak{C}_1, \ldots, \mathfrak{C}_4$ нам известны, и мы знаем закономерные модели этих классов. Задача построения систематики состоит в том, что бы найти такое подмножество признаков $\{x_{i_1}, x_{i_2}, \ldots, x_{i_n}\} \subset \{x_1, x_2, ..., x_{30}\}$, что бы для каждого

Класс	x_1	$\mathbf{x_2}$	x_7	$\mathbf{x_8}$	x_{10}	$\mathbf{x_{11}}$	x_{14}	$\mathbf{x_{15}}$	x_{20}	$\mathbf{x_{21}}$	x_{27}	$\mathbf{x_{28}}$	x_{30}
$\mathfrak{C}_1$	a_1	$y^2_{j_1}$		$y^8_{j_1}$		$y^{11}_{j_1}$		$y^{15}_{j_1}$		$y^{21}_{j_1}$		$y^{28}_{j_1}$	
	a_2	$y^2_{j_1}$		$y^8_{j_1}$		$y^{11}_{j_1}$		$y^{15}_{j_1}$		$y^{21}_{j_1}$		$y^{28}_{j_1}$	
$\mathfrak{C}_2$	a_3	$y^2_{j_2}$		$y^8_{j_2}$		$y^{11}_{j_2}$		$y^{15}_{j_2}$		$y^{21}_{j_2}$		$y^{28}_{j_2}$	
	a_4	$y^2_{j_2}$		$y^8_{j_2}$		$y^{11}_{j_2}$		$y^{15}_{j_2}$		$y^{21}_{j_2}$		$y^{28}_{j_2}$	
	a_5	$y^2_{j_2}$		$y^8_{j_2}$		$y^{11}_{j_2}$		$y^{15}_{j_2}$		$y^{21}_{j_2}$		$y^{28}_{j_2}$	
$\mathfrak{C}_3$	a_6	$y^2_{j_3}$		$y^8_{j_3}$		$y^{11}_{j_3}$		$y^{15}_{j_3}$		$y^{21}_{j_3}$		$y^{28}_{j_3}$	
	a_7	$y^2_{j_3}$		$y^8_{j_3}$		$y^{11}_{j_3}$		$y^{15}_{j_3}$		$y^{21}_{j_3}$		$y^{28}_{j_3}$	
$\mathfrak{C}_4$	a_8	$y^2_{j_4}$		$y^8_{j_4}$		$y^{11}_{j_4}$		$y^{15}_{j_4}$		$y^{21}_{j_4}$		$y^{28}_{j_4}$	
	a_9	$y^2_{j_4}$		$y^8_{j_4}$		$y^{11}_{j_4}$		$y^{15}_{j_4}$		$y^{21}_{j_4}$		$y^{28}_{j_4}$	

Таблица 1. Построение систематики классов

класса $\mathfrak{C}_1, \ldots, \mathfrak{C}_4$ набор значений этих признаков являлся порождающим. Эти признаки $x_2, x_8, x_{11}, x_{15}, x_{21}, x_{28}$ в таблице выделены серым цветом.

Определение 4. *Набор признаков $S = \langle x_{i_1}, x_{i_2}, ..., x_{i_n} \rangle$ будем называть **системообразующим** для классов $\{\mathfrak{C}_\mathrm{i}\}_{i \in I}$, если для каждого класса наборы значений этих признаков $\langle y^{i_1}_{j_1}, y^{i_2}_{j_2}, ..., y^{i_n}_{j_n} \rangle$ различны и являются порождающими совокупностями этих классов.*

В этом случае каждый класс будет однозначно определяться набором значений системообразующих признаков. Понятно, что наборы системообразующих признаков также определяются неоднозначно. Задача построения *систематики* и состоит в том, что бы найти наиболее компактный и информативный набор системообразующих признаков.

Систематика состоит в том, чтобы представить некоторым образом, например, таблицей, изменение значений системообразующих признаков при переходе от объектов одного класса к объектам другого класса. Значения остальных признаков объектов класса будут предсказываться по значениям системообразующих признаков данного класса. Изменение значений системообразующих признаков может удовлетворять некоторому закону, вследствие чего систематику можно представить некоторым специальным образом, например таблицей, чтобы этот закон был виден наглядно.

Определение 5. *Определим **закономерную модель систематики** как $M_S = \langle S, Z_S \rangle$, где S — набор системообразующих признаков, а Z_S — **закон систематики** — закон изменения значений системообразующих признаков из S при переходе от класса к классу. Этот закон может иметь достаточно разнообразный вид, как, например, таблица 2 или таблица Менделеева, но в нем должно быть отражено разнообразие различных значений системообразующих признаков для разных классов систематики. Каждому набору значений системообразующих признаков S соответствует некоторый класс $\mathfrak{C}$ и закономерная модель класса $\mathrm{M}_\mathfrak{C} = \langle \Omega_\mathfrak{C}, \mathrm{Z}_\mathfrak{C} \rangle$. Закон систематики Z_S является метазаконом по отношению к закономерностям классов $\mathrm{Z}_\mathfrak{C}$.*

Классы	$\mathbf{x}_2$	$\mathbf{x}_8$	$\mathbf{x}_{11}$	$\mathbf{x}_{15}$	$\mathbf{x}_{21}$	$\mathbf{x}_{28}$
Класс 1	$\langle y^2_{j_1}$	$y^8_{j_1}$	$y^{11}_{j_1}$	$y^{15}_{j_1}$	$y^{21}_{j_1}$	$y^{28}_{j_1}\rangle$
Класс 2	$\langle\ y^2_{j_2}$	$y^8_{j_2}$	$y^{11}_{j_2}$	$y^{15}_{j_2}$	$y^{21}_{j_2}$	$y^{28}_{j_2}\rangle$
Класс 3	$\langle\ y^2_{j_3}$	$y^8_{j_3}$	$y^{11}_{j_3}$	$y^{15}_{j_3}$	$y^{21}_{j_3}$	$y^{28}_{j_3}\rangle$
Класс 4	$\langle\ y^2_{j_4}$	$y^8_{j_4}$	$y^{11}_{j_4}$	$y^{15}_{j_4}$	$y^{21}_{j_4}$	$y^{28}_{j_4}\rangle$

Таблица 2. Закон систематики для таблицы 1

Закон систематики Z_S связан с законами классов, как это определено в определении С. А. Шрейдера [10]. Закономерностями первого типа в его определении являются закономерности классов $\mathrm{Z}_{\mathfrak{C}}$, а закономерностями второго типа – закон систематики Z_S.

Закон систематики для нашего примера можно представить таблицей 2. В ней представлены наборы порождающих совокупностей для всех классов. Например, для класса 1 этот набор имеет вид $\langle y^2_{j_2}, y^8_{j_8}, y^{11}_{j_{11}}, y^{15}_{j_{15}}, y^{21}_{j_{21}}, y^{28}_{j_{28}}\rangle$.

Рассмотрим критерий А. А. Любищева [5]. Системой по Любищеву является такое представление классификации объектов, когда по месту объекта в системе определяются все его признаки. В нашем определении значения признаков объектов определяются взаимодействием двух законов:

(1) закона систематики Z_S, используя который мы по положению объекта в системе (таблице) можем определить класс объекта и значения системообразующих признаков;
(2) по закономерностям класса $\mathrm{Z}_{\mathfrak{C}}$ и значениям системообразующих признаков мы можем определить все остальные признаки объекта.

Определение 6. *Определим* ***систематику*** *как набор* $\Sigma = \langle S, Z_S, \{Z_{\mathfrak{C}_i}\}_{i\in I}\rangle$.

Задача построения систематики состоит в том, чтобы выбрать наиболее совершенную систему, объясняющую свойства и строение объектов простейшим образом. Несмотря на субъективность выбора систематики, она является законом природы, потому что из неё можно предсказать все остальные свойства объектов. Таблица 2 по сравнению с таблицей 1 сжимает информацию практически без потерь, поскольку по закону систематики Z_S и закономерностям из $\{Z_{\mathfrak{C}_i}\}_{i\in I}$ мы можем восстановить всю таблицу 1.

Все предыдущие рассуждения проводились в предположении того, что классы нам известны и по ним можно найти порождающие совокупности признаков. В реальных задачах разбиение объектов на классы естествоиспытателю неизвестно. Как тогда строить систематику? Задача построения систематики состоит в этом случае в нахождении такого разбиения множества объектов на классы, что бы построенная на этих классах систематика была наиболее совершенной и простой. Такая систематика называется «естественной».

Однако анализ всех возможных разбиений множества объектов на классы и построение по ним систематики — вычислительно неприемлемая задача, поскольку всевозможных разбиений множества N объектов на классы 2^N.

Поэтому строить «естественную» классификацию надо отправляясь не от объектов, а от признаков и закономерностей, которым они удовлетворяют. Поэтому рассмотрим наше определение «естественной» классификации (номер 7).

В соответствии с этим определением надо найти такие группы закономерностей, которые согласованны по предсказанию и описывают разные классы объектов. Это можно сделать опираясь не на объекты, а на признаки, находя такие наборы значений признаков $\Omega_{\mathfrak{C}}$, которые взаимно предсказываются множеством закономерностей $\mathrm{Z}_{\mathfrak{C}}$. Формально эта задача сводится к обнаружению неподвижных точек взаимных предсказаний признаков по закономерностям, обнаруженным на всем множестве объектов. Приведем соответствующие определения.

3. Теория предметной области. Классы как неподвижные точки логических выводов в рамках теории

Вернемся к предметной области — *эмпирической системе* $M = \langle A,\ W\rangle$ сигнатуры $\Im = \langle \mathrm{P}_1, \ldots, \mathrm{P}_k\rangle$. Зафиксируем язык первого порядка L сигнатуры $\Im$, содержащей только одноместные предикаты. Пусть X — множество *переменных* a,b,...,c. Тогда высказывания вида $\mathrm{P}_i(a),\ i = 1, \ldots, k$, будем называть *атомарными формулами*. Обозначим через $\mathrm{U}(\Im)$ множество атомных формул. Атомарные формулы или их отрицания будем называть *литерами*, а множество всех литер обозначим через $\mathrm{Lit} = \mathrm{L}(\Im)$. Множеством *формул* $\Re(\Im)$ назовем замыкание всех литер относительно логических операций $\&, \vee, \neg, \Rightarrow$. *Теорией предметной области* $\mathrm{Th}(M)$ назовем множество всех предложений из $\Re(\Im)$ истинных на M. Предикаты из $W = \langle P_1, \ldots, P_k\rangle$ определим через признаки $P_i(a) = (x_i(a) = y^i_j),\ a \in A, i = 1, \ldots, N$.

Далее будем предполагать, что теория $\mathrm{Th}(M)$ — совокупность универсальных формул. Известно, что любая совокупность универсальных формул логически эквивалентна совокупности универсальных замыканий правил вида

$$C = (\mathrm{A}_1\&...\&\mathrm{A}_\mathrm{k} \Rightarrow \mathrm{A}_0),\ k \geq 0,\ \mathrm{A}_0 \notin \{\mathrm{A}_1, ..., \mathrm{A}_\mathrm{k}\}, \tag{1}$$

где $\mathrm{A}_0, \mathrm{A}_1, \ldots, \mathrm{A}_\mathrm{k}$ — литеры. Формулы вида $(\Rightarrow A_0)$ рассматриваются как правила $(\mathrm{T} \Rightarrow \mathrm{A}_0)$, где T — истина. В записи правил кванторная приставка всегда будет опускаться. Поэтому, не умаляя общности можно считать, что теория $\mathrm{Th}(M)$ есть множество правил вида (1).

Правило может быть истинным на эмпирической системе M только потому, что посылка правила всегда ложна. Также правило может быть истинно потому, что некоторое его логически более сильное «подправило», содержащее часть посылки, истинно на эмпирической системе. Эти наблюдения суммированы в следующей теореме.

