V ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ ЗНАНИЙ

Витяев Е.Е.,
Костин В.С.

Естественная классификация, систематика, онтология[1]

В работе приводится ряд определений «естественной» классификации и систематики, которые давались в разное время естествоиспытатели, а также даются формальные определения этих понятий. Показано, что эти формальные определения охватывают не только «естественную» классификацию и систематику, но и понятия «качество», «количество» и закон перехода количества в качество. Приводятся функции «естественной» классификации, позволяющие решать такие задачи как аналоговое прогнозирование, диагностирование, распространение, структурное прогнозирование, таксономическое прогнозирование и мерономическое прогнозирование. Описывается метод построения «естественных» классификаций, который позволяет решать задачи различных функций «естественной» классификации. Приводится иллюстративный пример построения «естественной» классификации и систематики на примере рукописных индексов. Приводится также пример решения реальной задачи классификации из области биоинформатики.

Ключевые слова: *Data Mining, классификация, естественная классификация, систематика, KDD&DM, интеллектуальный анализ данных.*

Введение. Что такое естественная классификация.

Данная работа является продолжением работ [Витяев Е.Е., 1983; Витяев Е.Е., Костин В.С., 1992; Витяев Е.Е. и др., 2005] посвященных понятию «естественной» классификации (см. [Scientific Discovery website])

В 1970-1980 гг. было организовано классификационное движение, в рамках которого развивалось и анализировалось понятие «*естественной*» *классификации* и был систематизирован опыт естествоиспытателей по созданию «естественных» классификаций. В то время В.Ю. Забродин систематизировал *критерии* "естественности" классификации, которые в различное время выдвигались естествоиспытателями [Забродин В.Ю., 1981]. Приведем эти критерии.

1.1 Смирнов Е.С. [Смирнов Е.С., 1938]: "Таксономическая проблема заключается в "индикации": от бесконечно большого числа признаков нам нужно перейти к ограниченному их количеству, которое заменило бы все остальные признаки";

1.2 Рутковский Л. [Рутковский Л., 1884]: "Чем в большем числе существенных признаков сходны сравниваемые предметы, тем вероятнее их одинаковость и в других отношениях";

1.3 Уэвель В. [Забродин В.Ю., 1981]: "Чем больше общих утверждений об объектах дает возможность сделать классификация, тем она естественней";

1.4 Любищев А.А. [Забродин В.Ю., 1981]: "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 организатора движения Кожара В.Л. также дали некоторые определения «естественной» классификации:

1.5 Забродин В.Ю. [Забродин В.Ю., 1981]: « «Естественной» является та, и только та классификация, которая выражает закон природы»;

1.6 Шрейдер С.А. [Шрейдер С.А., 1983]: «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 к 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

[1] Работа поддержана грантом РФФИ 08-07-00272-а; интеграционными проектами СО РАН №1 и №115, а также работа выполнена при финансовой поддержке Государственного контракта 2007-4-1.4-00-04 и Совета по грантам Президента РФ и государственной поддержке ведущих научных школ (проект НШ-335.2008.1)

объекта на все принадлежащие данному классу и, несколько сложнее, на объекты смежных классов»;

1.7 Витяев Е.Е. [Витяев Е.Е., 1983]: «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. У групп закономерностей могут быть общие закономерности, устанавливающие взаимосвязь признаков объектов из разных классов».

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

2. Онтология.

В последнее время интенсивно развивается понятие *онтологии*. Наиболее цитируемым определением этого понятия является определение Томаса Грубера [Thomas R. Gruber, 1993] *онтология - это спецификация концептуализации*.

Понятие «естественной» классификации предполагает заданной некоторую онтологию. Приведем уточнение понятия онтологии, необходимое для определения «естественной» классификации.

Онтология включает:

2.1 систему понятий;

2.2 аналитические выражения, фиксирующие связь понятий;

2.3 потенциально бесконечное множество признаков, характеристик, величин, характеризующие объекты;

2.4 априорные знания и законы, например, физические, устанавливающие взаимосвязь величин;

2.5 множество индуктивных законов, (закономерностей) устанавливающих взаимосвязи между потенциально бесконечным множеством признаков, характеристик, величин (оснований) объектов.

Аналитические выражения являются априорными. Индуктивные зависимости (закономерности), могут быть обнаружены некоторым методом Data Mining.

3. Основа «естественной» классификации - целостность объектов

Устойчивость целого обеспечивается характерным для него способом организации взаимодействующих частей: "Между частями органичного целого ... существует не простая функциональная зависимость, а значительно более сложная система разнокачественных связей - структурных, генетических, связей субординации, управления и т.п., в рамках которой причина одновременно выступает как следствие. ... Взаимосвязь частей такова, что она выступает не в виде линейного причинного ряда, а в виде своеобразного *замкнутого круга* (выд. Е.Е.), внутри которого каждый элемент связи является условием другого и обусловлен им" ([ФЭС], статья «часть и целое»).

В соответствии с нашим определением 1.7 «разбиение на классы должно производиться так, чтобы объекты одного класса подчинялись одним и тем же закономерностям, объекты разных классов подчинялись разным группам закономерностей» устойчивость целого определяется *системной взаимосвязью закономерностей*, которая специфична для данного целого (класса). Эта системная связь по «замкнутому кругу» определяет взаимосвязь частей и признаков объекта.

