

В работе анализируются основные понятия, связанные с возможностью познания некоторой предметной области методами интеллектуального анализа данных (Knowledge Discovery in Data Bases and Data Mining – KDD&DM) и машинного обучения (Machine Learning – ML) такие как: предметная область, предмет исследования, система понятий предметной области, онтология предметной области, извлечение информации из предметной области, знание, процесс познания, извлечение информации из данных, онтология методов KDD&DM и ML, извлечения знаний из данных и другие. Аргументируется, что для познания некоторой предметной области методами интеллектуального анализа данных необходимо, чтобы онтология предметной области была согласована с онтологией применяемого метода KDD&DM и ML. Для такого согласования необходимо сначала извлечь информацию из данных. Детально описывается процедура извлечения информации из наиболее распространенных типов данных, таких как парные сравнения, множественные сравнения, матричное представление бинарных отношений, матрицы упорядочений, матрицы близости и матрицы объект–признак. Кратко описывается оригинальный подход – Relational Data Mining – к извлечению знаний из данных, позволяющий обнаруживать знания опираясь не

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

1. Познание предметной области. Определим, что такое Предметная Область (ПО). Предметная область – это совокупность объектов предметной области, рассматриваемых с точки зрения некоторого предмета исследования – совокупности существенных свойств (атрибутов) и отношений объектов исследования, описываемых в некоторой системе понятий предметной области. Предмет исследования может быть задан онтологией предметной области – специфицирующей в некотором формальном языке множество рассматриваемых объектов, связи между ними, систему понятий, и свойства объектов. Предмет исследования и онтология определяют «взгляд», «точку зрения», с которой рассматриваются (описываются в системе понятий) объекты предметной области, отношения и их свойства.

Как отмечалось в [Витяев Е.Е., 2009] предметная область может быть задана эмпирической системой $\mathfrak{S} = \langle A, \Omega \rangle$, где A – объекты ПО, а Ω – множество отношений и операций, интерпретируемых в системе понятий ПО, и определяющих взаимосвязь объектов ПО. Система понятий онтологии может быть задана одноместными предикатами, которые также могут входить в Ω . Таким образом, множество Ω представляет собой онтологию предметной области, так как является формальной спецификацией связей между объектами, системы понятий и свойств объектов.

Для осуществления процесса познания необходимо понимание и интерпретация человеком предметной области и её онтологии, т.е. извлечение информации из предметной области. «Информация – это понимание (смысл, представление, интерпретация) возникающие в аппарате мышления человека, в результате получения им данных, взаимоувязанное с предшествующими знаниями и понятиями» [Фридланд А.Я., 2003]. Информация о предметной области и онтологии состоит из восприятия и интерпретации человеком объектов предметной области, связей между ними, системы понятий и свойств объектов. В результате такой интерпретации получаем знание о предметной области. «Знания – это воспринятая, осознанная и ставшая личностно значимой информация» [Бешенков С.А., Ракитина Е.А., 2002].

2. Информация, содержащаяся в атрибутах и свойствах объектов. Проанализируем, как следует задавать свойства и атрибуты объектов ПО в терминах онтологии Ω . Чтобы правильно извлекать информацию и знания из свойств и атрибутов необходимо их интерпретировать в системе понятий ПО. Сами по себе числовые значения величин смысла и информацию не содержат, смысл величин указывается в их интерпретации, например, 5 метров, 5 литров, 5 килограмм и т.д. Интерпретация чисел, в частности, определяет какие математические действия можно с ними осмысленно проводить, чтобы не получать бессмысленных результатов типа 1.5 дровосека, 1м. + 1кг., и т.д.

