АНАЛИЗ СТРУКТУРНЫХ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
2005 год Выпуск 174
УДК 518.74

ЕСТЕСТВЕННАЯ КЛАССИФИКАЦИЯ И СИСТЕМАТИКА КАК ЗАКОНЫ ПРИРОДЫ[1]

Е.Е. Витяев[2], Н.С. Морозова[3], А.С. Сутягин[3], К.А. Лапардин[3]

1. Введение. Что такое естественная классификация

Понятие естественной классификации, несмотря на его важность до сих пор не вошло в обиход современной науки. Понятие естественной классификации развивалось в 1970-1980 гг в рамках классификационного движения. В рамках этого направления был систематизирован опыт естествоиспытателей по созданию естественных классификаций, организовано несколько конференций и создана библиография. В данной работе, обобщающей опыт классификационного движения, предлагается формализация понятия естественной классификации.

В рамках классификационного движения В.Ю. Забродин систематизировал критерии «естественности» классификации, которые в различное время выдвигались естествоиспытателями [6]. Приведем эти критерии.

---

[1]Работа выполнена при финансовой поддержке РФФИ (проект № 05-07-90185в), Программой президента РФ для государственной поддержки научных школ (проект НШ-2112.2003.1) и Интеграционным проектом СО РАН (проект № 119).

[2]vityaev@math.nsc.ru; Институт математики им. С.Л. Соболева СО РАН

[3]Новосибирский государственный университет

1. Смирнов Е.С. [13]: «Таксономическая проблема заключается в "индикации": от бесконечно большого числа признаков нам нужно перейти к ограниченному их количеству, которое заменило бы все остальные признаки».

2. Рутковский Л. [12]: «Чем в большем числе существенных признаков сходны сравниваемые предметы, тем вероятнее их одинаковость и в других отношениях».

3. Уэвель В. (ссылку см. в работе[10]): «Чем больше общих утверждений об объектах дает возможность сделать классификация, тем она естественней».

4. Любищев А.А. [6]: «Наиболее совершенной системой является такая, где все признаки объекта определяются положением его 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 систем, является максимальным (в идеале это все его свойства)».

Участники классификационного движения по инициативе инициатора движения Кожара В.Л. [7,8] также дали некоторые определения естественной классификации:

5. Забродин В.Ю. (ссылку см. в работе [6]): «Естественной является та, и только та классификация, которая выражает закон природы».

6. Шрейдер С.А. [14]: «В многообразии объектов, образующих естественную классификацию, можно обнаружить два типа закономерностей:

• соотношения, связывающие "короткое" описание архетипа, достаточное для диагностирования принадлежности объекта к данному классу, с "полным" описанием. В сущности, эти законы, позволяющие на основании принадлежности объекта к некоторому естественному классу прогнозировать все его свойства;

• 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. Витяев Е.Е. [2,4]: «Разбиение на классы должно производиться так, чтобы объекты одного класса подчинялись одним и тем же закономерностям. Между классами существуют закономерности перехода от класса к классу. Объекты класса, кроме того, должны обладать некоторой целостностью. Целостность — взаимная согласованность закономерностей класса по взаимному предсказанию свойств объектов».

Далее мы введем определение естественной классификации и систематики объясняющее перечисленные выше свойства естественной классификации.

2. Онтологии и описание предметной области

В последнее время внимание различных исследователей привлекают онтологии. Это понятие заимствованно из философии. Точного определения этого понятия для задач искусственного интеллекта до сих пор нет. Емкое определение онтологии дал Thomas R. Gruber [15] как спецификацию концептуализации. Неформально онтология представляет собой описание предметной области. Оно состоит из системы понятий и определений новых понятий, описания предмета и методов исследования и априорного знания об объектах и методах исследования [11, 16].

Построение онтологий предполагает концептуализацию предметной области, которая включает в себя систему понятий и величин, а также систему законов аналитических и синтетических, связывающих между собой понятия и величины. Понятие естественной классификации предполагает заданной некоторую онтологию. Приведем определение онтологии необходимое для введения понятия естественной классификации.

Онтология состоит из:

1) системы понятий предметной области, которая

– содержит систему взаимосвязанных понятий, определяющих предмет рассмотрения и цели исследования и что именно интересует нас в объектах предметной области;

– содержит потенциально бесконечное множество признаков, величин (оснований), характеризующих объекты;

2) системы законов предметной области, включающей:

– аналитические выражения, фиксирующие связь понятий;

– законы, например, физические, устанавливающие взаимосвязь величин;

– множество индуктивных законов (закономерностей), устанавливающих взаимосвязи между потенциально бесконечным множеством признаков, характеристик (оснований) объектов предметной области.