Теорема 1 ([21]). *Правило* $\mathrm{C} = (\mathrm{A}_1\&...\&\mathrm{A}_\mathrm{k} \Rightarrow \mathrm{A}_0)$ *выводимо в исчислении высказываний из каждого из правил:*

1. $(\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h} \Rightarrow \neg\mathrm{A}_{i_0}) \vdash \mathrm{C},\ \{\mathrm{A}_{i_1}, ..., \mathrm{A}_{i_h}, \mathrm{A}_{i_0}\} \subset \{A_1, ..., A_k\},\ 0 \leq h < k;$

2. $(\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h} \Rightarrow \mathrm{A}_0) \vdash \mathrm{C},\ \{\mathrm{A}_{i_1}, ..., \mathrm{A}_{i_h}\} \subset \{A_1, ..., A_k\},\ 0 \leq h < k.$

Определение 7. *Подправилом правила C будем называть любое логически более сильное правило вида* 1 *или* 2*, определенные в теореме* 1*.*

Следствие 1. *Если подправило правила C истинно на M, тогда и правило C истинно на M.*

Определение 8. *Законом эмпирической системы M будем называть любое истинное на M правило C, каждое подправило которого уже не истинно на*

M. Правило ($\Rightarrow$ A_0) *истинно на M, если* $M \models A_0$. *Истинное на M правило* ($\Rightarrow$ A_0) *является законом на M.*

Пусть Law — множество всех законов на M. Тогда из Law логически следует теория $\mathrm{Th}(M)$.

Теорема 2 ([21])**.** *Law* $\vdash$ *Th(M) и для любого правила* $C \in Th(M)$ *существует его подправило являющееся законом на M.*

Доказательство. Правило $C \in Th(M)$ либо является законом и принадлежит Law, либо для него существует подправило, истинное на M. Возьмем это подправило, тогда снова оно либо является законом, либо для него есть подправило истинное на M и т. д. Получим закон, являющийся подправилом C. Тогда, в силу теоремы 1, оно выводимо из этого закона. □

Определим неподвижные точки логического вывода по законам из $\mathrm{Th}(M)$. Поскольку в сигнатуре нет функциональных символов и многоместных предикатов, то любой объект $b \in A$ порождает подмодель $B_b = \langle b, W\rangle$ модели $M = \langle A, W\rangle$, будем писать $B_b \subset M$.

Лемма 1. *Множество литер* $\mathrm{S}_b = \{\mathrm{L} \in \mathrm{Lit} \mid B_b \models L\}$*, истинных на подмодели* B_b*, замкнуто относительно логического следования по законам из Law.*

Доказательство. Пусть посылка H некоторого закона $\mathrm{C} = (\mathrm{A}_1 \& ... \& \mathrm{A}_\mathrm{k} \Rightarrow \mathrm{A}_0)$ содержится в S_b, то есть $B_b \models A_1, ..., B_b \models A_k$. Поскольку $M \models C$, то в силу свойств подмодели и универсальности высказывания C (содержит только кванторы всеобщности) получаем $B_b \models C$ и, значит, $B_b \models A_0$. Поэтому $\mathrm{A}_0 \in \mathrm{S}_b$ и S_b оказывается замкнутым относительно выводов по законам из Law □

Теорема 3. *Множество литер* N *замкнуто относительно логического следования по правилам из Law тогда и только тогда, когда* $N = \bigcap\limits_{\mathrm{S}_b \in \chi} \mathrm{S}_b$*, где* $\chi = \{S_b \mid N \subseteq S_b\} \neq \emptyset$.

Доказательство. Положим $X = \bigcap\limits_{\chi} S_b$. Пусть N замкнуто относительно логического следования по законам из Law. Включение $N \subseteq X$ очевидно. Рассмотрим произвольную литеру $x \in X$, тогда $\forall S \in \chi (x \in S)$. Покажем, что импликация $I = (\& N \Rightarrow x)$ истина на M и её посылка не является ложной на M: $M \not\models \neg(\& N)$ ввиду $\chi \neq \emptyset$ (здесь $\& N$ — конъюнкция литер из N). Пусть для некоторого объекта $b \in A, B_b \models N$. Тогда по построению $N \subseteq S_b$, $S_b \in \chi$ и $x \in S_b$. Тогда импликация I истинна на M и, в силу теоремы 2, существует закон $Q \in Law$ являющийся подправилом этой импликации, имеющий вид 2 теоремы 1 (поскольку $M \not\models \neg(\& N)$). Но так как множество N замкнуто относительно логического вывода по правилам из Law, то заключение закона Q также принадлежит N, т.е. $x \in N$ и, следовательно, $X \subseteq N$.

Обратно, пусть $N = \bigcap\limits_{S_b \in \chi} S_b$. Обозначим замыкание N с помощью логического вывода как $cl\,(N)$. Включение $N \subseteq cl\,(N)$ очевидно. Пусть $y \in cl\,(N)$, тогда в Law существует конечная последовательность импликаций, начиная с некоторой импликации вида $\& L \Rightarrow x$, $L \subset N$, которая приводит к y. Из $N = \bigcap\limits_{S_b \in \chi} S_b$ и леммы 1 следует, что $x \in S_b$, $\forall S_b \in \chi$, откуда следует, что $x \in \bigcap\limits_{\chi} S_b = N$. Тогда возьмем следующую импликацию, приводящую к y, и т. д. □

Определение 9. *Множество литер* N*, удовлетворяющее условиям теоремы 3 (замкнутое относительно логического следования по правилам из Law), будем называть неподвижной точкой оператора логического вывода по правилам из Law и обозначать* $N \uparrow \mathrm{Law}$*. Множество всех таких неподвижных точек обозначим через Cl(M).*

Переопределим основные понятия «естественной» классификации и систематики в терминах неподвижных точек.

Определение 10.
(1) *Множество всех неподвижных точек* $Cl(M)$ *будем называть множеством всех детерминированных классов эмпирической системы M.*
(2) *Каждый класс N выделяет в эмпирической системе M множество объектов принадлежащих классу* $M(N) = \{b \in A \mid N \subset \mathrm{S}_b\}$.
(3) *Закономерной моделью детерминированного класса* $C = \langle N, \mathrm{Z}_N \rangle$ *является множество литер* $N \in \mathrm{Cl}(M)$ *и множество* $\mathrm{Z}_N \subset Law$ *всех законов, в состав которых входят только литеры из N.*
(4) *Порождающим множеством некоторого класса* $N \uparrow \mathrm{Law}$ *будем называть такое подмножество литер* $L \subset N$*, что* $L \uparrow Law = N \uparrow Law$.
(5) *Набор S атомарных высказываний* $\mathrm{P}_j(a)$*, j=1,... s,* $s \leq k$ *будем называть системообразующим, если для каждого класса из Cl(M) есть порождающее множество литер, которые включают только атомарные высказывания из системообразующего набора.*
(6) *Систематику определим как набор* $\Sigma = \langle S, Z_S, \{Z_{N_i}\}_{N_i \in Cl} \rangle$*, где S — системообразующий набор атомарных высказываний,* Z_S *– закон систематики, определяющий порядок взятия отрицаний атомарных высказываний из S,* $\{Z_{N_i}\}_{N_i \in Cl}$ *– множества правил неподвижных точек из Cl(M).*

4. «Естественная» классификация и систематика для цифр индекса

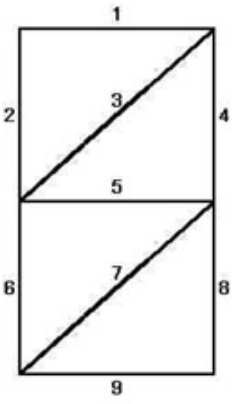


Рис. 1. Признаки.

Рассмотрим 10 цифр индекса 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Занумеруем признаки индексов как показано на рис. 1. Определим предикаты $P_1, \ldots, P_9$ означающие наличие (истина) или отсутствие (ложь) i-го элемента в начертании цифры. Тогда цифры индекса можно представить таблицей 1. Рассмотрим *эмпирическую систему* $M = \langle A, W \rangle$, где $W = \langle P_1, \ldots, P_9 \rangle$ – множество предикатов сигнатуры $\Omega = \{P_1, \ldots, P_9\}$, определенных на $A = \{0, 1, \ldots, 9\}$. Будем рассматривать цифры как классы. Найдем теорию $\mathrm{Th}(M)$ этой эмпирической системы в виде множества всех правил, истинных на всех цифрах индекса, а также все неподвижные точки этой теории. Для этого воспользуемся программой [18]. В результате получим 3743 правила и 10 неподвижных точек соответствующих цифрам индекса.

Далее для каждой цифры определим все закономерности, которые на ней выполняются. Например, для цифры 2 выполнены 529 закономерностей. Тогда для каждой цифры $a \in A$ получим закономерную модель класса $\mathrm{M_a} = \langle \mathrm{a}, \mathrm{Z_a} \rangle$.

Найдем для каждого класса минимальные порождающие совокупности. Для цифры 2 это будет, например, совокупность P_2, P_3. Значения остальных признаков восстанавливается по следующим законам:
$\neg P_3 \& \neg P_2 \Rightarrow P_1$, $\neg P_3 \& \neg P_3 \& P_1 \Rightarrow P_4$, $P_4 \& \neg P_2 \& P_1 \Rightarrow \neg P_5$,
$\neg P_3 \& \neg P_2 \& P_1 \Rightarrow \neg P_6$, $\neg P_6 \& \neg P_5 \& P_4 \& P_1 \Rightarrow P_7$, $P_7 \& \neg P_3 \& P_1 \Rightarrow \neg P_8$,
$P_7 \& P_3 \& P_1 \Rightarrow P_8$, $P_8 \& \neg P_6 \& \neg P_5 \& \neg P_2 \Rightarrow P_9$.

Порождающие совокупности определяются не единственным образом. Например, совокупность P_5, P_7 так же будет порождающей для цифры 2.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
0	1	1	0	1	0	1	0	1	1
1	0	0	1	1	0	0	0	1	0
2	1	0	0	1	0	0	1	0	1
3	1	0	1	0	1	0	1	0	0
4	0	1	0	1	1	0	0	1	0
5	1	1	0	0	1	0	0	1	1
6	0	0	1	0	1	1	0	1	1
7	1	0	1	0	0	1	0	0	0
8	1	1	0	1	1	1	0	1	1
9	1	1	0	1	1	0	1	0	0

Перейдём к построению систематики. Её закон Z_S представим в виде таблицы, в каждой строке которой стоит цифра класса и значения порождающих признаков. Для выбора минимальной системообразующей совокупности систематики, рассмотрим различные порождающие совокупности для каждого из классов. Так как $2^3 = 8$ меньше чем число классов, то 3-х признаков будет недостаточно для однозначного восстановления каждого класса. Поэтому рассматриваем различные комбинации из 4-х признаков. В результате получим минимальную системообразующую совокупность признаков систематики цифр $\{P_4, P_5, P_6, P_7\}$. Систематику цифр индекса можно тогда представить таблицей 2, где для каждой цифры указаны значения системообразующих признаков $\{P_4, P_5, P_6, P_7\}$, а так же минимальные определяющие совокупности.

0	1	0	1	0	$\{P_4, P_5, P_6\}$
1	1	0	0	0	$\{P_5, P_6, P_7\}$
2	1	0	0	1	$\{P_5, P_7\}$
3	0	1	0	1	$\{P_4, P_7\}$
4	1	1	0	0	$\{P_4, P_5, P_6, P_7\}$
5	0	1	0	0	$\{P_4, P_6, P_7\}$
6	0	1	1	0	$\{P_4, P_5, P_6\}$
7	0	0	1	0	$\{P_4, P_5\}$
8	1	1	1	0	$\{P_4, P_5, P_6\}$
9	1	1	0	1	$\{P_4, P_5, P_7\}$

По значениям системообразующих признаков определяется класс, а по минимальной определяющей совокупности восстанавливаются значения всех остальных признаков.