Если на закономерности смотреть как на систему аксиом, сформулированную в системе понятий онтологии, а на объекты как на модели этой системы аксиом, то системная связь закономерностей применимых к некоторому объекту (классу объектов) автоматически вытекает из целостности самого объекта (класса объектов). Системная связь закономерностей даёт *структурный закон строения объекта* (класса объектов). В нём «взаимосвязь частей объекта» проявляется в виде структурной закономерной связи по «замкнутому кругу», где каждый элемент объекта (признак, свойство, характеристика) является условием наличия другого элемента объекта.

Совокупность всех таких объектов-моделей для законов онтологии дает картину всех возможных объектов данной онтологии и позволяет предсказывать существование новых объектов, удовлетворяющих системе аксиом.

4. Классы - качественные состояния целостных объектов.

Формой проявления целого всегда является то или иное *качество*. При этом если для целостности объекта безразлично, какая именно система связей его определяет (лишь бы она была устойчивой), то для качества существенна ещё и конкретная организация связей. "Качество отражает устойчивое взаимоотношение составных элементов объекта, которое характеризует его специфику, дающую возможность отличать один объект от других... Вместе с тем качество выражает и то общее, что характеризует весь класс однородных объектов" ([0], ст. «Качество»). Итак, в основе определенного качественного состояния объекта лежит соответствующая система взаимосвязей его составных элементов – структурный закон, который обнаруживается в этой взаимосвязи.

Здесь необходимо отметить, что наиболее существенное сходство объектов проявляется не просто как некоторая эвристически заданная близость наборов значений признаков, как это используется в методах интеллектуального анализа данных, а как "закономерная форма связи вещей, явлений и процессов в составе целого" ([ФЭС], ст. «Общее»). И система взаимосвязи частей в составе целого отражается не просто в виде точек в многомерном признаковом пространстве, а главным образом в виде закономерностей, связывающих значения одних признаков со значениями других. "Наука движется от качественных оценок и описаний явлений к установлению количественных закономерностей, опираясь на последние, она получает возможность глубже исследовать качество" ([ФЭС], ст. «Количество»).

Количественные закономерности, характерные для данного качества, проявляются только до тех пор, пока сохраняется это качество, то есть в границах его *меры*. "Мера - это своего рода зона, в пределах которой данное качество может модифицироваться, сохраняя при этом свои существенные характеристики" ([ФЭС], ст. «Мера»). Если качество характеризуется некоторым структурным законом, то мера, как "единство качественных и количественных характеристик объекта" ([ФЭС], ст. «Мера»), должна содержать, кроме набора интервалов изменений значений признаков, также и весь набор количественных закономерностей – количественный структурный закон строения объекта, в котором качество объекта определяется структурой взаимосвязей частей (признаков) объекта.

5. Закон перехода количества в качество. Таксономическая структура.

До сих пор мы рассматривали количественные изменения внутри одного качества. А что происходит при выходе за пределы меры? С точки зрения системы взаимосвязанных элементов мы вправе ожидать реализации двух принципиально различных возможностей:

- во-первых, это разрушение целостности объекта с полным распадом до элементов;
- во-вторых, это преобразование системы связей, перезамыкание их в новую структуру, т.е. переход в новое качество.

В первом случае количественные изменения приводят к прекращению существования объекта по законам данной предметной области, а во втором - к переходу количества в качество. "Появление нового качества по существу означает появление предмета с новыми закономерностями и мерой, в которой заложена уже иная количественная определенность. ...Начало скачка от одного явления в другое характеризуется началом коренного преобразования всей системы связей между элементами целого, самой природы элементов. Завершение скачка означает образование единства качественно новых элементов и иной структуры целого" (ФЭС, ст. «Переход количественных изменений в качественные»).

Вспомним наше определение классификации: «Разбиение на классы должно производиться так, чтобы объекты одного класса подчинялись одним и тем же закономерностям, объекты разных классов подчинялись разным группам закономерностей». При переходе количества в качество «коренное преобразование всей системы связей между элементами целого» как раз и означает переход от одной группы закономерностей (одного структурного закона), которой подчиняются «взаимосвязь частей объекта» к другой, определяющей другое качество.

Таксономическая структура (структура классов) включает архетипы (описания) всех классов и закономерности качественных переходов между классами. В нашем определении: «У групп закономерностей могут быть общие закономерности, устанавливающие взаимосвязь признаков объектов из разных классов». Таксономическая структура определяется «закономерностями, устанавливающими взаимосвязь между признаками объектов разных классов» или как в определении 1.6.б Шрейдера «правила (закономерности – Е.Е.), показывающие как деформируются свойства объектов при переходе к смежным классам».

6. Формальное определение естественной классификации и систематики

Перейдем к формальному определению введенных понятий. Определим модель $M_a = \langle \Omega_a, Z_a \rangle$ объекта **а** как:

- Ω_a – множество значений всех понятий, признаков, характеристик и величин $x_1, x_2, ..., x_k, ...$, которые применимы к объекту **а** и принимают на нём определенные значения $y^1_{j_1}, y^2_{j_2}, ..., y^k_{j_k}, ...$ (истинности, числовые и т.д.);
- Z_a – множество законов и закономерностей онтологии вида:
  $(x^i_1 = y^{i_1}_{j_1}) \& (x^i_2 = y^{i_2}_{j_2}) \& ... \& (x^i_k = y^{i_k}_{j_k}) \Rightarrow (x^i_0 = y^{i_0}_{j_0})$,
  применимых к объекту **а**. Здесь $y^{i_1}_{j_1}, y^{i_2}_{j_2}, ..., y^{i_k}_{j_k}, y^{i_0}_{j_0}$ – значения признаков $x^i_1, x^i_2, ..., x^i_k, x^i_0$.
  Множество Z_a дает структурный закон строения объекта, его качество.

