

ЕСТЕСТВЕННАЯ КЛАССИФИКАЦИЯ И СИСТЕМАТИКА КАК ЗАКОНЫ ПРИРОДЫ

Витяев Е.Е.¹, Морозова Н.С.², Сутягин А.С.², Лапардин К.А.²

¹Институт математики им. С.Л.Соболева СО РАН, vityaev@math.nsc.ru

²Новосибирский государственный университет

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 разное время выдвигались естествоиспытателями [4]. Приведем эти критерии.

1. Смирнов Е.С. [11]: “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. [10]: “Чем в большем числе существенных признаков сходны сравниваемые предметы, тем вероятнее их одинаковость и в других отношениях”;
3. Уэвель В. [10]: “Чем больше общих утверждений об объектах дает возможность сделать классификация, тем она естественней”;
4. Любищев А.А. [4]: “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 системе, является максимальным (в идеале это все его свойства)”.

Участники классификационного движения по инициативе инициатора движения Кожара В.Л. также дали некоторые определения естественной классификации:

5. Забродин В.Ю. [4]: “Естественной” является та, и только та классификация, которая выражает закон природы”;
6. Шрейдер С.А. [12]: “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, это 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. [1,3]: “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 [13] как спецификацию концептуализации. Неформально онтология представляет собой описание предметной области. Такое описание, обычно и называемое концептуализацией, состоит из системы понятий и определений новых понятий, описания предмета и методов исследования и априорного знания об объектах и методах исследования [9].

Построение онтологий предполагает концептуализацию предметной области (ПрОбл), 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 заданную некоторую онтологию. Приведем определение онтологии необходимое для введения понятия естественной классификации.

Онтология состоит из:

Системы понятий ПрОбл, которая:

- содержит систему взаимосвязанных понятий, определяющих предмет рассмотрения и цели исследования и что именно интересует нас в объектах ПрОбл;
- содержит потенциально бесконечное множество признаков, величин (оснований) характеризующих объекты.

Системы законов ПрОбл, 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 ПрОбл.

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

Объекты ПрОбл являются целостными образованиями, соединяющими в себе понятия из Системы понятий и законы из Системы законов ПрОбл. Поэтому законы из Системы законов выполнены (с некоторой степенью вероятности, достоверности и т.д.) на объектах ПрОбл.

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

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 ПрОбл подмножество Z_a , законов и закономерностей, которые применимы к данному объекту. Это будут не

все закономерности Системы законов ПрОбл. 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 Класа* $M_{\mathfrak{C}} = \langle \Omega_{\mathfrak{C}}, Z_{\mathfrak{C}} \rangle$ как пересечение всех Закономерных Моделей объектов класса \mathfrak{C} .

Проанализируем критерий Е.С. Смирнова [11]. 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 набору значений порождающих признаков $\langle x_{i1}=x_{i1j1}, x_{i2}=x_{i2j2}, \dots, x_{im}=x_{imjm} \rangle$, где $x_{i1j1}, x_{i2j2}, \dots, x_{imjm}$ – значения признаков $x_{i1j1}, x_{i2j2}, \dots, x_{imjm}$ и закономерностям из $Z_{\mathfrak{C}}$ мы можем предсказать все остальные значения признаков $\Omega_{\mathfrak{C}}$ объектов класса. 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 $x_{i1}, x_{i2}, \dots, x_{iN}$, что для каждого класса из $\{\mathfrak{C}_{i \in I}\}$ набор значений признаков $\langle x_{i1}=x_{i1j1}, x_{i2}=x_{i2j2}, \dots, x_{iN}=x_{iNjN} \rangle$ является порождающим. Набор признаков $S = \langle x_{i1}, x_{i2}, \dots, x_{iN} \rangle$ будем называть *системообразующим*, если для каждого класса из $\{\mathfrak{C}_{i \in I}\}$ значения порождающего набора признаков $\langle x_{i1}=x_{i1j1}^{i1}, x_{i2}=x_{i2j2}^{i2}, \dots, x_{iN}=x_{iNjN}^{iN} \rangle$ 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, что бы 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 метазаконном 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 [12]. Закономерностями первого типа являются закономерности соответствующего класса $Z_{\mathfrak{C}}$, а закономерностями второго типа - закон систематики Z_S .