5. Обнаружение законов в вероятностном случае. Проблема статистической двусмысленности

В приводимом примере мы рассматривали только законы, истинные для всех объектов предметной области. Но в общем случае, когда обнаруживаются не только истинные законы, но и вероятностные (см. определение далее), то логический вывод и получаемые с его помощью предсказания могут быть противоречивы.

Проблема противоречивости предсказаний, получаемых из индуктивно выведенного знания, является *проблемой статистической двусмысленности*. Существует два типа предсказаний (объяснений): *дедуктивно-номологические* (D-N) и *индуктивно-статистические* (I-S). В D-N предсказаниях, используемые законы предполагаются истинными, в том время как в I-S предсказаниях они предполагаются статистическими.

$$\frac{\mathrm{L}_1, \ldots, \mathrm{L}_\mathrm{m}}{\dfrac{\mathrm{C}_1, \ldots, \mathrm{C}_\mathrm{n}}{\mathrm{G}}}$$

Дедуктивно-номологическая модель может быть представлена схемой слева, в которой $\mathrm{L}_1, \ldots, \mathrm{L}_\mathrm{m}$ — множество законов; $\mathrm{C}_1, \ldots, \mathrm{C}_\mathrm{n}$ — множество фактов; G — предсказываемое высказывание; доказуемо $\mathrm{L}_1, \ldots, \mathrm{L}_\mathrm{m}, \mathrm{C}_1, \ldots, \mathrm{C}_\mathrm{n} \vdash \mathrm{G}$; множество $\{\mathrm{L}_1, \ldots, \mathrm{L}_\mathrm{m}, \mathrm{C}_1, \ldots, \mathrm{C}_\mathrm{n}\}$ непротиворечиво; $\mathrm{L}_1, \ldots, \mathrm{L}_\mathrm{m} \nvdash \mathrm{G}$, $\mathrm{C}_1, \ldots, \mathrm{C}_\mathrm{n} \nvdash \mathrm{G}$; законы $\mathrm{L}_1, \ldots, \mathrm{L}_m$ содержат только кванторы всеобщности; множество фактов $\mathrm{C}_1, \ldots, \mathrm{C}_n$ — бескванторные формулы.

$$\frac{\dfrac{\mathrm{F} \Rightarrow \mathrm{G},\ \mathrm{p}(\mathrm{G};\mathrm{F}) = \mathrm{r}}{\mathrm{F}(a)}\,[r]}{\mathrm{G}(a)}$$

Индуктивно-статистическая модель аналогична дедуктивно-номологической с тем отличием, что множество статистических законов $\mathrm{L}_1, \ldots, \mathrm{L}_\mathrm{m}$ должно удовлетворять требованию максимальной специфичности RMS (Requirement of Maximal Specificity).

По Гемпелю [16] *требование максимальной специфичности* (RMS) определяется следующим образом. Правило $\mathrm{F} \Rightarrow \mathrm{G}$ является максимально специфичным при состоянии знания K, если для каждого класса H, для которого оба высказывания $\forall \mathrm{x}(\mathrm{H}(\mathrm{x}) \Rightarrow \mathrm{F}(\mathrm{x}))$ и $\mathrm{H}(a)$ принадлежат K, существует статистический закон $\mathrm{p}(\mathrm{G};\mathrm{H}) = \mathrm{r}'$ в K такой, что $\mathrm{r} = \mathrm{r}'$.

Идея RMS состоит в том, что, если F и H оба содержат объект a и H является подмножеством F, то H обладает более специфической информацией об объекте a, чем F и, следовательно, закон $\mathrm{p}(\mathrm{G};\mathrm{H})$ должен предпочитаться закону $\mathrm{p}(\mathrm{G};\mathrm{F})$. Однако, для максимально специфических правил закон $\mathrm{p}(\mathrm{G};\mathrm{H})$ должен иметь ту же вероятность, что и закон $\mathrm{p}(\mathrm{G};\mathrm{F})$.

Требование максимальной специфичности позволяет найти условия для вероятностных правил, при которых можно получать предсказания без противоречий, а также получать неподвижные точки не содержащие одновременно литеру и ее отрицание. Это ставит следующие задачи, которым будут посвящены следующие разделы.

(1) В общем случае формально определить требование максимальной специфичности.

(2) Определить такие максимально специфические правила, которые бы удовлетворяли требованию максимальной специфичности и не приводили к противоречивым предсказаниям.
(3) Выяснить, можно ли использовать эти максимально специфические правила для построения неподвижных точек так, чтобы они давали определения классов, как в детерминированном случае, и при этом не содержали одновременно литеру и ее отрицание.
(4) Разработать метод обнаружения максимально специфических правил и метод построения «естественной» классификации и систематики на основе таких неподвижных точек.

6. Вероятностный закон. Семантический вероятностный вывод

Обобщим понятия закона на вероятностный случай. Определим вероятность на эмпирической системе предметной области $M = \langle A, W\rangle$ как на генеральной совокупности. Для простоты, рассмотрим вероятность $\mu : A \to [0,1]$, определенную на A (более общие случаи рассмотрены в [15]):

$$\begin{gathered}\sum_{a\in A}\mu(a) = 1, \mu(a) \neq 0, a \in A,\\ \mu(B) = \sum_{b\in B}\mu(b), B \subseteq A.\end{gathered} \tag{2}$$

Вероятность μ^n на A^n, определяется естественным образом:

$$\mu^n(a_1,\ldots,a_n) = \mu(a_1)\times\cdots\times\mu(a_n).$$

Определим *интерпретацию* языка L как отображение $I : \Im \to W$, где каждому предикатному символу $\mathrm{P_j} \in \Im$ ставится в соответствие предикат $P_j \in W$, $j = 1,\ldots,k$. Отображение $\nu : X \to A$ назовем *означиванием*. Тогда композиция отображений $\nu I\varphi$, где $\varphi \in \Re(\Im)$, дает формулу, получающуюся из φ заменой предикатных символом сигнатуры $\Im$ на предикаты W посредством интерпретации I и заменой переменных из φ на объекты из A посредством означивания ν. *Вероятность* η *предложения* $\phi(a,\ldots,b) \in \Re(\Im)$ определим как

$$\eta(\phi(a,\ldots,b)) = \mu^n(\{(a_1,\ldots,a_n) \mid M \models \nu I\phi,\ \nu(a) = a_1,\ldots,\nu(b) = a_n\}). \tag{3}$$

Определение 11. *Вероятностным законом на* M *будем называть правило* $\mathrm{C} = (\mathrm{A_1}\&\ldots\&\mathrm{A_k} \Rightarrow \mathrm{A_0})$ *вида* (1)*, у которого* $\eta(\mathrm{A_1}\&\ldots\&\mathrm{A_k}) > 0$ *и условная вероятность* $\eta(C) = \eta(\mathrm{A_0}/\mathrm{A_1}\&\ldots\&\mathrm{A_k}) > 0$ *строго больше условных вероятностей всех его подправил вида* (1) *теоремы 1. Условная вероятность подправила вида* $C = (\Rightarrow \mathrm{A_0})$ *равна* $\eta(C) = \eta(\mathrm{A_0}/\ \mathrm{T}) = \eta(\mathrm{A_0})$*. Все правила вида* $(\Rightarrow \mathrm{A_0})$*,* $\eta(\mathrm{A_0}) > 0$ *являются вероятностными законами.*

Множество всех вероятностных законов обозначим через LP.

Определение 12. *Сильнейшим вероятностным законом (SPL-правилом) на* M*, назовем вероятностный закон* $\mathrm{C} = (\mathrm{A_1}\&\ldots\&\mathrm{A_k} \Rightarrow \mathrm{A_0})$*, который не является подправилом никакого другого вероятностного закона. Обозначим через SPL – множество всех SPL-правил.*

Докажем, что понятие вероятностного закона обобщает понятие закона истинного на M.

Теорема 4. *Law ⊂ SPL ⊂ LP.*

Доказательство. Второе включение следует из определения. Рассмотрим первое включение. Если правило $\mathrm{C} \in \mathrm{L}$ имеет вид $\mathrm{C} = (\Rightarrow \mathrm{A}_0)$, то оно принадлежит LP по определению. Предположим, что правило $\mathrm{C} = (\mathrm{A}_1\&...\&\mathrm{A}_\mathrm{k} \Rightarrow \mathrm{A}_0)$ является законом на M. Докажем, что $\eta(\mathrm{A}_1\&...\&\mathrm{A}_\mathrm{k}) > 0$. Если правило C является законом на M, то подправило $(\mathrm{A}_2\&...\&\mathrm{A}_\mathrm{k} \Rightarrow \neg\mathrm{A}_1)$ не всегда истинно на M и, значит, в некоторых случаях истинна конъюнкция $\mathrm{A}_2\&...\&\mathrm{A}_\mathrm{k}\&\mathrm{A}_1$ откуда следует, что $\eta(\mathrm{A}_2\&...\&\mathrm{A}_\mathrm{k}\&\mathrm{A}_1) > 0$. Тогда условные вероятности всех подправил определены, т.к. из $\{\mathrm{A}_{i_1}, ..., \mathrm{A}_{i_h}\} \subset \{A_1, ..., A_k\}$ следует $\eta(\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h}) \geq \eta(A_1\&...\&A_k) > 0$. Докажем, что $\eta(C) = 1$.

$$\eta(C) = \eta(A_0/A_1\&...\&A_k) = \eta(A_0\&A_1\&...\&A_k)/\eta(A_1\&...\&A_k)$$
$$= \eta(A_0\&A_1\&...\&A_k)/\eta(A_0\&A_1\&...\&A_k) + \eta(\neg A_0\&A_1\&...\&A_k).$$

Поскольку правило C истинно на M, то на M нет случаев, когда конъюнкция $(\neg\mathrm{A}_0\&\mathrm{A}_1\&...\&\mathrm{A}_\mathrm{k})$ истинна и, значит, $\eta(\neg\mathrm{A}_0\&\mathrm{A}_1\&...\&\mathrm{A}_\mathrm{k}) = 0$ и $\eta(C) = 1$.

Докажем, что условная вероятность каждого подправила правила C строго меньше $\eta(C) = 1$. Любое подправило $(\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h} \Rightarrow \mathrm{A}_0)$ правила C ложно на M. Это означает, что $\eta(\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h}\&\neg\mathrm{A}_0) > 0$. Тогда

$$\eta(\mathrm{A}_0/\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h}) = \eta(\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h}\&\mathrm{A}_0)/\eta(\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h})$$
$$= \eta(\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h}\&\mathrm{A}_0)/(\eta(\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h}\&\neg\mathrm{A}_0) + \eta(\mathrm{A}_{i_1}\&...\&\mathrm{A}_{i_h}\&\mathrm{A}_0)) < 1.$$

Поскольку $\eta(C) = 1$, то правило C не может быть подправилом никакого другого вероятностного закона, т. к. тогда его условная вероятность была бы строго меньше условной вероятности этого правила, что невозможно. □

Определение 13 ([20]). *Семантическим вероятностным выводом (SP-выводом) некоторого SPL-правила, предсказывающего литеру* A_0*, будем называть последовательность вероятностных законов* $\mathrm{C}_1 \sqsubset \mathrm{C}_2 \sqsubset ... \sqsubset \mathrm{C}_\mathrm{n}$*, такую что:*

(1) $\mathrm{C}_1 = (\Rightarrow A_0)$,
(2) $\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mathrm{n} \in \mathrm{LP}$, $\mathrm{C}_\mathrm{i} = (A_1^i\&...\&A_{k_i}^i \Rightarrow A_0)$, $k_i \geq 0$,
(3) C_i *— подправило* C_{i+1}, $\{A_1^i\&...\&A_{k_i}^i\} \subset \{A_1^{i+1}\&...\&A_{k_{i+1}}^{i+1}\}$, $k_i < k_{i+1}$,
(4) $\eta(C_{i+1}) > \eta(C_i)$,
(5) C_n *— SPL-правило.*

Рассмотрим множество всех SP-выводов, предсказывающих литеру A_0. Это множество можно представить *деревом семантического вероятностного вывода литеры* A_0.