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 $\mathfrak{C}$, в которой:

- $\Omega_{\mathfrak{C}}$ – множество значений всех понятий, признаков, характеристик и величин $x_1, x_2, ..., x_k, ...$, которые применимы к каждому объекту **а** класса $\mathfrak{C}$ и принимают на нём определенные значения $y^1_{j_1}, y^2_{j_2}, ..., y^k_{j_k}, ...$;
- $Z_{\mathfrak{C}} = \cap Z_{a \in \mathfrak{C}}$.

Проанализируем критерий Е.С. Смирнова [Смирнов Е.С., 1938]. 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 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 признаки, и составляет проблему «индикации» [Смирнов Е.С., 1938]. Такими значениями признаков в закономерной модели класса $M_{\mathfrak{C}}$ являются *порождающие совокупности значений признаков*. По набору значений порождающих признаков $\langle x^i_1 = y^{i_1}_{j_1}, x^i_2 = y^{i_2}_{j_2}, ..., x^i_m = y^{i_m}_{j_m} \rangle$, где $y^{i_1}_{j_1}, y^{i_2}_{j_2}, ..., y^{i_m}_{j_m}$ – значения признаков $x^i_1, x^i_2, ..., x^i_m$, 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.

Рассмотрим задачу построения *систематики*. Рассмотрим в качестве примера таблицу 1. В ней представлены объекты $a_1, a_2, ..., a_9$, разбитые на 4 класса и описываемые 30-ю признаками. Предположим, что классы $\mathfrak{C}_1, ..., \mathfrak{C}_9$ нам известны, и мы знаем закономерные модели этих классов. Задача построения систематики состоит в том, что бы найти такие признаки $x^i_1, x^i_2, ..., x^i_m$, среди 30-и признаков, что бы для каждого класса $\mathfrak{C}_1, ..., \mathfrak{C}_9$ набор значений этих признаков $\langle x_2 = y^2_{j_2}, x_8 = y^8_{j_8}, x_{11} = y^{11}_{j_{11}}, x_{15} = y^{15}_{j_{15}}, x_{21} = y^{21}_{j_{21}}, x_{28} = y^{28}_{j_{28}} \rangle$ являлся порождающим. Эти признаки x_2, x_8, x_{11}, x_{15}, x_{21}, x_{28} условно выделены цветом в таблице.

Набор признаков $S = \langle x^i_1, x^i_2, ..., x^i_m \rangle$ будем называть *системообразующим* для классов $\{\mathfrak{C}_{i \in I}\}$, если для каждого класса из $\{\mathfrak{C}_{i \in I}\}$ значения признаков системообразующего набора $\langle x^i_1 = y^{i_1}_{j_1}, x^i_2 = y^{i_2}_{j_2}, ..., x^i_m = y^{i_m}_{j_m} \rangle$ различны и являются порождающими совокупностями значений признаков. 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 [Борисова И.А., Загоруйко

Н.Г., 2004, 2005] также рассматривается задача нахождения минимального множества «существенных» признаков.

Таблица 1. Построение систематики классов.

Классы	x_1	$\mathbf{x_2}$	x_3	x_4	x_5	x_6	x_7	$\mathbf{x_8}$	x_9	x_{10}	$\mathbf{x_{11}}$	x_{12}	x_{13}	x_{14}	$\mathbf{x_{15}}$	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	$\mathbf{x_{21}}$	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	$\mathbf{x_{28}}$	x_{29}	x_{30}
Класс 1	a_1	$y^2_{j_2}$						$y^8_{j_8}$			$y^{11}_{j_{11}}$				$y^{15}_{j_{15}}$						$y^{21}_{j_{21}}$							$y^{28}_{j_{28}}$		
	a_2	$y^2_{j_2}$						$y^8_{j_8}$			$y^{11}_{j_{11}}$				$y^{15}_{j_{15}}$						$y^{21}_{j_{21}}$							$y^{28}_{j_{28}}$		
Класс 2	a_3	$y^2_{j_2}$						$y^8_{j_8}$			$y^{11}_{j_{11}}$				$y^{15}_{j_{15}}$						$y^{21}_{j_{21}}$							$y^{28}_{j_{28}}$		
	a_4	$y^2_{j_2}$						$y^8_{j_8}$			$y^{11}_{j_{11}}$				$y^{15}_{j_{15}}$						$y^{21}_{j_{21}}$							$y^{28}_{j_{28}}$		
	a_5	$y^2_{j_2}$						$y^8_{j_8}$			$y^{11}_{j_{11}}$				$y^{15}_{j_{15}}$						$y^{21}_{j_{21}}$							$y^{28}_{j_{28}}$		
Класс 3	a_6	$y^2_{j_2}$						$y^8_{j_8}$			$y^{11}_{j_{11}}$				$y^{15}_{j_{15}}$						$y^{21}_{j_{21}}$							$y^{28}_{j_{28}}$		
	a_7	$y^2_{j_2}$						$y^8_{j_8}$			$y^{11}_{j_{11}}$				$y^{15}_{j_{15}}$						$y^{21}_{j_{21}}$							$y^{28}_{j_{28}}$		
Класс 4	a_8	$y^2_{j_2}$						$y^8_{j_8}$			$y^{11}_{j_{11}}$				$y^{15}_{j_{15}}$						$y^{21}_{j_{21}}$							$y^{28}_{j_{28}}$		
	a_9	$y^2_{j_2}$						$y^8_{j_8}$			$y^{11}_{j_{11}}$				$y^{15}_{j_{15}}$						$y^{21}_{j_{21}}$							$y^{28}_{j_{28}}$		