Как говорилось в [Витяев Е.Е., 2009] интерпретация числовых значений – метры, литры, килограммы и т.д. привязана к соответствующей ПО и её онтологии. Физические величины, измерен-

¹ Работа поддержана грантом РФФИ 08-07-00272-а; интеграционными проектами СО РАН № 47, 115, 119, а также работа выполнена при финансовой поддержке Совета по грантам Президента РФ и государственной поддержке ведущих научных школ (проект НШ-335.2008.1)

ные в отличной от физики предметной области, теряют свою исходно физическую интерпретацию. Рассмотрим, например, такую физическую величину как температура. Шкалы температур в нефизических областях, например, при измерении температуры тела больного в медицине, температуры почвы в сельском хозяйстве, температуры воздуха в духовке в кулинарии и т.д., должны быть разные, хотя измеряться могут одним и тем же прибором – термометром. Далеко не всеми понимается тот факт, что шкала – это набор отношений и операций, которые имеет смысл производить с числовыми значениями величин в данной предметной области. Точнее, это те отношения и операции, которые интерпретируемы в онтологии ПО. Можно возразить, что термометр не может измерять ничего кроме температуры. Он действительно во всех случаях измеряет физическую температуру. Но зачем мы измеряем температуру? Ведь не затем чтобы согласно законам физики узнать, сколько в больном содержится тепла, и не затем, чтобы определить среднюю кинетическую энергию молекул почвы или курицы в духовке. Температура, как и любой другой прибор, нужен для *получения выводов (знаний) в системе понятий* (онтологии) той предметной области, к которой он относится. Для больного «температурный фактор служит наиболее общим и универсальным регулятором скорости химических реакций и активности ферментов, с повышением температуры в известной мере ускоряются и обменные процессы» [Лихорадка]. Для почв температура интерпретируется в системе понятий физиологии растений и деятельности микроорганизмов. Физическая величина температуры в других предметных областях *является косвенным измерением* некоей другой величины, интерпретируемой в системе понятий предметной области, которую мы и хотим измерить. Физическая температура больного – есть косвенное измерение медицинской величины – уровня обмена веществ, температура почвы измеряет состояние биохимических процессов в растениях и микроорганизмах, температура воздуха в духовке измеряет течение процесса свертывания белка и т.д. Какие отношения и операции над числовыми значениями температуры имеют смысл для всех этих величин, определяется уже этими интерпретациями и онтологиями соответствующих ПО. Например, для температуры больного интерпретируемы выделенные значения 36.7, 42. и отношение линейного порядка $<$.

Таким образом, для извлечения информации из атрибутов, свойств, признаков и величин ПО нужно определить (задача 1 в [Витяев Е.Е., 2009]) множество интерпретируемых в онтологии Ω математических отношений и операций и включить их в онтологию Ω .

Зачем нужно такое извлечение информации из атрибутов, свойств и величин ПО?

3. Познание предметной области методами интеллектуального анализа данных. Онтология методов интеллектуального анализа данных. Рассмотрим методы интеллектуального анализа данных (Knowledge Discovery in Data Bases and Data Mining (KDD&DM)) и машинного обучения (Machine Learning (ML)) с точки зрения их связи с процессом познания. Чтобы такое познание было осмысленно и интерпретируемо необходимо, чтобы методы KDD&DM и ML правильно использовали содержащуюся в данных информацию. Рассмотрим этот вопрос более подробно.

Анализ методов KDD&DM и ML [Витяев, 2006; E. Vityaev, B.Y. Kovalerchuk, 2008] показывает, что *методы* имеют свою *онтологию*, которая включает:

1. типы данных, с которыми работает метод;
2. язык оперирования и интерпретации данных;
3. класс гипотез, проверяемый методом и сформулированный в языке интерпретации данных.

Для того, чтобы познание ПО некоторым KDD&DM-методом было возможным и приводило к знаниям – интерпретируемым в онтологии ПО результатам, необходимо чтобы онтология метода и онтология ПО были согласованы между собой. Это означает, что:

1. типы данных, с которыми работает метод, должны интерпретироваться в онтологии Ω предметной области. Поэтому атрибуты, свойства и признаки, используемые в данных метода, должны быть интерпретируемы в онтологии Ω . Тем самым определяется *информация, извлекаемая из данных этим методом*, которая представляется множеством интерпретируемых в онтологии Ω математических отношений и операций;
2. язык оперирования данными, используемый методом в своей работе, также должен интерпретироваться в онтологии Ω . Это значит, что метод должен использовать в своей работе только интерпретируемые в онтологии Ω математические отношения и операции. Если это не так, что, как правило, и имеет место, то метод получает не вполне интерпретируемые и не являющиеся знаниями результаты. Человек не может осознать результаты математических действий, применённых методом, которые для него не имеют интерпретацию и, следовательно, бессмысленны с точки зрения системы понятий ПО;
3. класс проверяемых методом гипотез также должен интерпретироваться в онтологии ПО.