Лемма 2. *Любой вероятностный закон* $\mathrm{C} = (A_1\&...\&A_k \Rightarrow A_0)$ *принадлежит некоторому семантическому вероятностному выводу, предсказывающему литеру* A_0 *и, следовательно, дереву семантического вероятностного вывода литеры* A_0.

Доказательство. Если вероятностный закон имеет вид $\mathrm{C} = (\Rightarrow A_0)$, то проверяем является ли он сильнейшим вероятностным законом. Если да, то семантический вероятностный вывод найден, если нет, то существует вероятностный закон, для которого данный вероятностный закон является подправилом. Возьмем его в качестве следующего правила семантического вероятностного вывода

и снова проверим, является ли он сильнейшим вероятностным законом и т.д. Для вероятностного закона $\mathrm{C} = (A_1\&...\&A_k \Rightarrow A_0)$, $k \geq 1$ найдем подправило являющееся вероятностным законом. Такое подправило всегда существует, т.к. по крайней мере правило $\mathrm{C} = (\Rightarrow A_0)$ является вероятностным законом. Добавим его в качестве предшествующего правила семантического вероятностного вывода и продолжим процедуру. □

Определение 14. *Максимально специфическим законом* $\mathrm{MS}(A_0)$ *на* M *для предсказания литеры* A_0 *будем называть SPL-правило, имеющее максимальное значение условной вероятности среди всех SPL-правил дерева семантического вероятностного вывода литеры* A_0. *Если есть несколько правил с одинаковым максимальным значением, то все они являются максимально специфическими законами.*

Множество всех максимально специфических законов обозначим через MSR.

Предложение 1. *Law* $\subset$ *MSR* $\subset$ *SPL* $\subset$ *LP.*

7. Требование максимальной специфичности. Разрешение проблемы статистической двусмысленности.

Определим требование максимальной специфичности в общем случае. Будем предполагать, что в приведенной Гемпелем формулировке требования максимальной специфичности, высказывание H является предложением $\mathrm{H} \in \Re(\Im)$.

Определение 15. *Правило* $C = (F \Rightarrow G), F \in \Re(\Im), G \in Lit$ *удовлетворяет требованию максимальной специфичности (RMS), если из того что* $H \in \Re(\Im)$ *и* $F(a)\&H(a)$ *для какого-нибудь* $a \in A$ *следует, что правило* $C' = (F\&H \Rightarrow G)$ *имеет ту же вероятность, т.е.* $\eta(G/F\&H) = \eta(G/F) = r$.

Другими словами, RMS говорит о том, что не существует предложения $H \in \Re(\Im)$, которое увеличивало бы (или уменьшало, см. лемму ниже) условную вероятность $\eta(G/F) = r$ правила путем добавления его к посылке.

Лемма 3. *Если утверждение* $H \in \Re(\Im)$ *уменьшает условную вероятность правила* $\eta(G/F\&H) < \eta(G/F)$, *то* $\neg H$ *увеличивает её и* $\eta(G/F\&\neg H) > \eta(G/F)$.

Доказательство. Обозначим $\mathrm{a} = \eta(\mathrm{G\&F\&H})$, $\mathrm{b} = \eta(\mathrm{F\&H})$, $\mathrm{c} = \eta(\mathrm{G\&F\&}\neg\mathrm{H})$, $\mathrm{d} = \eta(\mathrm{F\&}\neg\mathrm{H})$. Тогда исходное неравенство $\eta(\mathrm{G/F\&H}) < \eta(\mathrm{G/\ F})$ перепишется как $\mathrm{a/b} < (\mathrm{a} + \mathrm{c})/(\mathrm{b} + \mathrm{d})$, из которого следует, что $(\mathrm{a} + \mathrm{c})/(\mathrm{b} + \mathrm{d}) < \mathrm{c/d} \Leftrightarrow \eta(\mathrm{G/F}) < \eta(\mathrm{G/F\&}\neg\mathrm{H})$. □

Лемма 4. *Для любого правила* $\mathrm{C} = (\mathrm{A_1\&...\&A_k} \Rightarrow \mathrm{A_0})$ *вида* (1), $\eta(\mathrm{A_1\&...\&A_k}) > 0$ *найдется вероятностный закон* $C' = (B_1\&...\&B_{k'} \Rightarrow \mathrm{A_0})$, $B_1\&...\&B_{k'} \subset \mathrm{A_1\&...\&A_k}$, $k' < k$ *на* M, *для которого* $\mu(C') \geq \mu(C)$.

Доказательство. Правило $\mathrm{C} = (\mathrm{A_1\&...\&A_k} \Rightarrow \mathrm{A_0})$, либо является вероятностным законом, либо для него существует подправило $R' = (P_1\& \ldots \&P_{k'} \Rightarrow \mathrm{A_0})$, $k' \geq 0$, $\{P_1, \ldots, P_{k'}\} \subset \{A_1, \ldots, A_k\}$, $k' < k$ такое, что $\mu(R') \geq \mu(A)$. Аналогично для правила R', либо оно является вероятностным законом, либо для него существует подправило с аналогичными свойствами и т.д. □

Теорема 5 ([20]). *Любой максимально специфический закон* $\mathrm{MS}(G) = (\mathrm{F} \Rightarrow \mathrm{G})$, $\mathrm{F} \in \Re(\Im)$, $\mathrm{G} \in \mathrm{Lit}$ *удовлетворяет требованию максимальной специфичности.*

Доказательство. Надо доказать, что для любого предложения $\mathrm{H} \in \Re(\Im)$, если $\mathrm{F}(a)\&\mathrm{H}(a)$, $a \in A$ истинно на M, то имеет место равенство $\eta(\mathrm{G/F\&H}) = \eta(\mathrm{G/F}) = \mathrm{r}$. Из условия истинности $\mathrm{F}(a)\&\mathrm{H}(a)$ на M следует, что $\eta(\mathrm{F\&H}) > 0$ и, следовательно, условная вероятность определена.

Рассмотрим случай, когда H является литерой (B или $\neg$B). Предположим противное, что $\eta(\mathrm{G/F\&H}) \neq \mathrm{r}$. Тогда, согласно лемме 3, будет выполнено одно из неравенств $\eta(F\&B \Rightarrow G) > r$ или $\eta(F\&\neg B \Rightarrow G) > r$ для одного из правил $(F\&B \Rightarrow G)$ или $(F\&\neg B \Rightarrow G)$. Тогда, согласно лемме 4, существует вероятностный закон C', являющийся подправилом и имеющий не меньшую условную вероятность, тогда $\eta(\mathrm{C}') > \mathrm{r}$. Отсюда по лемме 2 вероятностный закон C' принадлежит некоторому дереву семантического вероятностного вывода и имеет большее значение условной вероятности, чем максимально специфический закон $\mathrm{MS}(G)$, предсказывающий G, что противоречит максимальной специфичности $\mathrm{MS}(G)$.

Рассмотрим случай, когда предложение H является конъюнкцией двух литер, для которых теорема уже доказана. Предположим противное, что одно из неравенств $\eta(\mathrm{G/F\&B_1\&}B_2)>r$, $\eta(\mathrm{G/F\&}\neg\mathrm{B_1\&}B_2)>r$, $\eta(\mathrm{G/F\&B_1\&}\neg B_2)>r$, $\eta(\mathrm{G/F\&}\neg\mathrm{B_1\&}\neg B_2)>r$ выполнено. Тогда по лемме 4 и лемме 2 существует вероятностный закон C', принадлежащий дереву семантического вероятностного вывода, являющийся подправилом одного из этих правил и такой, что $\eta(\mathrm{C}') > \mathrm{r}$. Но это невозможно, так как правило $C = (\mathrm{F\&H} \Rightarrow \mathrm{G})$ максимально специфично. Следовательно, для всех этих неравенств мы можем имееть только равенство $=$ или неравенство $<$. Последний случай невозможен из-за следующего равенства:

$$r = \frac{\eta(\mathrm{G\&F})}{\eta(F)} = \frac{\eta(\mathrm{G\&F\&B_1\&}B_2) + \eta(\mathrm{G\&F\&}\neg\mathrm{B_1\&}B_2) + \eta(\mathrm{G\&F\&B_1\&}\neg B_2) + \eta(\mathrm{G\&F\&}\neg\mathrm{B_1\&}\neg B_2)}{\eta(F\&\mathrm{B_1\&}B_2) + \eta(F\&\neg\mathrm{B_1\&}B_2) + \eta(F\&\mathrm{B_1\&}\neg B_2) + \eta(F\&\neg\mathrm{B_1\&}\neg B_2)}.$$

Случай, когда предложение H является конъюнкцией нескольких атомов или их отрицаний доказывается индукцией.

В общем случае предложение $\mathrm{H} \in \Re(\Im)$ может быть представлено как дизъюнкция конъюнкций атомов или их отрицаний. Для завершения доказательства нам достаточно рассмотреть случай, когда предложение H является дизъюнкцией двух непересекающихся предложений $\mathrm{D} \vee \mathrm{E}$, $\eta(\mathrm{D\&E}) = 0$, для которых теорема уже доказана, т.е.

$$\eta(\mathrm{G/F\&D}) = \eta(\mathrm{G/F\&E}) = \eta(\mathrm{G/F}) = \mathrm{r}.$$

Тогда

$$\eta(\mathrm{G/F\&(D} \vee E)) = \frac{\eta(\mathrm{G\&F\&(D} \vee E))}{\eta(\mathrm{F\&(D} \vee E))} = \frac{\eta(\mathrm{G\&F\&D}) + \eta(\mathrm{G\&F\&}E)}{\eta(\mathrm{F\&D}) + \eta(\mathrm{F\&}E)} = r.$$

Случай дизъюнкции большего числа непересекающихся предложений доказывается по индукции. □

Для решения проблемы статистической двусмысленности докажем сначала, что не существует двух максимально специфических законов, противоречащих друг другу.

Теорема 6 ([20])**.** *Среди максимально специфических законов нет двух законов* $\mathrm{A}, \mathrm{B} \in \mathrm{MSR}$, $\mathrm{A} = (\mathrm{A_1\&...\&}A_k \Rightarrow \mathrm{G})$, $\mathrm{B} = (\mathrm{B_1\&...\&}B_m \Rightarrow \neg\mathrm{G})$, $\eta((\mathrm{A_1\&...\&}A_k)\&(\mathrm{B_1\&...\&}B_m)) > 0$, $k, m \geq 0$, $k > 0$ *или* $m > 0$, *предсказывающих противоречие* G *и* $\neg G$.

Для доказательства теоремы сначала докажем следующую лемму.

Лемма 5. *Если для правил* $\mathrm{A} = (\bar{\mathrm{A}} \Rightarrow \mathrm{G})$, $\mathrm{B} = (\bar{\mathrm{B}} \Rightarrow \neg\mathrm{G})$, $\bar{\mathrm{A}} = \mathrm{A}_1\&...\&A_k$, $\bar{\mathrm{B}} = \mathrm{B}_1\&...\&B_m$, $\eta(\bar{\mathrm{A}}\&\neg\bar{\mathrm{B}}) > 0$, $k \geq 0$, $m > 0$, *верно неравенство* $\eta(G/\bar{A}\&\neg\bar{B}) > \eta(G/\bar{A})$, *то существует правило имеющее строго большую условную вероятность, чем правило* A.