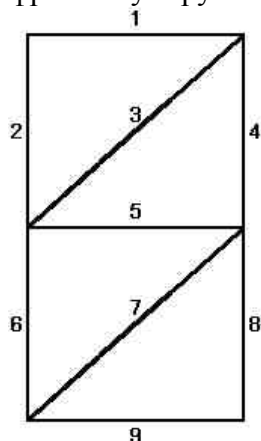
Рассмотрим критерий А.А.Любищева [4]. Системой по Любищеву является такое представление классификации объектов, где по месту объекта в системе определяются все его признаки. В нашем определении значения признаков некоторого объекта определяются взаимодействием двух законов – сначала законом систематики 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 $Z_{\mathcal{C}}$ мы можем определить все остальные свойства объекта.

Определим *Систематику* как набор $\Sigma = \langle S, Z_S, \{Z_{\mathcal{C}i}\}_{i \in I} \rangle$. Не все закономерности из Системы законов ПрОбл будут входить во множества закономерностей $Z_S, \{Z_{\mathcal{C}i}\}_{i \in I}$, т.к. 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 $\langle S, Z_S, \{Z_{\mathcal{C}i}\}_{i \in I} \rangle$.

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, что бы построенная на этих классах систематика была наиболее совершенной и простой.

4. Пример построения систематики. Распознавание цифр индекса.

Рассмотрим цифры индекса 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 таблицы 1:

Таблица 1. Представление цифр значениями признаков.

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	$\{P_4, P_5, P_7\}$

Будем рассматривать цифры как классы $\{C_{i \in I}\}$, $I = \{0, \dots, 9\}$. Найдем закономерные модели этих классов.

Для этого будем искать закономерности в виде импликативных детерминированных закономерностей, определение которых приведено ниже.

Рассмотрим $M = \{A, Q\}$ – модель сигнатуры $\Omega = \{P_1, \dots, P_9\}$, где A – генеральная совокупность объектов; $Q = \{P_1, \dots, P_9\}$ множество предикатов сигнатуры Ω , заданных на A ; P_i , $i=1, \dots, 9$ – предикатные символы сигнатуры Ω .

Импликативной детерминированной закономерностью назовем истинную на A формулу вида

$$F = (P_{i1}^{\varepsilon_1}(a) \& \dots \& P_{im}^{\varepsilon_m}(a) \Rightarrow P_{i0}^{\varepsilon_0}(a)),$$

где $\{P_{i1}, \dots, P_{im}, P_{i0}\} \subset \{P_1, \dots, P_9\}$, $\varepsilon = 1(0)$, если отношение берется без отрицания (с отрицанием), удовлетворяющую следующим условиям:

- а) среди атомарных отношений $P_{i1}^{\varepsilon_1}(a), \dots, P_{im}^{\varepsilon_m}(a), P_{i0}^{\varepsilon_0}(a)$ нет повторений и нет одновременно отношения и его отрицания;
- б) если из конъюнкции $P_{i1}^{\varepsilon_1}(a) \& \dots \& P_{im}^{\varepsilon_m}(a)$ удалить одно из отношений, либо заменить отношение $P_{i0}^{\varepsilon_0}(a)$ на 0 (ложь), то полученная формула станет ложной на A .

Найдем все импликативные детерминированные закономерности для Системы законов Предметной Области $I = \{0, \dots, 9\}$. Получим 3743 закономерности, найденные программой в таблице 1.

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

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

Для 2-ки это будет, например, совокупность $\{P_2, P_3\}$. Значения остальных признаков восстанавливается по следующим закономерностям:

$$\begin{aligned} \neg P_3 \& \neg P_2 &\Rightarrow P_1 \\ \neg P_3 \& \neg P_2 \& P_1 &\Rightarrow P_4 \\ P_4 \& \neg P_2 \& P_1 &\Rightarrow \neg P_5 \\ \neg P_3 \& \neg P_2 \& P_1 &\Rightarrow \neg P_6 \\ \neg P_6 \& \neg P_5 \& P_4 \& P_1 &\Rightarrow P_7 \\ P_7 \& \neg P_3 \& P_1 &\Rightarrow \neg P_8 \\ P_8 \& \neg P_6 \& \neg P_5 \& \neg P_2 &\Rightarrow P_9 \end{aligned}$$

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:

$$\begin{aligned} P_7 &\Rightarrow P_1 \\ P_7 \&\neg P_5 &\Rightarrow \neg P_2 \\ P_7 \&\neg P_5 &\Rightarrow P_4 \\ P_4 \&\neg P_2 \& P_1 &\Rightarrow \neg P_3 \\ \neg P_3 \&\neg P_2 &\Rightarrow P_9 \\ P_4 \&\neg P_2 &\Rightarrow \neg P_6 \\ P_9 \&\neg P_6 \& P_4 &\Rightarrow \neg P_8 \end{aligned}$$