Это означает, что класс проверяемых гипотез также должен выражаться через интерпретируемые в онтологии Ω математические отношения и операции. Например, решающие функции в распознавании образов, функции регрессии, формы кластеров в признаковом пространстве и т.д. должны содержать только интерпретируемые математические отношения и операции.

В настоящее время такого рода проверка на соответствие онтологии ПО и онтологии метода не проводится. Для того, что бы знать какая информация содержится в данных и, следовательно, какой метод KDD&DM и ML мы можем применить для обработки данных, нам необходимо *извлечь информацию из данных*. Для этого необходимо представить данные в виде эмпирических систем, в которых информация, содержащаяся в данных, будет представлена множеством отношений и операций, интерпретируемых в онтологии предметной области. После этого можно определить, в соответствии с этими отношениями и операциями, какой метод KDD&DM и ML можно применять для обработки этих данных.

4. Реляционный подход к извлечению знаний из информации, содержащейся в данных.

Даже если окажется, что никакой метод KDD&DM и ML не применим для обработки рассматриваемых данных (они не используют требуемые отношения и операции), то можно применить разработанный нами реляционный подход (Relational Data Mining) к методам извлечения знаний и систему «Discovery», реализующую этот подход [Витяев, 2006; Витяев, Москвитин, 1993; Vityaev, Kovalerchuk, 2004; Kovalerchuk, Vityaev, 2000]. Данный подход и система «Discovery» специально разработаны для обнаружения знаний путём использования информации, представленной эмпирическими системами. Система «Discovery», кроме того, обладает следующими преимуществами по сравнению с другими KDD&DM и ML методами.

Существующие методы не в состоянии поддерживать режим исследования данных, когда обнаруживаемая закономерность заранее неизвестна. Каждый KDD&DM-метод обнаруживает свой специфический класс гипотез, соответствующий его онтологии. Система «Discovery» может обнаружить произвольный класс гипотез, который захочет проверить эксперт, и тем самым она способна поддерживать режим исследования данных, когда класс гипотез может сильно изменяться пользователем (экспертом) в процессе исследования.

Система «Discovery» обнаруживает гипотезы, сформулированные в заданных экспертом (например, финансистом) терминах – множестве интерпретируемых в онтологии ПО отношений и операций. Интерпретируемость получаемых закономерностей очень важна при принятии ответственных решений в таких областях, как медицина или финансы. К примеру, если речь идет о крупном вложении капитала и у нас есть два прогноза об ожидаемой прибыли, полученные нейронными сетями и системой «Discovery», то доверие будет к тому прогнозу, который понятен и интерпретируем. Невозможно принимать ответственные решения, не понимая, как они получены. Прогнозы, получаемые на основании интерпретируемых правил понятны, и по ним можно принимать решения.

Другой важной задачей, которую решает система «Discovery», является задача максимально полного *извлечения знаний из данных*. Полнота извлечения знаний системой «Discovery» обеспечивается двумя путями:

1. использованием онтологии ПО и теории измерений для извлечения всей интерпретируемой информации из данных и представлением её эмпирической системой;
2. обнаружением практически любого класса гипотез в терминах извлеченной информации – выявленных отношений и операций в этой эмпирической системе.

Для решения этих задач система «Discovery» позволяет:

1. предоставлять пользователю возможность в диалоге с системой, задавать отношения и операции, которые система будет использовать, и которые интерпретируемы в онтологии ПО, что позволяет работать в информацией, извлеченной из данных, а не с самими исходными данными;
2. возможности задания любого класса гипотез, сформулированного в терминах этой информации – заданным самим же пользователем множеством отношений и операций.

На данный момент разработана достаточно «универсальная» версия системы «Discovery», позволяющая пользователю самому задавать класс обнаруживаемых закономерностей, извлекать из данных множество закономерностей заданного класса и использовать найденные закономерности для прогноза и принятия решений.