Доказательство. Распишем условную вероятность

$$\eta(G/\bar{A}\&\neg\bar{B}) = \eta(G/\bar{A}\&(\neg\mathrm{B}_1 \vee ... \vee \neg\mathrm{B}_m)).$$

Представим дизъюнкцию $\neg\mathrm{B}_1 \vee ... \vee \neg\mathrm{B}_m$ как дизъюнкцию конъюнкций $\bigvee\limits_{i=(1,...1,0)}^{i=(0,...,0)}(\mathrm{B}_1^{i_1}\&...\&\mathrm{B}_m^{i_m})$, где $i = (i_1, ..., i_m)$, $i_1, ..., i_m \in \{0, 1\}$, ноль означает наличие отрицания у соответствующего атома, а единица - отсутствие отрицания. Дизъюнкция не включает $(1,\ldots,1)$, соответствующий конъюнкции $\mathrm{B}_1\&...\&\mathrm{B}_m$.

Тогда условная вероятность $\eta(G/\bar{A}\&(\neg\mathrm{B}_1 \vee ... \vee \neg\mathrm{B}_m))$ перепишется как

$$\eta\left(G/\bigvee_{i=(0,...,0)}^{i=(1,...1,0)}(\bar{A}\&\mathrm{B}_1^{i_1}\&...\&\mathrm{B}_m^{i_m})\right).$$

Докажем, что если $\eta(G/\bar{A}\&\neg\bar{B}) > \eta(G/\bar{A})$, то также будет выполнено одно из неравенств

$$\eta(G/\bar{A}\&B_1^{i_1}\&...\&B_m^{i_m}) > \eta(G/\bar{A}), (i_1, ..., i_m) \neq (1, ..., 1).$$

Предположим противное, что одновременно выполнены все неравенства

$$\eta(G/\bar{A}\&B_1^{i_1}\&...\&B_m^{i_m}) \leq \eta(G/\bar{A}), (i_1, ..., i_m) \neq (1, ..., 1)$$

в тех случаях, когда $\eta(\bar{A}\&B_1^{i_1}\&...\&B_m^{i_m}) > 0$. Поскольку $\eta(\bar{A}\&\neg\bar{B}) > 0$, то есть случаи, когда $\eta(\bar{A}\&B_1^{i_1}\&...\&B_m^{i_m}) > 0$.

Тогда

$$\eta(G\&\bar{A}\&B_1^{i_1}\&...\&B_m^{i_m}) \leq \eta(G/\bar{A})\eta(\bar{A}\&B_1^{i_1}\&...\&B_m^{i_m}), (i_1, ..., i_m) \neq (1, ..., 1),$$

$$\eta\left(G/\bigvee_{i=(0,...,0)}^{i=(1,...1,0)}(\bar{A}\&\mathrm{B}_1^{i_1}\&...\&\mathrm{B}_m^{i_m})\right) = \frac{\eta\left(\bigvee\limits_{i=(0,...,0)}^{i=(1,...1,0)}(G\&\bar{A}\&\mathrm{B}_1^{i_1}\&...\&\mathrm{B}_m^{i_m})\right)}{\eta\left(\bigvee\limits_{i=(0,...,0)}^{i=(1,...1,0)}(\bar{A}\&\mathrm{B}_1^{i_1}\&...\&\mathrm{B}_m^{i_m})\right)} =$$

$$\frac{\sum\eta(G\&\bar{A}\&\mathrm{B}_1^{i_1}\&...\&\mathrm{B}_m^{i_m})}{\sum\eta(\bar{A}\&\mathrm{B}_1^{i_1}\&...\&\mathrm{B}_m^{i_m})} \leq \frac{\eta(G/\bar{A})\sum\eta(\bar{A}\&\mathrm{B}_1^{i_1}\&...\&\mathrm{B}_m^{i_m})}{\sum\eta(\bar{A}\&\mathrm{B}_1^{i_1}\&...\&\mathrm{B}_m^{i_m})} = \eta(G/\bar{A})$$

что противоречит неравенству $\eta(G/\bar{A}\&\neg\bar{B}) > \eta(G/\bar{A})$. Поэтому наше предположение не верно и существует правило вида

$$\bar{A}\&B_1^{i_1}\&...\&B_m^{i_m} \Rightarrow G, (i_1, ..., i_m) \neq (1, ..., 1)$$

имеющее строго большую оценку условной вероятности, чем A. □

Вернемся к теореме. Предположим противное, что есть максимально специфические законы $\mathrm{A}, \mathrm{B} \in MSR$, $\mathrm{A} = (\bar{\mathrm{A}} \Rightarrow \mathrm{G})$, $\mathrm{B} = (\bar{\mathrm{B}} \Rightarrow \neg\mathrm{G})$, $\bar{\mathrm{A}} = \mathrm{A}_1\&...\&A_k$, $\bar{\mathrm{B}} = \mathrm{B}_1\&...\&B_m$, $\eta(\bar{\mathrm{A}}\&\bar{\mathrm{B}}) > 0$, $k, m \geq 0$, $k > 0$ или $m > 0$. Докажем, что в этом случае существует правило, имеющее строго большую условную вероятность, чем одно из правил A, B. Тогда по лемме 4 и лемме 2 будет существовать

максимально специфический закон, имеющий строго большую условную вероятность, что противоречит их максимальной специфичности.

Предположим противное, что все другие правила предсказывающие те же предикаты G и $\neg G$ имеют условную вероятность не большую, чем правила $\mathrm{A}, \mathrm{B} \in \mathrm{MSR}$.

По условию теоремы $k > 0$ или $m > 0$, либо $\bar{A} \neq \emptyset$, либо $\bar{B} \neq \emptyset$.

I. Если $\bar{B} \neq \emptyset$, то рассмотрим правило $\bar{\mathrm{A}}\&\bar{\mathrm{B}} \Rightarrow G$. По предположению $\eta(G/\bar{A}\&\bar{B}) \leq \eta(G/\bar{A})$.

Рассмотрим три случая:

(1) $\eta(G/\bar{A}\&\bar{B}) < \eta(G/\bar{A})$, $\eta(\bar{A}\&\neg\bar{B}) \neq 0$. Тогда по лемме 3 $\eta(G/\bar{A}\&\neg\bar{B}) > \eta(G/\bar{A})$. Тогда в силу леммы 5 существует правило, имеющее строго большую условную вероятность, чем правило A. Поэтому в этом случае теорема верна.
(2) случай $\eta(G/\bar{A}\&\bar{B}) = \eta(G/\bar{A})$, будет рассмотрен далее.
(3) случай $\eta(G/\bar{A}\&\bar{B}) < \eta(G/\bar{A})$, когда $\eta(\bar{A}\&\neg\bar{B}) = 0$ и $\bar{A} \neq \emptyset$ сводиться к случаю 2. Так как из $\eta(\bar{A}\&\bar{B}) > 0$ и, следовательно, $\eta(\bar{A}) > 0$ получаем, что

$$\eta(G/\bar{A}) = \frac{\eta(G\&\bar{A})}{\eta(\bar{A})} = \frac{\eta(G\&\bar{A}\&\bar{B}) + \eta(G\&\bar{A}\&\neg\bar{B})}{\eta(\bar{A}\&\bar{B}) + \eta(\bar{A}\&\neg\bar{B})} = \frac{\eta(G\&\bar{A}\&\bar{B})}{\eta(\bar{A}\&\bar{B})} = \eta(G/\bar{A}\&\bar{B}).$$

Если $\bar{A} = \emptyset$ и $\eta(\bar{A}\&\neg\bar{B}) = 0 = \eta(\neg\bar{B}) = 0$, то, поскольку $\eta(\bar{A}\&\bar{B}) > 0$ и, следовательно, $\eta(\bar{\mathrm{B}}) > 0$, получаем, что $\eta(\bar{B}) = 1$. Тогда

$$\eta(\neg G) = \frac{\eta(\neg G\&\bar{B}) + \eta(\neg G\&\neg\bar{B})}{\eta(\bar{B}) + \eta(\neg\bar{B})} = \frac{\eta(\neg G\&\bar{B})}{\eta(\bar{B})} = \eta(\neg G/\bar{B}),$$

что невозможно в силу того, что правило $\mathrm{B} = (\bar{\mathrm{B}} \Rightarrow \neg\mathrm{G})$ является максимально специфическим законом для которого $\eta(\neg G) < \eta(\neg G/\bar{B})$.

II. Если $\bar{A} \neq \emptyset$, то рассмотрим другое правило $\bar{\mathrm{A}}\&\bar{\mathrm{B}} \Rightarrow \neg G$. Также по предположению $\eta(\neg G/\bar{A}\&\bar{B}) \leq \eta(\neg G/\bar{B})$. Проводя аналогичные рассуждения, как в случае правила $\bar{\mathrm{A}}\&\bar{\mathrm{B}} \Rightarrow G$, получим, что либо теорема верна, либо надо рассмотреть оставшийся случай

$$\eta(\neg G/\bar{A}\&\bar{B}) = \eta(\neg G/\bar{B}), \text{ когда } \bar{A} \neq \emptyset,\ \eta(\bar{A}) > 0$$

Рассмотрим оставшиеся случаи равенств

$$\eta(G/\bar{A}\&\bar{B}) = \eta(G/\bar{A}), \text{ когда } \bar{B} \neq \emptyset \text{ и } \eta(\bar{\mathrm{B}}) > 0$$
$$\eta(\neg G/\bar{A}\&\bar{B}) = \eta(\neg G/\bar{B}), \text{ когда } \bar{A} \neq \emptyset,\ \eta(\bar{A}) > 0.$$

Рассмотрим сумму

$$\eta(G/\bar{A}\&\bar{B}) + \eta(\neg G/\bar{A}\&\bar{B}) = 1 = \eta(G/\bar{A}) + \eta(\neg G/\bar{B}).$$

Поскольку либо $\bar{A} \neq \emptyset$, либо $\bar{B} \neq \emptyset$, то, по крайней мере, одно из правил $\bar{\mathrm{A}} \Rightarrow \mathrm{G}$, $\bar{\mathrm{B}} \Rightarrow \neg\mathrm{G}$ является вероятностным законом, удовлетворяющим, либо условию $\eta(G/\bar{A}) > \eta(G)$, либо условию $\eta(\neg G/\bar{B}) > \eta(\neg G)$. Тогда получим противоречие $1 = \eta(G/\bar{A}) + \eta(\neg G/\bar{B}) > \eta(G) + \eta(\neg G) = 1$. □

Далее мы также докажем, что любой I-S вывод непротиворечив для любых законов $\mathrm{L}_1, \ldots, \mathrm{L}_\mathrm{m} \in \mathrm{MSR}$.

8. Неподвижные точки выводов по MSR правилам

Определение 16. *Определим оператор непосредственного следования Pr по правилам из* $P \subset$ MSR *на наборе литер* L *следующим образом:*

$$\Pr_P(L) = L \cup \{\mathrm{A}_0 \mid \mathrm{C} = (\mathrm{A}_1\&...\&\mathrm{A}_\mathrm{k} \Rightarrow \mathrm{A}_0),\ \{\mathrm{A}_1, \ldots, \mathrm{A}_\mathrm{k}\} \subset L,\ C \in P\}$$

Определение 17. *Неподвижной точкой оператора Pr непосредственного следования относительно набора литер* L *назовем замыкание* $\Pr_P^\infty(L)$ *этого множества литер относительно оператора непосредственного следования, такое что* $\Pr_P(\Pr_P^\infty(L)) = \Pr_P^\infty(L)$.

Определение 18. *Набор литер* $L = \{L_1, \ldots, L_\mathrm{k}\}$ *назовем совместным, если* $\eta(\mathrm{L}_1\&...\&L_\mathrm{k}) > 0$.