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, вследствие 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 $\mathfrak{C}$ и закономерная модель класса $M_{\mathfrak{C}} = \langle \Omega_{\mathfrak{C}}, Z_{\mathfrak{C}} \rangle$. 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 [Шрейдер С.А., 1983]. Закономерностями первого типа являются закономерности соответствующего класса $Z_{\mathfrak{C}}$, а закономерностями второго типа – закон систематики Z_S.

Закон систематик в любом случае можно представить следующей таблицей. Приведем её на примере классификации из таблицы 1. Здесь закон систематика представлен таблицей наборов порождающих совокупностей значений признаков для всех классов. Например, для класса 1 этот набор имеет вид $\langle y^1_{j_2}, y^1_{j_8}, y^1_{j_{11}}, y^1_{j_{15}}, y^1_{j_{21}}, y^1_{j_{28}} \rangle$.

Таблица 2.

Классы	$\mathbf{x_2}$	$\mathbf{x_8}$	$\mathbf{x_{11}}$	$\mathbf{x_{15}}$	$\mathbf{x_{21}}$	$\mathbf{x_{28}}$
Класс 1	$\langle y^2_{j_2}$	$y^8_{j_8}$	$y^{11}_{j_{11}}$	$y^{15}_{j_{15}}$	$y^{21}_{j_{21}}$	$y^{28}_{j_{28}} \rangle$
Класс 2	$\langle y^2_{j_2}$	$y^8_{j_8}$	$y^{11}_{j_{11}}$	$y^{15}_{j_{15}}$	$y^{21}_{j_{21}}$	$y^{28}_{j_{28}} \rangle$
Класс 3	$\langle y^2_{j_2}$	$y^8_{j_8}$	$y^{11}_{j_{11}}$	$y^{15}_{j_{15}}$	$y^{21}_{j_{21}}$	$y^{28}_{j_{28}} \rangle$
Класс 4	$\langle y^2_{j_2}$	$y^8_{j_8}$	$y^{11}_{j_{11}}$	$y^{15}_{j_{15}}$	$y^{21}_{j_{21}}$	$y^{28}_{j_{28}} \rangle$

Рассмотрим критерий А.А.Любищева [Забродин В.Ю., 1981]. Системой по Любищеву является такое представление классификации объектов, где по месту объекта в системе определяются все его признаки. В нашем определении значения признаков объектов определяются взаимодействием двух законов:

- закона систематики Z_S, используя который мы по положению объекта в системе (таблице) можем определить класс объекта и значения системообразующих признаков;
- по закономерностям класса $Z_{\mathfrak{C}}$ и значениям системообразующих признаков этого класса и мы далее можем определить все остальные свойства объекта.

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 как системы, она является законом природы потому что из неё можно предсказать потенциально бесконечное количество свойств объектов.

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.

Такая классификация объектов, которая строится с целью построения наиболее совершенной систематики, называется «естественной» классификацией.

7. Метод построения «естественной» классификации

Перейдем теперь к рассмотрению основного механизма построения «естественной» классификации. Проанализируем таблицу 1. В каждом классе $\mathfrak{C}$, по закономерностям класса $Z_{\mathfrak{C}}$ и по порождающим значениям признаков, предсказываются значения всех остальных признаков объектов класса. В качестве порождающих значений признаков можно брать различные системы признаков и их значений и каждый раз значения всех остальных признаков будут предсказываться по порождающим значениям признаков. Фактически, класс определяется не порождающими признаками, а системой взаимосвязанных и взаимосогласованных по предсказанию закономерностей $Z_{\mathfrak{C}}$, что соответствует нашему определению «естественной» классификации: «Объекты одного класса, кроме того, должны обладать некоторой целостностью. Целостность определим как взаимную *согласованность закономерностей* каждой группы *по предсказанию свойств объектов*». Согласованность закономерностей $Z_{\mathfrak{C}}$ по предсказанию значений признаков каждого класса означает, что по значениям некоторых признаков $\langle x^i_1 = y^{i_1}_{j_1}, x^i_2 = y^{i_2}_{j_2}, ..., x^i_m = y^{i_m}_{j_m}\rangle$ объектов класса предсказываются (по закономерностям из $Z_{\mathfrak{C}}$) значения некоторых других признаков $\langle x^k_1 = y^{k_1}_{j_1}, x^k_2 = y^{k_2}_{j_2}, ..., x^k_m = y^{k_m}_{j_m}\rangle$ этого же класса и, наоборот, по значениям признаков $\langle x^k_1 = y^{k_1}_{j_1}, x^k_2 = y^{k_2}_{j_2}, ..., x^k_m = y^{k_m}_{j_m}\rangle$ предсказываются (по закономерностям из $Z_{\mathfrak{C}}$) значения признаков $\langle x^i_1 = y^{i_1}_{j_1}, x^i_2 = y^{i_2}_{j_2}, ..., x^i_m = y^{i_m}_{j_m}\rangle$. Таким образом, значения признаков некоторого класса предсказывают друг друга по закономерностям из $Z_{\mathfrak{C}}$ как бы по «замкнутому кругу».