4. Извлечение информации из данных. Для извлечения информации из данных и дальнейшего применения какого-либо KDD&DM и ML метода покажем, как из таких известных типов дан-

ных, как парные сравнения, множественные сравнения, матричное представление бинарных отношений, матрицы упорядочений, матрицы близости и матрицы объект–признак, *может быть извлечена информация*, представленная эмпирической системой. Такие типы данных встречаются в таких областях, как экспертное оценивание, социология, психология, психофизика, геология, медицина, сельское хозяйство и т. д. Все эти области характеризуются тем, что в них встречаются признаки и величины самой разнообразной природы.

Для извлеченной информации в виде эмпирических систем приведем соответствующие результаты теории измерений [Krantz D.H. et al, 1971; 1989; 1990; Пфанцагль И., 1976] показывающие, какие числовые представления для данных эмпирических систем существуют. Для обнаружения систем аксиом теории измерений можно применить систему «Discovery».

4.1. Парные сравнения. Результаты, полученные по методу парных сравнений, можно представить в виде четырехмерной матрицы (x_{ijst}) [Девид Г., 1978], где i, j – номера сравниваемых объектов, взятых из некоторого множества $A = \{a_1, \dots, a_m\}$, $s = 1, \dots, n$ – номер экспертов, сравнивающих объекты из A ; $t = 1, \dots, r_s$ – номер сравнения (пары объектов одним и тем же экспертом могут сравниваться r_s раз). Обозначим объект a_i , сравниваемый экспертом s в сравнении с номером t , через a_i^{st} . Тем самым мы предполагаем, что сам объект и эксперт могут изменяться от сравнения к сравнению. Значение $x_{ijst} = 0(1)$, если объект a_i^{st} предпочтительнее, чем объект a_j^{st} .

Методы парного сравнения используются в социологии в экспертных оценках, психологии и в других областях. Целью этих методов является получение полного упорядочения объектов множества A .

Определим, какие эмпирические системы извлекают информацию из данных, полученных методами парных сравнений. Матрицу (x_{ijst}) можно понимать как матричную запись значений истинности n бинарных отношений предпочтения P_1, \dots, P_n соответствующих предпочтениям n экспертов: $P_s(a_i^{st}, a_j^{st}) \Leftrightarrow (x_{ijst} = 1)$. Кроме того, у нас определено отношение равенства = между объектами. Равенство $a_i^{st} = a_j^{st}$ определено для объектов a_i^{st}, a_j^{st} , сравниваемых экспертом s в сравнении t , и истинно тогда и только тогда, когда эти объекты совпадают.

Определим ещё отношение эквивалентности \sim , указывающее, что в разных сравнениях с разными экспертами участвует один и тот же объект из $A = \{a_1, \dots, a_m\}$, $a_i^{s_1t_1} \sim a_j^{s_2t_2} \Leftrightarrow i = j$. Множеством отношений Ω в этом случае является множество $\Omega = \{=, \sim, P_1, \dots, P_n\}$. Определим эмпирическую систему, являющуюся представлением матрицы (x_{ijst}) . Пусть $A = \{a_i^{st}\}$. Только одно отношение \sim из V определено на всем множестве A . Отношения P_s определены только на таких парах объектов $a_i^{s_1t_1}, a_j^{s_2t_2}$, для которых $t_1 = t_2, s_1 = s_2$. Введем для отношений из Ω третье значение истинности «не определено». Доопределим отношения $=, P_1, \dots, P_n$ на всем множестве A с помощью этого значения. Тем самым мы определили предикаты из Ω на всем множестве A , что дает нам эмпирическую систему $\mathcal{S} = \langle A, \Omega \rangle$.

4.2. Множественные сравнения [Шмерлинг Д.С., 1978]. Пусть дано множество объектов $A = \{a_1, \dots, a_m\}$. Группе из n экспертов поочередно предъявляются все возможные наборы из k объектов множества A . Каждый эксперт должен упорядочить каждый набор в соответствии с некоторым предпочтением. Обозначим через a_i^{tsl} тот факт, что объект с номером i в наборе с номером t экспертом s был поставлен на l -е место, $i = 1, \dots, m$; $s = 1, \dots, n$; $t = 1, \dots, C_m^k$; $l = 1, \dots, k$. Множество полученных упорядоченных наборов обозначим через $R = \{\langle a_{i_1}^{tsl}, a_{i_2}^{ts2}, \dots, a_{i_k}^{tsk} \rangle\}$.