Определение 19. *Набор литер* L *непротиворечив, если он не содержит одновременно атом* G *и его отрицание* $\neg G$.

Предложение 2. *Если* L *совместно, то* L *непротиворечиво.*

Доказательство. Если L совместно, то не может существовать атом $G \in L$ и его отрицание $\neg G \in L$, поскольку тогда $\eta(\&L) \le \eta(G\&\neg G) = 0$, где $\&L$ – конъюнкция литер из L □

Покажем, что непосредственное следование сохраняет свойство совместности.

Теорема 7 ([4])**.** *Если* L *совместно, то* $\Pr_P(L)$ *также совместно,* P $\subset$ MSR.

Доказательство. Надо доказать, что при каждом применении какого-либо закона из P $\subset$ MSR мы снова получаем совместный набор литер. Предположим противное, что при применении некоторого закона $\mathrm{A} = (\mathrm{A}_1\&...\&A_k \Rightarrow \mathrm{G})$, $\{\mathrm{A}_1, ..., A_k\} \subset L$, $k > 1$ к набору литер $L = \{L_1, ..., L_\mathrm{k}\}$ мы получим литеру G, для которой $\mu(\mathrm{L}_1\&...\&L_n\&G) = 0$.

Поскольку для законов MSR выполнены неравенства $\eta(\mathrm{G}/\mathrm{A}_1\&...\&A_k) > \eta(\mathrm{G})$, $\eta(\mathrm{A}_1\&...\&A_k) > 0$, $\eta(\mathrm{G}) > 0$, то $\eta(\mathrm{G}\&\mathrm{A}_1\&...\&A_k) > \eta(\mathrm{G})\eta(\mathrm{A}_1\&...\&A_k) > 0$.

Добавим отрицание литер $\{B_1, ..., B_t\} = \{\mathrm{L}_1\&...\&L_n\}\backslash\{\mathrm{A}_1, ..., A_k\}$ в правило A, получим правило $(\mathrm{A}_1\&...\&A_k\&\neg(B_1\&...\&B_t) \Rightarrow \mathrm{G})$.

Обозначим $\&\mathrm{A}_i = \mathrm{A}_1\&...\&A_k$, $\&B_j = B_1\&...\&B_t$, $\&\mathrm{L} = \mathrm{L}_1\&...\&L_n$.

По предположению $\mu(\mathrm{L}_1\&...\&L_n\&G) = 0$ и $\eta(\&\mathrm{A}_i\&(\&B_j)) = \eta(\&\mathrm{L}) > 0$. Докажем, что в этом случае $\eta(\&\mathrm{A}_i\&\neg(\&B_j)) > 0$. Предположим противное, что $\eta(\&\mathrm{A}_i\&\neg(\&B_j)) = 0$, тогда

$$\eta(\mathrm{G}\&(\&\mathrm{A}_i)\&\neg(\&B_j)) \le \eta(\&\mathrm{A}_i\&\neg(\&B_j)) = 0.$$

Откуда следует, что

$$0 = \mu(\&\mathrm{L}\&G) = \eta(\mathrm{G}\&(\&\mathrm{A}_i)\&(\&B_j)) = \eta(\mathrm{G}\&(\&\mathrm{A}_i) - \eta(\mathrm{G}\&(\&\mathrm{A}_i)\&\neg(\&B_j))) =$$
$$\eta(\mathrm{G}\&\mathrm{A}_1\&...\&A_k) > \eta(\mathrm{G})\eta(\mathrm{A}_1\&...\&A_k) > 0.$$

Получили противоречие. Тогда

$$\eta(\mathrm{G}/\&\mathrm{A}_i\&\neg(\&B_j)) = \frac{\eta(\mathrm{G}\&(\&\mathrm{A}_i)\&\neg(\&B_j))}{\eta(\&\mathrm{A}_i\&\neg(\&B_j))} = \frac{\eta(\mathrm{G}\&(\&\mathrm{A}_i)) - \eta(\mathrm{G}\&(\&\mathrm{A}_i)\&(\&B_j))}{\eta(\&\mathrm{A}_i) - \eta(\&\mathrm{A}_i\&(\&B_j))} =$$
$$\frac{\eta(\mathrm{G}\&(\&\mathrm{A}_i)) - \eta(\&\mathrm{L}\&\mathrm{G})}{\eta(\&\mathrm{A}_i) - \eta(\&\mathrm{A}_i\&(\&B_j))} = \frac{\eta(\mathrm{G}\&(\&\mathrm{A}_i))}{\eta(\&\mathrm{A}_i) - \eta(\&\mathrm{A}_i\&(\&B_j))} > \frac{\eta(\mathrm{G}\&(\&\mathrm{A}_i))}{\eta(\&\mathrm{A}_i)} = \eta(G/\mathrm{A}_1\&...\&A_k).$$

Тогда в силу лемм 2, 4, 5 мы получим, что существует вероятностный закон с большей условной вероятностью, чем правило A, что противоречит максимальной специфичности правила A. □

Следствие 2. *Если L совместно, то* $\mathrm{Pr}_P(L)$ *непротиворечиво,* $\mathrm{P} \subset \mathrm{MSR}$.

Следствие 3. *(Решение проблемы статистической двусмысленности) I-S вывод непротиворечив для любого множества законов* $\mathrm{P} = \{\mathrm{L}_1, \ldots, \mathrm{L}_\mathrm{m}\} \subset \mathrm{MSR}$ *и множества фактов* $\{C_1, ..., C_n\}$.

Следствие 4. *Неподвижные точки* $\mathrm{Pr}^\infty_P(L)$ *для совместного набора литер L совместны и непротиворечивы.*

Множество всех неподвижных точек $L = \mathrm{Pr}^\infty_{MSR}(N)$, полученных по всем максимально специфическим законам на всех совместных наборах литер N, обозначим через Class(M).

Следствие 5. *Поскольку* $\mathrm{Law} \subset MSR$, *для любого класса* $N \uparrow \mathrm{Law} \in \mathrm{Cl(M)}$ *существует класс* $\mathrm{Pr}^\infty_{MSR}(N)$ *из* Class(M) *такой что* $N \uparrow \mathrm{Law} \subset \mathrm{Pr}^\infty_{MSR}(N)$.

9. «Естественная» классификация как неподвижные точки предсказаний по максимально специфическим правилам.

Таким образом, нам удалось решить первые три задачи обобщения неподвижных точек на вероятностный случай, обозначенные в конце раздела 5.

Определим понятие «естественной» классификации и систематики, приведенные в определении 10, через неподвижные точки по максимально специфическим законам.

Определение 20.
(1) *Множество всех неподвижных точек Class(M) будем называть множеством всех классов эмпирической системы M.*
(2) *Каждый класс L выделяет в эмпирической системе M множество объектов принадлежащих классу* $M(L) = \{b \in A \mid L \subset \mathrm{S}_b\}$.
(3) *Закономерную модель класса* $C = \langle L, \mathrm{Z}_L\rangle$, $L \in \mathrm{Class(M)}$ *определим как множество литер L, неподвижной точки оператора, и множество всех правил* $\mathrm{Z}_L \subset MSR$, *применимых к литерам из L.*
(4) *Порождающим множеством некоторого класса* $C = \langle L, \mathrm{Z}_L\rangle$ *будем называть такое подмножество литер* $N \subset L$, *что* $L = \mathrm{Pr}^\infty_{MSR}(N)$.
(5) *Набор S атомарных высказываний* $\mathrm{P}_j(a)$, *j=1,... s,* $s \leq k$ *будем называть системообразующим, если для каждого класса из Class(M) есть порождающее множество литер, которые включают только атомарные высказывания из системообразующего набора.*
(6) *Систематику определим как набор* $\Sigma = \langle S, Z_S, \{Z_{L_i}\}_{L_i \in Class(M)}\rangle$, *где S – системообразующий набор атомарных высказываний,* Z_S *– закон систематики, определяющий порядок взятия отрицаний для атомарных высказываний из S,* $\{Z_{L_i}\}_{L_i \in Class(M)}$ *– множества правил для неподвижных точек из Class(M).*

Определение классов как неподвижных точек по максимально специфическим законам является одновременно вероятностным обобщением формальных понятий, как это показано в работе [4]. Поэтому полученные классы можно также рассматривать как веротяностные формальные понятия.

Если эмпирическая система $M = \langle A, W\rangle$ как генеральная совокупность нам известна, как это имеет место в анализе формальных понятий, то обнаружив все классы мы обнаружим вероятностные формальные понятия.

Однако нашей целью является разработка метода «естественной» классификации как методов кластеризации. В этом случае эмпирическая система $M = \langle A, W\rangle$ как генеральная совокупность нам не известна, а известна только выборка данных из генеральной совокупности. Поэтому для разработки метода «естественной» классификации надо рассмотреть вопросы построения «естественной» классификации и неподвижных точек по выборкам данных.

10. Метод «естественной» классификации на данных.

Под выборкой из генеральной совокупности $M = \langle A, W\rangle$ будем понимать подмодель $B_B = \langle B, W\rangle$, $B_B \subset M$, где B – случайная выборка объектов из генеральной совокупности A. На выборке определим вероятность μ_B, приписав каждому элементу выборки вероятность $\mu_B(a) = 1/|B|$. По ней определяется вероятность η_B на высказываниях из $\Re(\Im)$. На выборке $B_B = \langle B, W\rangle$, как на подмодели, можно получить теорию $\mathrm{Th}(B)$, множество законов $\mathrm{Law}(B)$, множество сильнейших вероятностных законов $\mathrm{SPL}(B)$ и множество максимально специфических законов $\mathrm{MSR}(B)$.

Предложение 3. $\mathrm{Th}(M) \subset \mathrm{Th}(B)$.

Поскольку каждое множество литер S_b, $b \in B$ совместно, т. к. получено на реальном объекте, имеющем ненулевую вероятность, то множеством всех неподвижных точек на выборке B_B будет $\mathrm{Class}(B) = \{L = \mathrm{Pr}^{\infty}_{MSR(B)}(N) | N \subset S_b, b \in B\}$. Классы из Class(B) также будут вероятностными формальными понятиями [4].

Но нас интересуют не вероятностные формальные понятия, определенные на выборке $B_B \subset M$, а возможность обнаружения и распознавания неизвестных нам классов генеральной совокупности по классам, обнаруженным на выборке. В этом случае вероятность μ на генеральной совокупности нам не известна, а известна только вероятность μ_B.

Максимально специфические законы $\mathrm{MSR}(B)$ на выборке B_B могут не являться таковыми на генеральной совокупности. Их можно рассматривать как аппроксимации законов $\mathrm{MSR}(M)$ в следующем смысле. Вспомним определение максимально специфических законов через семантический вероятностный вывод – мы в процессе вывода последовательно наращивали посылку правила, строго увеличивая его условную вероятность и включая как можно больше релевантной информации, в соответствии с требованием максимальной специфичности, чтобы обеспечить максимально вероятное и непротиворечивое предсказание. Добавляя в правило только предикаты, строго увеличивающие условную вероятность, мы также исключаем из рассмотрения случайные предикаты, не имеющие отношение к предсказанию. По выборке B_B заранее нельзя сказать, какой из сильнейших вероятностных законов $\mathrm{SPL}(B)$ будет таковым и на M и может дать максимально специфический закон из $\mathrm{MSR}(M)$. Но множество сильнейших вероятностных законов $\mathrm{SPL}(B)$ является аппроксимацией множества законов $\mathrm{MSR}(M)$. Дерево семантического вероятностного вывода можно строить по выборке наращивая посылку и поверяя некоторым статистическим критерием строгое увеличение условной вероятности. В своих работах

[18,22] мы используем для проверки вероятностных неравенств семантического вероятностного вывода точный критерий независимости Фишера для таблиц сопряженности. Используя этот критерий, мы можем обнаружить по выборке $B_B = \langle B, W\rangle$ множество $\mathrm{LP}_\alpha(B)$ вероятностных законов с доверительным уровнем α, в которых каждое вероятностное неравенство будет статистически значимо с уровнем доверия α. По множеству $\mathrm{LP}_\alpha(B)$ мы можем найти множество $\mathrm{SPL}_\alpha(B)$ сильнейших вероятностных законов, обнаруженных с доверительным уровнем α на выборке.