Таким образом, если мы не знаем классы, то искать их надо, обнаруживая наборы взаимно предсказывающихся значений признаков $\langle y^{i_1}_{j_1}, y^{i_2}_{j_2}, ..., y^{i_N}_{j_N}\rangle$. Однако, если мы не знаем классы, то и не знаем закономерности классов $\{Z_{\mathfrak{C}i}\}_{i\in I}$. Поэтому у нас в распоряжении есть только множество всех закономерностей $Z = \cup Z_{\mathfrak{C}i\in I}$, которые можно обнаружить на всех объектах (объединении всех объектов классов). Множество закономерностей Z можно обнаружить системой Discovery [Витяев Е.Е., 1976, 2006; Витяев Е.Е., Москвитин А.А., 1993; Kovalerchuk B., Vityaev E., 2000; Vityaev E., Kovalerchuk B., 2004, 2008]. Если по закономерностям из Z обнаружить набор взаимно предсказывающихся значений признаков $\langle y^{i_1}_{j_1}, y^{i_2}_{j_2}, ..., y^{i_N}_{j_N}\rangle$, то закономерности класса $Z_{\mathfrak{C}}$ есть те и только те закономерности из Z, которые применимы к значениям признаков набора $\langle y^{i_1}_{j_1}, y^{i_2}_{j_2}, ..., y^{i_N}_{j_N}\rangle$. Как теперь найти эти наборы значений признаков $\langle y^{i_1}_{j_1}, y^{i_2}_{j_2}, ..., y^{i_N}_{j_N}\rangle$ и определить, что некоторый объект **a** принадлежит классу, определяемому этим набором значений. Объекты могут содержать ошибки в измерениях значений признаков.

Для этого, используя закономерности из Z, мы последовательно выполним одну из следующих коррекций некоторого объекта **a**:

a) возьмём набор значений признаков $\langle y_a^1, y_a^2, ..., y_a^N \rangle$, который описывает объект **a**.
b) дополним этот набор новыми значениями признаков, которые с высокой вероятностью предсказываются (по закономерностям из Z) по имеющимся значениям признаков и сами хорошо предсказывают другие значения признаков;
c) удаляем те значения признаков, которые противоречат предсказаниям, сделанных по закономерностям из Z по имеющимся значениям признаков.
d) шаги a), b) продолжаем до тех пор, пока не будут включены все необходимые значения и не будут удалены все случайные. Контролировать этот процесс должен определённый критерий качества взаимосогласованности закономерностей по предсказанию на наборе $\langle y_a^1, y_a^2, ..., y_a^N \rangle$, который приведён в [Витяев Е.Е., 1983].

Эту процедуру будем называть процедурой *идеализации*. Она даёт наборы классов $\langle y_{j_1}^{i_1}, y_{j_2}^{i_2}, ..., y_{j_N}^{i_N} \rangle$. "Сотри случайные черты, и ты увидишь - мир прекрасен" – эти слова Александра Блока как нельзя лучше характеризуют процесс идеализации. Проведя процедуру *идеализации* для всех имеющихся объектов, мы получим все наборы классов $\langle y_{j_1}^{i_1}, y_{j_2}^{i_2}, ..., y_{j_N}^{i_N} \rangle$. Таким образом, каждый класс $\mathfrak{C}$ будет характеризоваться своим набором значений признаков $\langle y_{j_1}^{i_1}, y_{j_2}^{i_2}, ..., y_{j_N}^{i_N} \rangle_{\mathfrak{C}}$ и соответствующим ему множеством закономерностей $Z_{\mathfrak{C}}$. Те объекты, которые в процессе идеализации будут приводиться к набору значений некоторого класса $\mathfrak{C}$, будут относиться к этому классу. Множество объектов отнесенных к классу $\mathfrak{C}$ обозначим через $A_{\mathfrak{C}}$. Таким образом, полученная «естественная» классификация будет содержать множество наборов значений признаков $\langle y_{j_1}^{i_1}, y_{j_2}^{i_2}, ..., y_{j_N}^{i_N} \rangle_{\mathfrak{C}}$, множества закономерностей $Z_{\mathfrak{C}}$ и классы объектов $\mathfrak{C}_i$, $A_{\mathfrak{C}}, i \in I$. На основании полученной «естественной» классификации можно построить систематику, как было описано выше.

8. Онтологии

Вернемся к понятию онтологии. Определение онтологии, принятое в литературе, включает только пункты 2.1-2.4 определения онтологии §2. Индуктивно выведенные (обнаруженные) законы и закономерности из пункта 2.5 не включаются обычно в определение онтологии. Как видно из предыдущих рассуждений, для построения «естественной» классификации и систематики достаточно к определению онтологии в традиционном смысле (без п. 2.5) добавить индуктивно выведенные законы и закономерности. Для этого достаточно предположить, как было отмечено в предыдущем пункте, что мы применяем к имеющимся данным метод обнаружения закономерностей – систему Discovery [Витяев Е.Е., 1976, 2006; Витяев Е.Е., Москвитин А.А., 1993; Kovalerchuk B., Vityaev E., 2000; Vityaev E., Kovalerchuk B., 2004, 2008]. Отсюда следует формула:

«**онтология + индуктивные закономерности (система Discovery) = «естественная» классификация и систематика**».