Целью методов множественного сравнения является построение результирующего упорядочения объектов по полученным упорядочениям из R . Поставим в соответствие каждому эксперту s отношение предпочтения $P_s(a_{i_1}^{tsl_1}, a_{i_2}^{tsl_2}) \Leftrightarrow l_1 < l_2$. Определим два отношения эквивалентности \sim и \sim_t :

$$a_{i_1}^{t_1s_1l_1} \sim a_{i_2}^{t_2s_2l_2} \Leftrightarrow i_1 = i_2;$$

$$a_{i_1}^{t_1s_1l_1} \sim_t a_{i_2}^{t_2s_2l_2} \Leftrightarrow t_1 = t_2;$$

и отношение равенства =

$$a_{i_1}^{t_1s_1l_1} = a_{i_2}^{t_2s_2l_2},$$

истинное тогда и только тогда, когда в сравнении объектов из набора с номером t экспертом s объекты с именами $a_{i_1}^{t_1s_1l_1}$ и $a_{i_2}^{t_2s_2l_2}$ равны между собой. Множеством отношений для методов множественного сравнения будем множество $\Omega = \{=, \sim, \sim_t, P_1, \dots, P_n\}$. Информация извлеченная из данных R задается эмпирической системой, определенной на множестве $A = \{a_i^{tsl}\}$, $s = 1, \dots, n$; $s =$

$1, \dots, C_m^k; i = 1, \dots, m; l = 1, \dots, k$. Отношения из Ω доопределяются на всем множестве A с помощью значения «не определено». В результате мы получили эмпирическую систему $\mathfrak{S} = \langle A, \Omega \rangle$.

4.3. Матричное представление бинарных отношений. Бинарное отношение $P(a, b)$, определенное на множестве объектов $A = \{a_1, \dots, a_m\}$, задается матрицей (e_{ij}) , $i, j = 1, \dots, m$; где $e_{ij} = 1(0)$ означает, что $P(a_i, a_j)$ истинно (ложно). Такой матрицей можно задать произвольное бинарное отношение на множестве A . Такое представление широко используется в работе [Миркин Б.Г., 1980, Шрейдер С. А., 1983] ввиду его привычности и простоты. Наиболее часто используются отношения эквивалентности, квазипорядка, частичного порядка и лексикографического порядка.

Матрица бинарного отношения фиксирует некоторое бинарное отношение P , которое включается во множество Ω эмпирической системы $\mathfrak{S} = \langle A, \Omega \rangle$, $\Omega = \{P\}$.

Приведем результаты теории измерений, относящиеся к бинарным отношениям.

4.3.1. Отношение толерантности:

$$P(a, a); \\ P(a, b) \Leftrightarrow P(b, a).$$

4.3.2. Отношение эквивалентности:

$$P(a, a); \\ P(a, b) \Leftrightarrow P(b, a); \\ P(a, b) \& P(b, c) \Rightarrow P(a, c).$$

4.3.3. Отношение частичного порядка, для любых $a, b, c \in A$:

$$P(a, a); \\ P(a, b) \& P(b, c) \Rightarrow P(a, c).$$

Числового представления не существует.

4.3.4. Отношение интервального упорядочения, для любых $a, b, c, d \in A$:

$$\neg P(a, a); \\ P(a, b) \& P(c, d) \Rightarrow (P(a, d) \vee P(c, b)).$$

Числовое представление существует. Существуют две вещественнозначные функции $U, s: A \rightarrow \mathbb{R}^+$, такие, что для любых $a, b \in A$

$$P(a, b) \Leftrightarrow (U(a) + s(a)) < U(b).$$

4.3.5. Отношение полупорядка. Отношение P называется отношением полупорядка, если оно является отношением интервального порядка и для любых $a, b, c, d \in A$ удовлетворяет аксиоме

$$P(a, b) \& P(b, c) \Rightarrow P(a, d) \vee P(d, c).$$

Числовое представление существует. Существует вещественнозначная функция $U: A \rightarrow \mathbb{R}$ такая, что для любых $a, b \in A$

$$P(a, b) \Leftrightarrow (U(a) + 1) < U(b).$$

4.3.6. Отношение древовидного порядка. Отношение P называется отношением древовидного порядка, если оно является отношением строгого частичного порядка и для любых $a, b, c \in A$ удовлетворяет аксиоме

$$P(a, b) \& P(a, c) \Rightarrow (P(b, c) \vee P(c, b)).$$

Числового представления не существует.