Аналитические выражение являются априорными. Индуктивные зависимости могут быть явно представлены в системе законов предметной области или могут быть обнаружены некоторым методом Data Mining на множестве объектов предметной области. Аналитические выражения имеют статус определений и могут рассматриваться как аксиомы предметной области. Закономерности тоже могут быть выражены в виде некоторых логических утверждений и имеют некоторую дополнительную характеристику своей выполнимости — вероятности, достоверности и т.д.

Объекты предметной области являются целостными образованьями, соединяющими в себе понятия из системы понятий и законы из системы законов проблемной области. Поэтому законы из системы законов выполнены (с некоторой степенью вероятности, достоверности и т.д.) на объектах предметной области.

Если на систему законов смотреть как на систему аксиом предметной области, сформулированную в системе понятий, которой должны удовлетворять объекты предметной области, то объекты являются объектами-моделями системы аксиом. Совокупность всех таких объектов-моделей системы аксиом дает картину всех возможных объектов предметной области в данной системе понятий и позволяет предсказывать существование новых объектов, удовлетворяющих системе аксиом.

3. Определение естественной классификации и систематики

Определим модель M_a объекта a. В нее входит множество Ω_a значений всех понятий, признаков, характеристик и величин, которые применимы к объекту и принимают на нем определенные значения (истинности, числовые). Выделим из системы законов предметной области подмножество Z_a законов и закономерностей, которые применимы к данному объекту. Это будут не

все закономерности системы законов. Например, закономерности вида IF...THEN... не применимы к объекту, если посылка правила не выполнена на объекте. Подмножество Z_a дает закономерную структуру объекта. Модель $M_a = \langle \Omega_a, Z_a \rangle$ назовем закономерной моделью объекта.

Рассмотрим некоторый класс $\mathfrak{C}$ объектов. Определим *закономерную модель класса* $M_{\mathfrak{C}} = \langle \Omega_{\mathfrak{C}}, Z_{\mathfrak{C}} \rangle$ как пересечение всех закономерных моделей объектов класса $\mathfrak{C}$.

Проанализируем критерий Е.С. Смирнова [13]. Разнообразие классов всегда несопоставимо меньше разнообразия комбинаций значений признаков и, следовательно, между значениями признаков должно существовать огромное количество закономерных связей. Если число классов составляет, например, сотни, а признаки бинарные, то независимыми среди них могут быть только около 10 признаков: $1024 = 2^{10}$. При классификации животных, растений, почв и т.д. естествоиспытатели могут использовать огромное, потенциально бесконечное, множество признаков и характеристик. Но среди них только десяток признаков может быть в известной степени независим, а остальные признаки связаны между собой закономерностями так, что из десятка признаков предсказываются значения всех остальных признаков. Найти признаки, из которых предсказываются все остальные и составляет проблему индикации. Такими значениями признаков в закономерной модели класса $M_{\mathfrak{C}}$ являются порождающие совокупности значений признаков. По закономерностям из $Z_{\mathfrak{C}}$ и набору значений порождающих признаков $\langle x_{i_1} = x_{i_1 j_1},\ x_{i_2} = x_{i_2 j_2}, \ldots, x_{i_m} = x_{i_m j_m} \rangle$, где $x_{i_1 j_1}, x_{i_2 j_2}, \ldots, x_{i_m j_m}$ — значения признаков $x_{i_1}, x_{i_2}, \ldots, x_{i_m}$, мы можем предсказать все остальные значения признаков $\Omega_{\mathfrak{C}}$. Понятно, что набор значений порождающих признаков определяется неоднозначно.

Предположим, что все классы $\{\mathfrak{C}_{i \in I}\}$ нам известны и мы знаем все закономерные модели этих классов $M_{\mathfrak{C}_i}$. Рассмотрим задачу построения систематики. Будем искать такие порождающие наборы признаков $x_{i_1 j_1}, x_{i_2 j_2}, \ldots, x_{i_N j_N}$, что для каждого класса $\{\mathfrak{C}_{i \in I}\}$ набор значений признаков $\langle x_{i_1} = x^{i_1}_{j_1},\ x_{i_2} = x^{i_2}_{j_2}, \ldots, x_{i_N} = x^{i_N}_{j_N} \rangle$ является порождающим. Набор призна-

ков $S = \langle x_{i_1}, x_{i_2}, \ldots, x_{i_N}$ будем называть *системообразующим*, если для каждого класса из $\{\mathfrak{C}_{i \in I}\}$ значения порождающего набора признаков $\langle x_{i_1} = x_{j_1}^{i_1}, x_{i_2} = x_{j_2}^{i_2}, \ldots, x_{i_N} = x_{j_N}^{i_N} \rangle$ различны. В этом случае каждый класс будет однозначно определяться набором значений системообразующих признаков. Понятно, что наборы системообразующих признаков также определяются неоднозначно. Задача и состоит в том, чтобы найти наиболее компактный и информативный набор системообразующих признаков. В работах Загоруйко Н.Г. и Борисовой И.А. [1, 17] также ставится задача нахождения минимального множества «существенных» признаков.