Покажем в следующем параграфе, что полученная таким образом «естественная» классификация и систематика позволяют решать достаточно широкий круг задач.

9. Функции «естественной» классификации и систематики.

«Естественная» классификация позволяет решать задачи часть из которых не решаются традиционными методами анализа данных. Перечислим функции «естественной» классификации по работе [Кожара В.Л., 1982] и опишем метод их решения, основанный на построении «естественной» классификации и систематики.

6.1 Аналоговое прогнозирование. Под **аналоговым прогнозированием** понимается предсказание свойств объектов на основании аналогии, то есть перенесение на объект существенных свойств другого объекта – его аналога. При этом возникает два вопроса: как ищется аналог, и какие свойства можно считать существенными.

Существующие методы классификации опираются на представление о сходстве объектов в признаковом пространстве и группировании близких объектов в некоторые сгустки точек в этом многомерном пространстве. Различение существенных и несущественных, необходимых и случайных признаков не имеет четкого обоснования в такой постановке. Кроме того, задачи классификации и распознавания образов не являются методами вывода по аналогии.

В «естественной» классификации вывод по аналогии основывается на том, что объекты одного класса подчиняются одному и тому же структурному закону своего строения (системе взаимосогласованных по предсказанию закономерностей). В этом случае они тождественны в части существенных признаков, описывающих закон строения объекта.

Различают два вида аналогового прогнозирования: *диагностирование* и *распространение*.

6.1.1 Диагностирование. Задачей диагностирования является определение принадлежности объекта некоторому классу. В анализе данных эта задача называется задачей распознавания. Как и классификация, распознавание основывается, чаще всего, на сходстве объектов в признаковом пространстве.

С точки зрения «естественной» классификации диагностирование производится иначе, проведением процедуры идеализации.

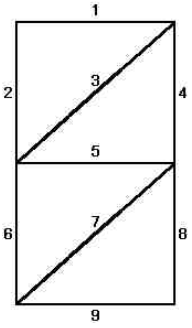
6.1.2 Распространение. Распространение означает перенос свойств объекта на класс, когда существенные свойства некоторого объекта класса переносятся на весь класс. В этом случае класс сам по себе обладает определённым набором свойств – набор класса $\langle y^{i_1}_{j_1}, y^{i_2}_{j_2}, ..., y^{i_N}_{j_N} \rangle$, присущих всем объектам класса. Прямого аналога этой задачи нет в анализе данных. Похожей задачей является определение эталона класса как срединного по положению в пространстве признаков. Но в этом случае нет различия между существенными и случайными признаками. В «естественной» классификации эта задача решается проведением процедуры идеализации объекта, дающей набор $\langle y^{i_1}_{j_1}, y^{i_2}_{j_2}, ..., y^{i_N}_{j_N} \rangle$ значений признаков класса.

6.2 Структурное прогнозирование. Поскольку «естественная» классификация описывает объекты одновременно на двух уровнях – систематики и классификации, то она включает в себя структуры двух типов – таксономическую (структура классов) и мерономическую (структурный закон строения объектов класса). Соответственно эти структуры описываются закономерности двух типов а) и б), как описано в определении Шрейдера 1.6. Для восстановления пропущенных элементов этих структур нужно использовать соответствующие им закономерности. Соответственно возникают задачи **таксономического** или **мерономического структурного прогнозирования**.

6.2.1 Таксономическое прогнозирование. Задачу восстановления отсутствующих звеньев в систематики невозможно решить без использования закономерностей таксономической структуры. Таксономическая структура включает в себя архетипы классов (наборы значений $\langle y^{i_1}_{j_1}, y^{i_2}_{j_2}, ..., y^{i_N}_{j_N} \rangle$ классов) и закономерные связи признаков между архетипами различных классов (закона систематики Z_S). Дефекты таксономической структуры устраняются использованием закона систематики Z_S. В том числе возможно предсказание существования новых классов. Аналога такой задачи нет в анализе данных.

6.2.2 Мерономическое прогнозирование. Восстановление мерономической структуры объектов класса $\mathfrak{C}$ осуществляется по закономерностям класса $Z_{\mathfrak{C}}$. Для прогнозирования значений недостающих (неизвестных) признаков нет необходимости проводить процедуру идеализации – достаточно осуществить предсказание этих признаков закономерностями из $Z_{\mathfrak{C}}$. В анализе данных эта задача называется задачей предсказания (прогнозирования), а также задачей заполнения пробелов в данных. Для её решения применяются различные методы, основанные, как правило, на близости в признаковом пространстве.

10. Пример построения систематики цифр индекса.