Вывод по правилам $\mathrm{SPL}_\alpha(B)$ может быть противоречив, поэтому для построения неподвижных точек по множеству $\mathrm{SPL}_\alpha(B)$ необходимо использовать более слабый критерий согласованности законов по предсказанию, допускающий наличие противоречивых предсказаний. Для этого определим специальный оператор непосредственного следования $\mathrm{Pr}\Phi_{SPL_\alpha}(L)$. Поскольку дальнейшие рассуждения относятся к выборке B, то ее как параметр будем опускать. Оператор $\mathrm{Pr}\Phi_{SPL_\alpha}(L)$ учитывает взаимную согласованность законов по предсказанию следующим образом.

Определим множество законов, *подтверждающихся* на наборе литер L как

$$\mathrm{Sat(L)} = \{\mathrm{C} \mid C \in SPL_\alpha, \mathrm{C} = (\mathrm{A_1\&...\&A_k} \Rightarrow A_0), \{\mathrm{A_1}, ..., \mathrm{A_k}\} \subset L, \mathrm{A_0} \in L\},$$

а также множество законов *опровергающихся* на наборе литер L как

$$\mathrm{Fal(L)} = \{C \mid C \in SPL_\alpha, \mathrm{C} = (\mathrm{A_1\&...\&A_k} \Rightarrow A_0), \{\mathrm{A_1}, ..., \mathrm{A_k}\} \subset L, \neg\mathrm{A_0} \in L\}.$$

Определим критерий Kr взаимной согласованности законов из SPL_α по предсказанию на наборе L:

$$Kr_{SPL_\alpha}(L) = \sum_{C \in Sat(L)} \nu(C) - \sum_{C \in Fal(L)} \nu(C),\ \nu(C) = -\log(1 - \eta_B(C)).$$

Функция $-\log(1 - \eta_B(C))$ учитывает не саму вероятность, а её близость к 1.

Оператор $\mathrm{Pr}\Phi_{SPL_\alpha}(L)$ работает следующим образом: он либо добавляет новую литеру к набору L, либо удаляет одну из литер из набора L. При этом, каждый раз должна строго увеличиваться взаимная согласованность применимых к набору L законов, т.е. каждый раз должно выполняться неравенство $Kr_{SPL_\alpha}(\mathrm{Pr}\Phi_{SPL_\alpha}(L)) > Kr_{SPL_\alpha}(L)$. В противном случае набор L остается без изменений и является неподвижной точкой. В обоих случаях нас интересует такое добавление/удаление элемента, которое максимально увеличивает критерий $Kr_{SPL_\alpha}(L)$. Изменения критерия, при добавлении/удалении элемента равны соответственно:

$$\delta^+(L) = \max_{A_0 \in \mathrm{Pr}_{SPL_\alpha}(L),\ A_0 \notin L} \{Kr_{SPL_\alpha}(L \cup A_0) - Kr_{SPL_\alpha}(L)\},$$

$$\delta^-(L) = \max_{A_0 \in \mathrm{Pr}_{SPL_\alpha}(L),\ A_0 \in L} \{Kr_{SPL_\alpha}(L \backslash A_0) - Kr_{SPL_\alpha}(L)\}.$$

Оператор $\mathrm{Pr}\Phi_{SPL_\alpha}(L)$ добавляет/удаляет тот элемент из набора L, который максимизирует соответствующее значение. Добавляемые/удаляемые элементы определяются следующим образом:

$$(A_0)^+ = \arg\max_{A_0 \in \mathrm{Pr}_{SPL_\alpha}(L),\ A_0 \notin L} (Kr_{SPL_\alpha}(L \cup A_0)),$$

$$(A_0)^- = \arg\max_{A_0 \in \mathrm{Pr}_{SPL_\alpha}(L),\ A_0 \in L} (Kr_{SPL_\alpha}(L \backslash A_0)).$$

При каждом применении оператор $\Pr\Phi_{SPL_\alpha}(L)$ добавляет/удаляет тот элемент, который максимально увеличивает критерий, т.е. добавляет элемент $(A_0)^+$, если $\delta^+(L) > \delta^-(L)$, $\delta^+(L) > 0$, и удаляет элемент $(A_0)^-$, если $\delta^-(L) > \delta^+(L)$, $\delta^-(L) > 0$.

Таким образом, оператор $\Pr\Phi_{SPL_\alpha}(L)$ определяется следующим образом:

$$Pr\Phi_{SPL_\alpha}(L) = \left\{ \begin{array}{l} L \cup (A_0)^+, \text{ if } \delta^+(L) > \delta^-(L), \delta^+(L) > 0 \\ L \backslash (A_0)^-, \text{ if } \delta^-(L) \geq \delta^+(L), \delta^-(L) > 0 \\ L, \text{ else.} \end{array} \right\}$$

Неподвижная точка $\Pr\Phi_{SPL_\alpha}(L) = L$ получается в третьем случае, когда добавление/удаление элемента не увеличивает критерий.

Множество всех таких неподвижных точек, полученных на всех совместных наборах литер L по закономерностям SPL_α определим как множество «естественных» классов $\mathrm{Class}_\alpha(M)$.

Чтобы получить закономерную модель класса $L \in \mathrm{Class}_\alpha(M)$ надо определить множество закономерностей Z_L взаимно предсказывающих признаки класса. Такими закономерностями будут закономерности Sat(L). В неподвижной точке L опровергающиеся предсказания по закономерностям Fal(L) перекрываются подтверждающимися предсказаниями по закономерностям Sat(L). Поэтому закономерности Fal(L) не адекватны признакам класса и исключаются из рассмотрения. Мы рассматриваем закономерности из SPL_α, которые не вошли ни в одно из множеств Sat(L) какого либо класса $L \in \mathrm{Class}_\alpha(M)$, как полученные в результате переобучения, поэтому мы удаляем их из SPL_α.

Определение 21. *Закономерной моделью класса* $L \in \mathrm{Class}_\alpha(M)$ *будем называть набор* $C = \langle L, Sat(L)\rangle$.

Для «естественных» классов определения порождающего множества, системообразующего набора и систематики из определения 20 остаются без изменений.

11. Распознавание «естественных» классов

Закономерные модели классов позволяют распознавать их на контрольных объектах, выбранных из генеральной совокупности. Для закономерных моделей класса $C = \langle L, Sat(L)\rangle$ определим закономерные матрицы. Для каждой литеры $A_0 \in L$ оценим силу ее предсказания по закономерностям из Sat(L) и литерам из L:

$$Kr_{Sat(\mathrm{L})}(A_0) = \sum_{C \in Sat(L),\ C=(A_1\&\ldots\&A_k \Rightarrow A_0),\ A_1,\ldots,A_k \in L} \nu(C).$$

Определение 22. *Закономерную матрицу класса определим как набор* $M_C = \langle L, \{Kr_{Sat(\mathrm{L})}(A_0)\}_{A_0 \in L}\rangle$.

Предложение 4. *Для любого класса* $C = \langle L, Sat(L)\rangle$ *выполнено равенство*

$$Kr_{SPL_\alpha}(L) = \sum_{A_0 \in L} Kr_{Sat(L)}(A_0).$$

Используя закономерные матрицы можно распознавать новые объекты генеральной совокупности и относить их к «естественным» классам. Определим

оценку принадлежности объекта $b \in B$ к классу $C = \langle L, Sat(L)\rangle$. Для этого надо оценить в какой степени объект $b \in B$ удовлетворяет (подчиняется) закономерностям класса $Sat(L)$.

Определение 23. *Оценкой принадлежности объекта $b \in B$ к классу $C = \langle L, Sat(L)\rangle$ будем называть величину*

$$Score(b/C) = \sum_{A_0 \in (L \cap S_b)} Kr_{Sat(L\cap S_b)}(A_0) - \sum_{A_0 \in L,\ \neg A_0 \in S_b} Kr_{Sat(L\cap S_b)}(A_0).$$

Распознавание по закономерным матрицам осуществляется так же, как и по весовым матрицам [19]. Для этого надо для каждого класса подсчитать $Score(b/C)$ всех объектов позитивной и негативной выборки и вычислить значение порога, которое дает требуемые значения ошибок первого и второго рода. После этого можно распознавать новые объекты генеральной совокупности и относить их к классу, если значение $Score(b/C)$ этого объекта выше порога.

12. Классификация нуклеотидных последовательностей ДНК.

Данный подход был апробирован на задаче классификации и распознавания сайтов связывания транскрипционных факторов [22], которую мы далее приводим. Позиционно-весовые матрицы являются наиболее часто используемым методом распознавания участков связывания транскрипционных факторов. Вместо весовых матриц мы будем использовать закономерные матрицы.

Рассмотрим выборку сайтов $\mathrm{D} = \{a_1, \ldots, a_m\}$, представленную выровненными последовательностями нуклеотидов длины n, где в i-ой позиции стоит признак $\mathrm{x}_i(a) \in \{\mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}\}$, $\mathrm{i} = 1, \ldots, \mathrm{n}$, $a \in D$. Для описания классов будем использовать наборы значений признаков $x_{i_1}(a), \ldots, x_{i_m}(a)$, $i_1 < \ldots < i_m$. Они могут быть представлены матрицами. Например, последовательность

$$[A][A][C][A][G][C][T][A][C][A][G][G][T][A][A][G][G][G][G][C][T]$$

можно представить матрицей.

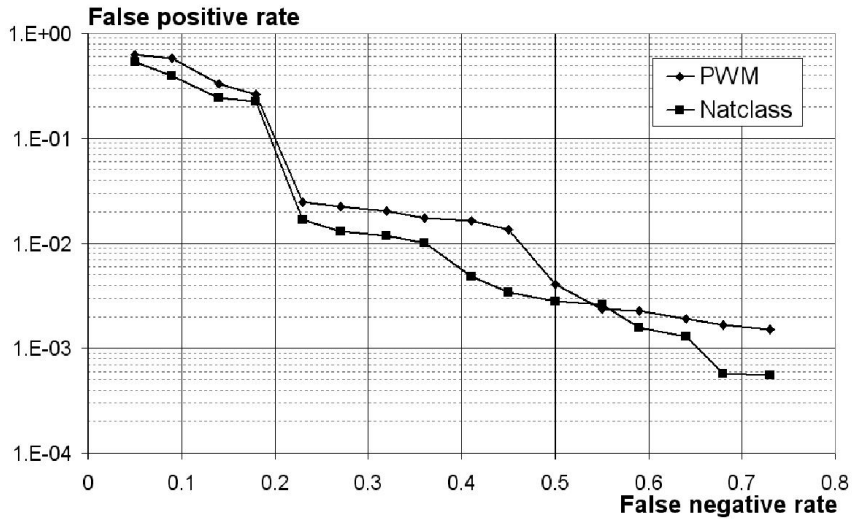
A	1	1	0	1	0	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0
T	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
G	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	1	1	1	0	0
C	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0

Определим предикаты и литеры $P_{ij}^{\varepsilon}(a) = (x_i(a) = j)^{\varepsilon}$, $j \in \{A, T, G, C\}$, где $\varepsilon = 1(0)$, если предикат не имеет (имеет) отрицание. Тогда выборкой из генеральной совокупности будет модель $B = \langle D,\ W\rangle$, где W – множество всех предикатов.