Глядя на закономерности видно, что в порождающих $\{P_5, P_7\}$ закономерная модель двойки проще. Она будет выглядеть следующим образом: $M_2 = \langle \Omega_2, Z_2 \rangle = \{\{1,0,0,1,0,0,1,0,1\}, \{P_7, \neg P_5, P_7 \Rightarrow P_1, P_7 \&\neg P_5 \Rightarrow \neg P_2, P_7 \&\neg P_5 \Rightarrow P_4, P_4 \&\neg P_2 \& P_1 \Rightarrow \neg P_3, \neg P_3 \&\neg P_2 \Rightarrow P_9, P_4 \&\neg P_2 \Rightarrow \neg P_6, P_9 \&\neg P_6 \& P_4 \Rightarrow \neg P_8\}\}$. По минимальной определяющей совокупности каждой цифры мы можем построить ее закономерную модель.

Перейдём к построению закономерной модели систематики. Её закон Z_5 представим в виде таблицы, в каждой строке которой стоят название классов и значения признаков. Для выбора минимальной определяющей совокупности систематики, рассмотрим различные сочетания определяющих совокупностей классов.

Максимальная по количеству признаков минимальная определяющая совокупность у цифры 8 (минимальное количество признаков 3). Значит, определяющая совокупность систематики состоит не меньше, чем из трех признаков. Минимальные определяющие совокупности классов не всегда позволяют выявить минимальную совокупность систематики. Например, минимальные определяющие совокупности 3-ки это $\{P_3, P_7\}$, $\{\neg P_4, P_7\}$, тогда как определяющие совокупности, состоящие из 3 признаков, для этого же класса не содержат 7-го 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 3 признаков будет недостаточно для однозначного восстановления класса. Поэтому рассматриваем все возможные комбинации из 4-х 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 таблицей 2, а так же наборы закономерностей для каждого класса и набор признаков класса.

Таблица 2. Систематика цифр.

	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

По значениям признаков определяется класс. А далее, с помощью минимальных совокупностей, для каждого класса по закономерностям восстанавливаются значения всех остальных признаков.

Благодарности

Эта работа поддержана грантом РФФИ 05-07-90185в, Интеграционным проектом СО РАН №119, Программой президента Российской Федерации поддержки научных школ 2112.2003.1

References

1. Борисова И.А., Загоруйко Н.Г. "Естественная классификация"// Сборник трудов ИАИ-2004, Киев, 2004 г., с 33-42.
2. Витяев Е.Е. Классификация как выделение групп объектов, удовлетворяющих разным множествам согласованных закономерностей. – В кн.: Анализ разнотипных данных (Выч. сист. 99), Новосибирск, 1983, с. 44-50.
3. Витяев Е.Е. Метод обнаружения закономерностей и метод предсказания. – В кн.: Эмпирическое предсказание и распознавание образов (Выч. сист. 67), Новосибирск, 1976, 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. 107-115.
5. Витяев Е.Е. Семантический подход к созданию баз знаний. Семантический вероятностный вывод наилучших для предсказания ПРОЛОГ-программ по вероятностной модели данных. // Логика и семантическое программирование (Выч. сист. 146), Новосибирск, 1992, с.19-49.
6. Забродин В.Ю. О критериях естественной классификации. – НТИ, сер.2, 1981, №8.
7. Кожара В.Л. Анализ информативно насыщенных таксономических структур как способ выявления географических закономерностей // Дисс. Канд. Геогр. Н., 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 классификаций. Вопросы философии, 1976, №12.
11. Россеева О.И., Загоруйко Ю.А., Сергеев И.П. Онтологии как средство организации эффективного поиска в сети Internet.
12. Рутковский Л. Элементарный учебник логики. – СПб., 1884.
13. Смирнов Е.С. Конструкция вида таксономической точки зрения // Зоол. Журн. – 1938, т. 17, №3, с. 387-418.
14. Шрейдер С.А. Систематика, типологии, классификация. – В кн.: Теория и методология биологических классификаций, М., Наука, 1983.
15. Thomas R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing // International Workshop on Formal Ontology. 1993. March, Padova, Italy.
16. Christopher Welty, Nicola Guarino (2001) Supporting ontological analysis of taxonomic relationships, Data & Knowledge Engineering, v.39 n.1, p.51-74
17. Zagoruiiko N., Borisova I. "Principles of natural classification"// Pattern Recognition and Image Analysis, 2005, Vol.15, No.1, p.27-29.