Рассмотрим цифры индекса как набор из 9 объектов. Определим предикаты P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, означающие наличие i-го признака в начертании цифры. Занумеруем признаки как показано на рисунке. Например, $P_1(3) = 1$ означает, что в начертании цифры 3 есть линия номер 1. Определим значения всех предикатов (признаков) и представим их в виде таблицы 3. Будем рассматривать цифры как классы $I = \{0,..,9\}$. Рассмотрим модель, включающую все цифры $M = \{I, Q\}$, где $Q = \{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9\}$ – множество предикатов (признаков). Для построения «естественной» классификации обнаружим на M все детерминированные закономерности вида:

$$P_{i_1}^{\varepsilon_1}(a) \& \ldots \& P_{i_m}^{\varepsilon_m}(a) \Rightarrow P_{i_0}^{\varepsilon_0}(a),$$

где $\{P_{i_1}^{\varepsilon_1}(a),\ldots,P_{i_m}^{\varepsilon_m}(a),P_{i_0}^{\varepsilon_0}(a)\} \subset \{P_1,\ldots,P_9\}$; ε = 1(0), если отношение берется без отрицания (с отрицанием), которые удовлетворяют условиям:

a) среди выражений $\{P_{i_1}^{\varepsilon_1}(a),\ldots,P_{i_m}^{\varepsilon_m}(a),P_{i_0}^{\varepsilon_0}(a)\}$ нет повторений и нет одновременно отношения и его отрицания;
b) если из конъюнкции $P_{i_1}^{\varepsilon_1}(a) \& \ldots \& P_{i_m}^{\varepsilon_m}(a)$ удалить одно из отношений, то полученная формула станет ложной на A;
c) если заменить отношение $P_{i_0}^{\varepsilon_0}(a)$ на «ложь», то полученная формула станет ложной на A.

В общем случае закономерности не являются детерминированными. Более общее определение закономерностей дано в [Витяев Е.Е., 1976, 2006; Kovalerchuk B., Vityaev E., 2000; Vityaev E., Kovalerchuk B., 2004, 2008].

Обнаружим детерминированные закономерности системой Discovery. Получим множество закономерностей Z, включающее 3743-и закономерности найденные системой по таблице 3.

В данном примере классы нам известны – это все цифры I. Для каждого класса $\mathfrak{C} \in I$ найдем закономерности из Z, которые на нём выполняются и получим множества закономерностей классов $Z_{\mathfrak{C}}$. Например, для цифры 2 будет выполнено 529 закономерностей. Закономерной моделью цифры 2 будет модель $M_2 = \langle 2, Z_2 \rangle$.

Для построения систематики цифр, определим для каждого класса минимальные определяющие совокупности. Для цифры 2 это будет, например, совокупность $\{P_2, P_3\}$, т.к. значения остальных признаков предсказываются по следующим закономерностям из Z_2:

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

Как уже упоминалось, определяющие совокупности определяются не единственным образом, например, набор $\{P_5, P_7\}$ также будет определяющей совокупностью, по значениям которого будут предсказываются значения остальных признаков по закономерностям из Z_2:

$P_7 \Rightarrow P_1$
$P_7 \& \neg P_5 \Rightarrow \neg P_2$
$P_7 \& \neg P_5 \Rightarrow P_4$
$P_4 \& \neg P_2 \& P_1 \Rightarrow \neg P_3$
$\neg P_3 \& \neg P_2 \Rightarrow P_9$

	P_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

Таблица 3. 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	$\{P_4,P_5,P_7\}$

Таблица 4. Систематика цифр.

$P_4\&\neg P_2\Rightarrow\neg P_6$

$P_9\&\neg P_6\& P_4\Rightarrow \neg P_8$

Глядя на закономерности видно, что для порождающего набора $\{P_5,P_7\}$ закономерная модель двойки проще. Она будет выглядеть следующим образом:

$M_2 = \langle 2, Z_2\rangle = \{\{1,0,0,1,0,0,1,0,1\}, \{P_7,\neg P_5, P_7\Rightarrow P_1, P_7\&\neg P_5\Rightarrow\neg P_2, P_7\&\neg P_5\Rightarrow P_4, P_4\&\neg P_2\& P_1\Rightarrow\neg P_3, \neg P_3\&\neg P_2\Rightarrow P_9, P_4\&\neg P_2\Rightarrow\neg P_6, P_9\&\neg P_6\& P_4\Rightarrow \neg P_8\}\}$.

Построим закономерную модель систематики. Её закон Z_S представим в виде таблицы. Для выбора минимальной определяющей совокупности систематики, рассмотрим различные сочетания определяющих совокупностей классов.

Максимальная по количеству признаков минимальная определяющая совокупность у цифры 8 (минимальное число порождающих признаков равно 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 для цифры 3 это $\{P_3, P_7\}$, $\{\neg P_4, P_7\}$, тогда как определяющие совокупности, состоящие из 3 признаков, не содержат 7-го признака. Следовательно, нужно рассматривать не только все определяющие совокупности длины 2, но и определяющие совокупности длинны 3 и более 3 признаков.

Так как $2^3 = 8$ меньше, чем число классов 10, то 3-х признаков будет заведомо недостаточно для однозначного восстановления класса. Поэтому нужно рассмотреть различные комбинации из 4-х признаков. В результате получим, что минимальная определяющая совокупность признаков для систематики цифр это $\{P_4, P_5, P_6, P_7\}$ (см. таблицу 4). В этом случае она определяется единственным образом.

Систематика цифр индекса, показанная в таблице 4, содержит в первом столбце цифру, в 2-5 столбцах значения предикатов P_4, P_5, P_6, P_7 для этой цифры и в последнем столбце – минимальный порождающий набор предикатов из числа P_4, P_5, P_6, P_7. Каждый класс, кроме того, характеризуется своим множеством закономерностей $Z_i, \in I$, по закономерностям которого восстанавливаются значения остальных признаков.