Систематика состоит в том, чтобы представить некоторым образом, например, таблицей, как изменяются наборы 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, в следствии чего систематику можно представить некоторым специальным образом, чтобы этот закон был виден наглядно. 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 признаков из S при переходе от класса к классу. Каждому набору значений системообразующих признаков S соответствует некоторый класс $M_{\mathfrak{C}} = \langle \Omega_{\mathfrak{C}}, Z_{\mathfrak{C}} \rangle$. Тогда закон систематики Z_S является метазаконом по отношению к закономерностям класса $Z_{\mathfrak{C}}$. Закон систематики Z_S связан с законами классов как это определено в определении данном С.А. Шрейдером [14]. Закономерностями первого типа являются закономерности соответствующего класса $Z_{\mathfrak{C}}$, а закономерностями второго типа — закон семантики Z_S.

Рассмотрим критерий А.А. Любищева [6]. Системой по Любищеву является такое представление классификации объектов, где по месту объекта в системе определяются все его признаки. В нашем определении значения признаков некоторого объекта определяются взаимодействием двух законов — сначa-

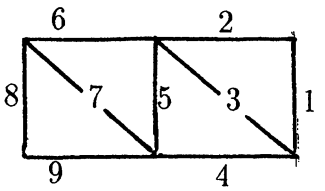
ла законом семантики Z_S, используя который мы по положению объекта в системе можем определить значения системообразующих признаков и класс, к которому принадлежит этот объект, и далее по значениям системообразующих признаков этого класса и по закономерностям класса $Z_{\mathfrak{C}}$ мы можем определить все остальные свойства объекта.

Определим *систематику* как набор $\Sigma = \langle S, Z_S, \{Z_{\mathfrak{C}_i}\}_{i\in I}\rangle$. Не все закономерности системы законов предметной области будут входить во множества закономерностей $Z_S, \{Z_{\mathfrak{C}_i}\}_{i\in I}$, так как эти множества зависят от выбора порождающих признаков. Задача и состоит в том, чтобы выбрать наиболее совершенную систему объясняющую свойства и строение объектов простейшим образом. Систематика как закон природы определяется набором $\langle S, Z_S, \{Z_{\mathfrak{C}_i}\}_{i\in I}\rangle$.

Предположим теперь, что нам неизвестно разбиение объектов на классы. Тогда систематику надо строить по закономерным моделям объектов, а не классов. Задача построения систематики сводится в этом случае к нахождению такого разбиения множества объектов на классы, чтобы построенная на этих классах систематика была наиболее совершенной.

4. Пример построения систематики. Распознавание цифр индекса

Рассмотрим цифры индексы как набор из 9 объектов. Предикат P_i означает наличие i-го элемента в начертании цифры. Занумеруем признаки следующим образом:



Тогда данные удобно представить в виде табл. 1.

Таблица 1

Значения признаков цифр

Цифра	Признаки цифр								
	$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$	$\mathbf{P_4}$	$\mathbf{P_5}$	$\mathbf{P_6}$	$\mathbf{P_7}$	$\mathbf{P_8}$	$\mathbf{P_9}$
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

Будем рассматривать цифры как классы $\{\mathfrak{C}_{i\in I}\}$, $I = \{0,\ldots,9\}$. Найдем закономерные модели этих классов. Для этого будем искать импликативные детерминированные закономерности. Рассмотрим $M = \{A, Q\}$ — модель сигнатуры $\Omega = \{P_1,\ldots,P_9\}$, где A — генеральная совокупность объектов; $Q = \{P_i,\ldots,P_9\}$ — множество предикатов сигнатуры Ω, заданных на A; P_i, $i = 1,\ldots,9$, — предикатные символы сигнатуры Ω.