4.3.7. Отношение квазипорядка, для любых $a, b, c \in A$ удовлетворяет аксиомам

$$P(a, a); \\ P(a, b) \& P(b, c) \Rightarrow P(a, c).$$

Числового представления не существует.

4.3.8. Отношение слабого порядка, для любых $a, b, c \in A$ удовлетворяет аксиомам

$$P(a, b) \vee P(b, a); \\ P(a, b) \& P(b, c) \Rightarrow P(a, c).$$

Если упорядоченная система $\langle A; P \rangle$ имеет счетную базу, то числовое представление существует [Шрейдер С. А., 1983].

Не все из приведенных отношений имеют числовые представления. Поэтому не всегда данные, содержащие бинарные отношения, можно представить в некотором числовом пространстве.

Рассмотрим, какие существуют методы обработки бинарных отношений. Большинство методов используют для обработки матриц расстояния или меры близости между матрицами. Эти расстояния и меры вводятся, либо на основании систем аксиом, либо из статистических предположений и свойств отношений.

4.4. Матрицы упорядочений: (r_{ij}) , $i = 1, \dots, m$; $j = 1, \dots, n$; r_{ij} – оценка i -го объекта по j -му признаку. Такие матрицы могут выражать, либо упорядочения k объектов n экспертами, либо упорядочения k объектов по n ранговым признакам. Такие матрицы обрабатываются методами многомерного шкалирования [Терехина А. Ю., 1983], а также методами обработки матричного представления бинарных отношений.

Поставим в соответствие каждому признаку j отношение P_j , определенное следующим образом:

Получим множество Ω бинарных отношений $\Omega = \{P_1, \dots, P_n\}$. Пусть $A = \{a_1, \dots, a_m\}$ – множество объектов, на которых получена матрица упорядочений. Тогда эмпирической системой будет модель $\mathfrak{S} = \langle A, \Omega \rangle$.

4.5. Матрицы близости. Пусть дано некоторое множество объектов $A = \{a_1, \dots, a_m\}$. Матрицей близости для этих объектов называется матрица (r_{ij}) , $i, j = 1, \dots, m$; r_{ij} – числовые оценки меры близости (сходства или различия) в порядковой шкале (имеет смысл только сравнение величин $r_{i_1j_1} < r_{i_2j_2}$). Такие матрицы возникают в различных областях при сравнении или оценке экспертом двух объектов в некотором отношении.

Матрицы близости обрабатываются методами многомерного неметрического шкалирования [Терехина А. Ю., 1983]. Целью этих методов является представление объектов точками в некотором метрическом пространстве (Евклидовом или Римановом) минимальной размерности так, чтобы расстояния t_{ij} между ними с точностью до порядка соответствовали бы величинам r_{ij} . Некоторые из этих методов в том же метрическом пространстве, называемом в этом случае объединенным психологическим пространством, представляют также и экспертов. Экспертам ставятся в соответствие точки, прямые или какие-либо другие подмножества метрического пространства. Каждый метод исходит из некоторой модели взаимодействия объекта и субъекта.

После применения методов многомерного шкалирования мы получаем представление данных в метрическом пространстве. Эти данные можно записать в виде матрицы объект-признак, которые рассмотрены ниже.

Определим на множестве A отношение

$$P(a_{i_1}, a_{i_2}, a_{i_3}, a_{i_4}) \Leftrightarrow r_{i_1i_2} < r_{i_3i_4}.$$

Так как это отношение определено на всем множестве A , то эмпирической системой будет модель $\mathfrak{S} = \langle A, \Omega \rangle$.