11. Применение «естественной» классификации в биоинформатике.

В работе [Vityaev E.E., 2008] получено применение «естественной» классификации для анализа регуляторных районов генов. Для обработки данных была использована программа построения «естественных» классификаций нуклеотидных последовательностей генов DNAClass, разработанная К.А. Лапардиным.

По выборке выровненных последовательностей сайта связывания транскрипционных факторов SF1 и SREBP были обнаружены «естественные» классы в виде последовательности значений признаков класса $\langle y^{i_1}_{j_1}, y^{i_2}_{j_2}, \ldots, y^{i_N}_{j_N}\rangle$. Для сайта SF1 она имела вид

[T/C][C][A][A][G][G][T/C][C][A][G],

где [T/C] означает, что на первом месте в последовательности может стоять, как нуклеотид T, так и C. Используя закономерности, обнаруженные для классов, проводилось распознавание этого сайта в новых последовательностях. Были получены хорошие оценки ошибок второго рода, при 50% ошибке первого рода: $2*10^{-5}$ для сайта SF1 и 10^{-5} для сайта SREBP. Этот результат превосходит точность стандартно применяемых для распознавания сайтов весовых матриц.

Литература

Борисова И.А., Загоруйко Н.Г. "Естественная классификация" // Сборник трудов ИАИ-2004, Киев, 2004 г., с 33-42.

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.

Витяев Е.Е. Метод обнаружения закономерностей и метод предсказания. – В кн.: Эмпирическое предсказание и распознавание образов (Выч. сист. 67), Новосибирск, 1976, с. 54-68.

Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов: Моногр. // Новосибирский гос. ун-т. Новосибирск, 2006. 293 с.

Витяев Е.Е., Костин В.С. Естественная классификация как закон природы // Интелектуальные системы и методология. (Материалы научно-практического симпозиума "Интелектуальная поддержка деятельности в сложных предметных областях"), вып.4, Новосибирск, 1992, с. 107-115.

Витяев Е.Е., Морозова Н.С., Сутягин А.С., Лапардин К.А. Естественная классификация и систематика как законы природы // Анализ структурных закономерностей (Вычислительные системы вып. 174), Новосибирск, 2005, с.80-92

Витяев Е.Е., Москвитин А.А. Введение в теорию открытий. Программная система DISCOVERY. // Логические методы в информатике (Вычислительные системы, вып. 148), Новосибирск, 1993, с.117-163.

Забродин В.Ю. О критериях естественной классификации. – НТИ, сер.2, 1981,№8.

Кожара В.Л. Анализ информативно насыщенных таксономических структур как способ выявления географических закономерностей // Дисс. Канд. Геогр. Н., М., 1989.

Кожара В.Л. Функции классификации // Теория классификаций и анализ данных, Новосибирск, 1982, ч. 1.

Мальцев А.И. Алгебраические системы, М., Наука, 1970.

Мейен С.В., Шрейдер С.А. Методологические аспекты теории классификаций. Вопросы философии, 1976, №12.

Россеева О.И., Загорулько Ю.А., Сергеев И.П. Онтологии как средство организации эффективного поиска в сети Internet.

Рутковский Л. Элементарный учебник логики. – Спб., 1884.

Смирнов Е.С. Конструкция вида таксономической точки зрения // Зоол. Журн. – 1938, т. 17, №3, с. 387-418.

Шрейдер С.А. Систематика, типологии, классификация. – В кн.: Теория и методология биологических классификаций, М., Наука, 1983.

ФЭС. Философский энциклопедический словарь. М. "Советская энциклопедия" 1989. 816с.

Scientific Discovery website
http://www.math.nsc.ru/AP/ScientificDiscovery/Related_papers.html#Natural Classification

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

Kovalerchuk B., Vityaev E. Data Mining in Finance: Advances in Relational and Hybrid methods. (Kluwer international series in engineering and computer science; SECS 547), Kluwer Academic Publishers, 2000, p.308.

Evgenii Vityaev, Boris Kovalerchuk, Empirical Theories Discovery based on the Measurement Theory. Mind and Machine, v.14, #4, 551-573, 2004.

E. Vityaev, B.Y. Kovalerchuk, Relational Methodology for Data Mining and Knowledge Discovery. Intelligent Data Analysis. Special issue on "Philosophies and Methodologies for Knowledge Discovery and Intelligent Data Analysis" eds. Keith Rennolls, Evgenii Vityaev. v.12(2), IOS Press, 2008, pp. 189-210.

Vityaev E.E., Lapardin K.A., Khomicheva I.V., Proskura A.,L. Transcription factor binding site recognition by regularity matrices based on the natural classification method. Intelligent Data Analysis. Special issue on "Machine learning and bioinformatics" eds. Nikolai Kolchanov, Evgenii Vityaev. v.12(5), IOS Press, 2008 (in press)

Christopher Welty , Nicola Guarino (2001) Supporting ontological analysis of taxonomic relationships, Data & Knowledge Engineering, v.39 n.1, p.51-74

Zagoruiko N., Borisova I. "Principles of natural classification"// Pattern Recognition and Image Analysis, 2005, Vol.15, No.1, p.27-29.