Импликативной детерминированной закономерностью [3, 5] назовем истинную на A формулу вида $F = (P_{i_1}^{\varepsilon_1}(a)\ \&\ldots\ldots\&\ P_{i_m}^{\varepsilon_m}(a) \Rightarrow P_{i_0}^{\varepsilon_0}(a))$, где $\{P_{i_1},\ldots,P_{i_m},P_{i_0}\} \subset \{P_1,\ldots,P_9\}$, $\varepsilon = 1(0)$, если отношение берется без отрицания (с отрицанием), удовлетворяющую следующим условиям:

а) среди атомарных отношений $P_{i_1}^{\varepsilon_1}(a),\ldots,P_{i_m}^{\varepsilon_m}(a), P_{i_0}^{\varepsilon_0}(a)$ нет повторений и нет одновременно отношения и его отрицания;

б) если из конъюнкции $P_{i_1}^{\varepsilon_1}(a)\ \&\ \ldots\ \&\ P_{i_m}^{\varepsilon_m}(a)$ удалить одно из отношений, либо заменить отношение $P_{i_0}^{\varepsilon_0}(a)$ на 0 (ложь), то полученная формула станет ложной на A.

Найдем все импликативные детерминированные закономерности для цифр $I = \{0, \ldots, 9\}$. Получим 3743 закономерности, найденные программой в табл. 1.

Далее для каждого класса выделим закономерности, которые на нем выполняются. Например, для цифры 2 будут выполнены 529 закономерности.

По табл. 1 (набор значений признаков) и набору закономерностей можно получить закономерную модель класса. Выделим для каждого класса минимальные определяющие совокупности.

Для двойки это будет, например, совокупность $\{P_2, P_3\}$. Значения остальных признаков восстанавливаются по следующим закономерностям:

$\neg P_3 \;\&\; P_2 \Rightarrow P_1,$
$\neg P_3 \;\&\; \neg P_2 \;\&\; P_1 \Rightarrow P_4,$
$P_4 \;\&\; \neg P_2 \;\&\; P_1 \Rightarrow \neg P_5,$
$\neg P_3 \;\&\; \neg P_2 \;\&\; P_1 \Rightarrow \neg P_6,$
$\neg P_6 \;\&\; \neg P_5 \;\&\; P_4 \;\&\; P_1 \Rightarrow P_7,$
$P_7 \;\&\; \neg P_3 \;\&\; P_1 \Rightarrow \neg P_8,$
$P_8 \;\&\; \neg P_6 \;\&\; \neg P_5 \;\&\; \neg P_2 \Rightarrow P_9.$

Как уже упоминалось, определяющие совокупности выделяются не единственным образом, например, $\{P_5, P_7\}$ тоже будет определяющей совокупностью, для которой значения остальных признаков восстанавливается по следующим закономерностям:

$P_7 \Rightarrow P_1,$
$P_7 \;\&\; \neg P_5 \Rightarrow \neg P_2,$
$P_7 \;\&\; \neg P_5 \Rightarrow P_4,$
$P_4 \;\&\; \neg P_2 \;\&\; P_1 \Rightarrow \neg P_3,$
$\neg P_3 \;\&\; \neg P_2 \Rightarrow P_9,$
$P_4 \;\&\; \neg P_2 \Rightarrow \neg P_6,$
$P_9 \;\&\; \neg P_6 \;\&\; P_4 \Rightarrow \neg P_8.$

Глядя на закономерности видим, что в порождающих $\{P_5, P_7\}$ закономерная модель двойки проще. Она будет выглядеть следующим образом: $M_2 = \langle \Omega_2, Z_2 \rangle = \{\{1,0,0,1,0,0,1,0,1\}, \{P_7, \neg P_5, \quad P_7 \Rightarrow P_1, \quad P_7 \;\&\; \neg P_5 \Rightarrow \neg P_2, \quad P_7 \;\&\; \neg P_5 \Rightarrow P_4, P_4 \;\&\; \neg P_2 \;\&\; P_1 \Rightarrow \neg P_3, \quad \neg P_3 \;\&\; \neg P_2 \Rightarrow P_9, \quad P_4 \;\&\; \neg P_2 \Rightarrow \neg P_6,$

P_9 & $\neg P_6$ & $P_4 \Rightarrow \neg P_8\}\}$. По минимальной определяющей совокупности каждой цифры мы можем построить ее закономерную модель.

Перейдем к построению закономерной модели систематики. Ее закон Z_S представим в виде таблицы, в каждой строке которой стоят названия классов и значения признаков. Для выбора минимальной определяющей совокупности семантики рассмотрим различные сочетания определяющих совокупностей классов.