В теории измерений эмпирические системы, включающие подобные четырехместные отношения, обозначаются как $M = \langle A^*; \leq \rangle$, где $A^* \subset A \times A$, \leq – бинарное отношение упорядочения, определенное на A^* . Приведем некоторые результаты теории измерений, относящиеся к таким эмпирическим системам.

4.5.1. Шкала положительных разностей [Krantz D.H. et al, 1971, 1989, 1990; с. 147]. Существует гомоморфизм $\Phi : A^* \rightarrow \text{Re}$, $A \neq \emptyset$, такой, что для любых пар (a, b) , (b, c) , (c, d) из A^* :

$$(a, b) \leq (c, d) \Leftrightarrow \Phi(a, b) \leq \Phi(c, d), \\ \Phi(a, c) = \Phi(a, b) + \Phi(b, c).$$

Отображение Φ единственно с точностью до положительного множителя (шкала отношений).

4.5.2. Шкала алгебраических разностей [Там же; с. 151]: $A^* = A \times A$. Существует гомоморфизм $\Phi : A \rightarrow \text{Re}$ такой, что для любых $a, b, c, d \in A$

$$(a, b) < (c, d) \Leftrightarrow (\Phi(a) - \Phi(b)) < (\Phi(c) - \Phi(d)).$$

Отображение Φ , обладающее этим свойством, единственно с точностью до лог-линейных преобразований (шкала интервалов).

4.5.3. Шкала разностей равных конечных промежутков [Там же; с. 168]: $A^* = A \times A$, A – конечно, $A^* \neq \emptyset$. Существует гомоморфизм $\Phi : A \rightarrow \text{N}$ (натуральные числа), такой, что для любых $a, b, c, d \in A$

$$(a, b) \leq (c, d) \Leftrightarrow \Phi(a) - \Phi(b) \leq \Phi(c) - \Phi(d).$$

Отображение Φ единственно с точностью до линейных преобразований (шкала интервалов).

4.5.4. Шкала абсолютных разностей: [Там же; с. 172]: $A^* = A \times A$. Существует гомоморфизм $\Phi : A \rightarrow \text{Re}$ такой, что

$$P_j(a_{i_1}, a_{i_2}) \Leftrightarrow r_{i_1j} < r_{i_2j}$$

$$(a, b) < (c, d) \Leftrightarrow |\Phi(a) - \Phi(b)| < |\Phi(c) - \Phi(d)|.$$

Отображение Φ единственно с точностью до линейных преобразований (шкала интервалов).

6. Матрица объект-признак (x_{ij}) , $i = 1, \dots, m$; $j = 1, \dots, n$; x_{ij} – числовое значение j -го признака на i -м объекте. Признаки могут быть самыми произвольными как количественными, так и качественными. Тот факт, что такая матрица получена в результате некоторых измерений (опросов, экспериментов, обследований и т. д.), говорит о том, что существует n приборов или измерительных процедур, сопоставляющих каждому из m объектов числовые значения $x_{ij} = x_j(a_i)$ соответствующих признаков.

Данные такого типа имеют наибольшее распространение: анкетирование, тестирование, разнообразные социологические опросы, экспертное оценивание, карты обследований, геологоразведка, экспериментальные данные и т. д. Большинство известных методов предназначено для обработки именно таких данных.

Сопоставим каждому признаку x_i словарь Ω_i . Рассмотрим два случая:

1. Прибор x_i является хорошо изученным прибором, например, измеряющим некоторую физическую величину, и решаемая задача относится к области физики. Тогда множество Ω отношений и операций известно [Krantz D.H. et al, 1971, 1989, 1990] и эмпирической системой будет модель $\mathfrak{S} = \langle A, \Omega \rangle$.

2. Эмпирическая система прибора x_i не полностью или не достаточно точно определена, либо решаемая задача не относится к области физики. Такие измерения называют приборными [Пфанцagl И., 1976] или косвенными измерениями. Примерами таких измерений являются различные результаты тестирования, социологического опроса, балльные оценки, субъективные оценки и т. д. Все эти величины характеризуются тем, что предметная область, в рамках которой они рассматриваются, недостаточно разработана и поэтому эмпирические системы величин не полностью известны (хотя сам прибор, как, например, физические приборы известны хорошо). В этом случае, как уже говорилось, прибор или тестирование дают нам косвенные измерения интересующих нас величин.