На выборке B обнаружим закономерности SPL_α, классы $\mathrm{Class}_\alpha(M)$ и их закономерные модели. Каждый класс представим двояким образом: матрицей предикатов и закономерной матрицей. Матрица предикатов для некоторого класса $\mathrm{Class}_\alpha(M) = \{P_{ij}^{\varepsilon}\}$ определяется аналогично матрице последовательности и в ней для каждого предиката P_{ij}^{ε} в столбце i и строке j стоит значение $1(0)$, если $\varepsilon = 1(0)$, остальные клетки пусты. Закономерной матрицей для закономерной модели класса $C = \langle L, Sat(L)\rangle$, $L = \{P_{ij}^{\varepsilon}\}$ будет матрица $M_C = \langle L, \{Kr_{Sat(\mathrm{L})}(A_0)\}_{A_0 \in L}\rangle$, в которой для каждого предиката $P_{ij}^{\varepsilon} \in L$ в столбце i и строке j стоит значение $Kr_{Sat(\mathrm{L})}(P_{ij}^{\varepsilon})$.

Для анализа и распознавания были использованы выборки сайтов EGR1 (early growth response factor 1), SF1 (steroidogenic factor-1) и STEBP (sterol regulatory element binding protein). Обучающие выборки была взята из базы данных TRRD [17].

Данные сайта EGR1 состояли из 22 нуклеотидных последовательностей. Ввиду ограниченности данных, точность распознавания системой «естественной» классификации NatClass по сравнению с оптимизированными весовыми матрицами (positional weight matrix – PWM) была осуществлена скользящим контролем [19]. Для оптимизации весовых матриц мы запускали их на последовательностях разной длинны, как описано в [19]. В результате мы определили, что длина 10 последовательности дает наибольшую точность прогноза. Мы подготовили позитивную обучающую выборку EGR1 длинны 10. При скользящем контроле в качестве обучающих данных бралась 21 последовательность и одна оставалась на контроль. Обученный метод применялся к оставшейся последовательности и вычислялась оценка ложнопозитивных оценок. Для контроля было случайно сгенерировано 100 000 последовательностей с теми же частотами появления нуклеотидов, что и в исходных последовательностях. Результаты эксперимента отображены на рис. 2, причем закономерные матрицы дают лучшие результаты, чем весовые матрицы при любых уровнях ошибки первого/второго рода.



Весовые матрицы содержат неявное предположение о независимости вклада каждой нуклеотидной позиции в связующую способность сайта. В некоторых работах подчеркивается [11-12], что нуклеотиды в сайтах связывания не могут считаться независимыми. Это предположение неверно и противоречит биологическим моделям. В отличии от весовых матриц система «естественной» классификации NatClass обнаруживает множество зависимостей между нуклеотидами (и отрицанием нахождения их в определенном месте). Детальный анализ закономерностей позволяет обнаружить существенные для связывания позиции.

После подсчета точности мы запустили программу NatClass на всем множестве в 22 последовательности с теми же параметрами, что и при определении

точности. Система обнаружила множество SPL_α состоящее из 2354 закономерностей и в точности один класс [G][C][G][G][G][G][G][C][G][G], покрывающий всю позитивную выборку.

В качестве примера закономерности рассмотрим следующее правило:

(1=g & 7=g & 8=c) ⇒ {3=-c (0.96) & 3=-a (0.84) & 3=g (0.56) & 3=t (0.28)}

Здесь условие правила содержит конъюнкцию нуклеотидов или их отрицаний в определенных позициях (отрицание означает, что нуклеотид не должен находиться в данной позиции). Заключение правила содержит указание предсказываемой позиции и перечень нуклеотидов и их вероятностей, с которой они могут находиться в этой позиции. Приведенное правило означает, что если нуклеотид g находится в первой и седьмой позициях, а нуклеотид c в восьмой, то в третьей позиции не должен стоять нуклеотид «c» с вероятностью 0.96, не должен стоять нуклеотид «a» с вероятностью 0.84, должен стоять нуклеотид «g» с вероятность 0.56 и должен стоять нуклеотид «t» с вероятностью 0.28.

Из 2354 закономерностей только 2032 описывают данный класс и входят во множество $Sat(L)$ данного класса (мы считаем, что каждое вышеприведенное правило описывает четыре закономерности). Среди них 78 закономерностей предсказывает «c» в восьмой позиции, 69 закономерностей предсказывает «c» во второй позиции, 42 закономерностей предсказывает «g» в первой позиции и 28 закономерностей предсказывает «g» в шестой позиции (см. таблицу 3). Кроме того, из 78-и закономерностей, предсказывающих нуклеотид «c» в восьмой позиции, 51 закономерность (65,4%) содержит «g» в пятой позиции. В таблице 3 показаны наиболее часто обнаруживаемые зависимости между нуклеотидами и их отрицаниями в определенных позициях.

Тип закономерности	Количество
Все	2032
. . . ⇒8=c	78
. . . ⇒2=c	69
. . . ⇒1=g	42
. . . ⇒6=g	28
. . . & 5=g &. . . ⇒8=c	51 (65,4%)
. . . & 4=g &. . . ⇒2= c	39 (56,5%)
. . . & 5=g & 7=g . . . ⇒8=c	17 (21,8%)
. . . & 5=-t &8=c . . . ⇒6=g	13 (46,4%)
. . . & 3=-a &6=g . . . ⇒1=g	10 (23,8%)

Таблица 3. Описание класса для сайта EGR1 в виде основе закономерной матрицы

Последовательность [G][C][G][G][G][G][G][C][G][G] класса максимизирует критерий Kr_{SPL_α} и соответствует консенсусу сайта EGR1, что вместе с закономерностями хорошо согласуется с биологическими данными.

Для сайтов SF1 и SREBP были выбраны соответственно 54 и 38 последовательности. Оптимальные длины последовательностей для метода весовых матриц составили 13 (для SF1) и 18 (для SREBP) нуклеотидов. Были подготовлены

позитивные выборки для сайтов SF1, SREBP соответствующей длинны. Сравнение точности системы NatClass и весовых матриц проводилась бутстрепом в соответствии с работой [14]. Для этого 15 раз генерировалась позитивная выборка, включающих 90% позитивных образцов. Остальные 10% были использованы для контроля. Для каждого контроля рассчитывались закономерные матрицы и по ним оценивались значения $Score(b/C)$ ложнопозитивных (ЛП) прогнозов на случайно сгенерированных выборках (с учетом частот нуклеотидов) достаточно большого объёма (порядка 1 000 000 последовательностей). Далее мы ранжировали объединенное множество контрольных сайтов в соответствии со значениями ЛП. Таблица 4 показывает уровни ЛП при уровне ложнонегативных срабатываний 50%.

Таблица 4. Точность методов естественной классификации и весовых матриц на SF1 и SREBP.

Сайт	Элементов в выборке	Негативных случайных элементов на каждую итерацию	Число закономерностей в классе	Ложно-негативные прогнозы	Ложно-позитивные прогнозы (NatClass)	Ложно-позитивные прогнозы (PWM)
SF1	54	1 000 000	1670	27	2e-005	6.87e-5
SREBP	38	100 000	789	19	0/110000 $<$ 1E-005	8.32e-4

Для сайта SF1 был найден класс [T/C][C][A][A][G][G][T/C][C][A][G]. Здесь T/C означает возможность появления обоих нуклеотидов в данной позиции.

References

[1] Е.Е. Витяев, *Классификация как выделение групп объектов, удовлетворяющих разным множествам согласованных закономерностей*, Выч. сист., **99** (1983), 44–50. MR0784596

[2] Е.Е. Витяев, Н.С. Морозова, А.С. Сутягин, К.А. Лапардин, *Естественная классификация и систематика как законы природы*, Вычислительные системы, **174** (2005), 80–92.

[3] Е.Е. Витяев, В.С. Костин, *Естественная классификация как закон природы*, Интелектуальные системы и методология, Материалы научно-практического симпозиума "Интелектуальная поддержка деятельности в сложных предметных областях", **4** (1992), 107–115.

[4] Е.Е. Витяев, В.В. Мартынович, *Вероятностные формальные понятия на контекстах с отрицаниями*, Информационные технологии в гуманитарных исследованиях, **19** (2014), 5–20.

[5] В.Ю. Забродин, *О критериях естественной классификации*, НТИ, сер.2,**8** (1981).

[6] В.Л. Кожара, *Функции классификации*, Теория классификаций и анализ данных, Новосибирск, 1982, ч. 1.

[7] С.В. Мейен, С.А. Шрейдер, *Методологические аспекты теории классификаций*, Вопросы философии, **12** (1976).

[8] Л. Рутковский, *Элементарный учебник логики*, Спб., 1884.

[9] Е.С. Смирнов, *Конструкция вида таксономической точки зрения*, Зоол. Журн., **17**:3 (1938), 387–418.

[10] С.А. Шрейдер, *Систематика, типологии, классификация*, В кн.: Теория и методология биологических классификаций, М.: Наука, 1983.

[11] Y. Barash, G. Elidan, F. Friedman, and T. Kaplan, *Modeling dependencies in protein-DNA binding sites*, RECOMB, (2003), 28–37.

[12] P.V. Benos, M.L. Bulyk, G.D. Stormo, *Additivity in protein-DNA interactions: how good an approximation is it?*, Nucleic Acids Res., **30** (2002), 4442–4451.

[13] *Classification and Clustering* (1977), Ed. By J. Van Ryzin, Academic Press, Stateplace, New York. MR0443145

[14] B. Efron and G. Gong, *A leisurely look at the bootstrap the jackknife and resampling*, American Statistician, **37**:1 (1983), 36–48. MR0694281

[15] J.Y. Halpern, *An analysis of first-order logic of probability*, Artificial Intelligence, **46**:3 (1990), 311–350. MR1084887

[16] C.G. Hempel, *Maximal Specificity and Lawlikeness in Probabilistic Explanation*, Philosophy of Science, **35** (1968), 116–133.

[17] N.A. Kolchanov, E.V. Ignatieva, E.A. Ananko, O.A. Podkolodnaya, I.L. Stepanenko, T.I. Merkulova, M.A. Pozdnyakov, N.L. Podkolodny, A.N. Naumochkin, A.G. Romashchenko, *Transcription Regulatory Regions Database (TRRD): its status in 2002*, Nucleic Acid Res., **30** (2002), 312–317.

[18] B. Kovalerchuk, E. Vityaev, *Data Mining in finance: Advances in Relational and Hybrid Methods*, (2000), Kluwer Academic Publishers. Zbl 0944.91027

[19] V. Levitsky, E. Ignatieva, G. Vasiliev, N. Limova, T. Busygina, T. Merkulova, N. Kolchanov, *The SiteGA tool for recognition and context analysis of transcription factor binding sites: significant dinucleotide features besides the canonical consensus exemplified by SF-1 binding site*, Bioinformatics of Genome Regulation and Structure II, Springer Science+Business Media, Inc. (2006), 31–41.

[20] E. Vityaev, *The logic of prediction*, Mathematical Logic in Asia, (2006), 263–276. MR2294299

[21] E. Vityaev, B. Kovalerchuk, *Empirical Theories Discovery based on the Measurement Theory*, Mind and Machine, **14**:4 (2004), 551–573.

[22] E.E. Vityaev, K.A. Lapardin, I.V. Khomicheva, A.L. Proskura, *Transcription factor binding site recognition by regularity matrices based on the natural classification method*, Intelligent Data Analysis, **12** (2008), 495–512.

Vityaev Evgenii Evgen'evich
Sobolev Institute of Mathematics,
pr. Koptyuga, 4,
630090, Novosibirsk, Russia
E-mail address: evgenii.vityaev@math.nsc.ru

Martinovich Vitalii Valer'evich
Novosibirsk State University,
Pirogova, 2,
630090, Novosibirsk, Russia
E-mail address: vilco@ya.ru