Максимальная по количеству признаков минимальная определяющая совокупность у цифры 8 (минимальное количество признаков) равно 3. Значит, определяющая совокупность систематики состоит не меньше, чем из трёх признаков. Минимальные определяющие совокупности классов не всегда позволяют выявить минимальную совокупность систематики. Например, минимальные определяющие совокупности тройки — это $\{P_3, P_7\}$, $\{\neg P_4, P_7\}$, тогда как определяющие совокупности, состоящие из трех признаков, для этого же класса не содержат седьмого признака. Следовательно, стоит рассматривать не только все определяющие совокупности длины 2, но и определяющие совокупности длиной не более 3 признаков для каждого класса.

Так как $2^3 = 8$ меньше, чем число классов, то трех признаков будет недостаточно для однозначного восстановления класса. Поэтому рассматриваем все возможные комбинации из четырех признаков. В результате получим, что минимальная определяющая совокупность признаков для систематики это $\{P_4, P_5, P_6, P_7\}$. В этом случае она определяется единственным образом.

Систематика для классов цифр индекса — это закон систематики, представленный табл. 2, а так же наборы закономерностей для каждого класса и набор признаков класса.

Таблица 2

Систематика цифр

Цифры	Значения порождающих признаков				Порождающие совокупности
	P_4	P_5	P_6	P_7	
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 признаков определяется класс. С помощью порождающих совокупностей, для каждого класса восстанавливаются значения всех остальных признаков.

Литература

1. БОРИСОВА И.А., ЗАГОРУЙКО Н.Г. Естественная классификация //Сб. тр. 4-го российско-украинского сем. "Интеллектуальный анализ информации" (ИАИ-2004), Киев, 19-21 мая 2004 г. — Киев. — 2004. — С. 33-42.

2. ВИТЯЕВ Е.Е. Классификация как выделение групп объектов, удовлетворяющих разным множествам согласованных закономерностей //Анализ разнотипных данных. — Новосибирск, 1983. — Вып. 99: Вычислительные системы. — С. 44-50.

3. ВИТЯЕВ Е.Е. Метод обнаружения закономерностей и метод предсказания // Эмпирическое предсказание и распознавание образов. — Новосибирск, 1976. — Вып. 67: Вычислительные системы. — С. 54-68.

4. ВИТЯЕВ Е.Е., КОСТИН В.С. Естественная классификация как закон природы //Интеллектуальные системы и методология (Материалы научно-практического симпозиума "Интеллектуальная поддержка деятельности в сложных предметных областях"), вып. 4. — Новосибирск, 1992. — С. 107–225.

5. ВИТЯЕВ Е.Е. Семантический подход к созданию баз данных. Семантический вероятностный вывод наилучших для предсказания ПРОЛОГ-программ по вероятностной модели данных //Логика и семантическое программирование. — Новосибирск, 1992. — Вып. 146: Вычислительные системы. — С. 19–49.

6. ЗАБРОДИН В.Ю. О критериях естественной классификации. //Научно-техническая информация. — 1981. — Сер. 2, № 8. — С. 22-24.

7. КОЖАРА В.Л. Анализ информативно насыщенных таксономических структур как способ выявления географических закономерностей //Дисс. канд. геогр. наук. — М., 1989.

8. КОЖАРА В.Л. Функции классификации //Теория классификаций и анализ данных. — Новосибирск, 1982. Ч. 1.

9. МАЛЬЦЕВ А.И. Алгебраические системы. — М.: Наука. — 1970.

10. МЕЙЕН С.В., ШРЕЙДЕР С.А. Методологические аспекты теории классификации //Вопросы философии. — 1976. — № 12.

11. РОССЕЕВА О.И., ЗАГОРУЛЬКО Ю.А., Сергеев И.П. Организация эффективного поиска на основе онтологий. http://www.dialog-21.ru/Archive/2001/vilume2/2 49.htm

12. РУТКОВСКИЙ Л. Элементарный учебник логики. — Санкт-Петербург, 1884.

13. СМИРНОВ Е.С. Конструкция вида таксономической точки зрения //Зоол. журн. — 1938. — Т. 17, № 3. — С. 387–418.

14. ШРЕЙДЕР С.А. Систематика, типология, классификация //Теория и методология биологических классификаций. — М.: Наука. — 1983.

15. GRUBER Thomas R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing //International Workshop on Formal Ontology. — 1993. March, Padova, Italy.

16. CHRISTOPHER W., GUARINO N. Supporting ontological analysis of taxonomic relationships //Data & Knowledge Engineering. — 2001. — Vol. 39, № 1. — P. 51–74.

17. ZAGORUIRO N., BORISOVAI. Principles of natural classification //Pattern Recognition and Image Analysis. — 2005. — Vol. 15, № 1. — P. 27–29.

Поступила в редакцию
19 сентябрь 2005 года