Рассмотрим, как можно определить словарь Ω_i приборных измерений.

Для любого числового отношения $R(y_1, \dots, y_k)$, определенного на Re (множестве действительных чисел), можно определить следующее эмпирическое отношение на множестве объектов A :

$$P_j^R(a_1, \dots, a_k) \Leftrightarrow R(x_j(a_1), \dots, x_j(a_k)).$$

Это отношение может не иметь эмпирической интерпретации. Прибор $x_j(a)$ имеет эмпирическую интерпретацию, но связь его значений отношением R может уже не иметь эмпирическую интерпретацию. Поэтому нужно найти такие числовые отношения на Re , для которых отношение P_j^R имело бы эмпирическую интерпретацию. Предположим, что мы перебрали некоторые, наиболее распространенные числовые отношения и нашли, что отношения $P_j^{R_1}, \dots, P_j^{R_k}$ имеют эмпирическую интерпретацию. Данное множество отношений не пусто, так как, по крайней мере, отношение $P_j^=$ имеет эмпирическую интерпретацию. Если имеет смысл величина $x_j(a_1)$, то смысл отношения

$$P_j^=(a_1, a_2) \Leftrightarrow x_j(a_1) = x_j(a_2)$$

состоит в том, что на объектах a_1 и a_2 величина x_j принимает одно и то же значение. Отношение $P_j^=$, как правило, является отношением эквивалентности. В теории измерений известно много систем аксиом, использующих для некоторых величин только отношение $P_j^=$, и приводящих, тем не менее, к сильным шкалам. Определим словарь Ω_i приборного измерения x_j как множество $\{P_j^{R_1}, \dots, P_j^{R_k}\}$. В качестве множества отношений и операций эмпирической системы $\mathfrak{S} = \langle A, \Omega \rangle$ будет множество $\Omega = \Omega_1 \cup \dots \cup \Omega_n$.

Из приводимых примеров можно понять, как другие типы данных могут быть представлены эмпирическими системами. Общим аргументом в пользу универсальности извлечения информации с помощью эмпирических систем является методологический принцип теории измерений, состоящий в том, что отношения первичны, а свойства (числовые представления) вторичны. Свойства – это сжатое, закодированное числами представление отношений.

Литература

- Витяев Е.Е. Компьютерное познание. Информационные технологии в гуманитарных исследованиях, Вып. 13, ИАЭТ СО РАН, Новосибирск, 2009, стр. 81-87
- Фридланд А.Я. Информатика: процессы, системы, ресурсы. -М.: БИНОМ. Лаборатория знаний, 2003.
- Бешенков С.А., Ракитина Е.А. Моделирование и формализация. Методическое пособие. – М.: Лаборатория Базовых Знаний, 2002. – 336с.
- Лихорадка // Малая медицинская энциклопедия, М.

- Витяев Е.Е.** Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов. Новосибирск, 2006. 293 с.
- Е. 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 Renolls, Evgenii Vityaev. v.12(2), IOS Press, 2008, pp. 189-210
- Krantz D.H., Luce R.D., Suppes P., Tversky A.** Foundations of Measurement. Acad. Press, N.Y.; L. 1971; 1989; 1990. Vol. 1-3.
- Пфанцгль И.** Теория измерений. М.: Мир, 1976. 248 с.
- Девид Г.** Метод парных сравнений. М.: Статистика, 1978. 150 с.
- Шмерлинг Д. С.** О построении моделей парных и множественных сравнений со связями // Прикладной многомерный статистический анализ. М., 1978. С. 164-189.
- Миркин Б.Г.** Анализ качественных признаков и структур. М.: Статистика, 1980. 316 с.
- Шрейдер С. А.** Систематика, типологии, классификация // Теория и методология биологических классификаций, М.: Наука, 1983.
- Терехина А. Ю.** Методы многомерного шкалирования и визуализации данных: (Обзор) // Автоматика и телемеханика. 1973. № 7. С. 80-94.