

Фрактальность ограничений сосуществования свойств в задачах машинного обучения

Д.Е. Самойлов^{1,2}, В.А. Семенова², С.В. Смирнов²

¹Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086

²Институт проблем управления сложными системами РАН, ул. Садовая 61, Самара, Россия, 443020

Аннотация. Рассматривается априорное знание, устанавливающее, как факты обнаружения у эмпирически наблюдаемых объектов одних свойств влияют на наличие других. Учет таких ограничений сосуществования свойств у объектов изучаемой предметной области необходим для корректного формирования контекстов задач машинного обучения в условиях неполноты и противоречивости исходных данных. В статье представлена новая многоуровневая рекурсивная (фрактальная) модель таких ограничений. Модель строится на основе точки зрения о логико-понятийном происхождении ограничений с использованием элементарных бинарных отношений на множестве измеряемых свойств - несовместимости и обусловленности. Описывается эвристический метод и алгоритм для однозначной («четкой») аппроксимации «мягкого» контекста при действии ограничений сосуществования свойств.

1. Введение

Одной из определяющих сущностей машинного обучения является обучающая выборка. Как правило, она структурируется в форме таблицы «объекты-свойства» [1, 2] или ее экстенсивных расширений [2, 3]. Таблица содержит результаты измерений априори заданного набора свойств у всех объектов обучающей выборки и часто именуется многозначным формальным контекстом (ФК) задачи обучения (и, шире, анализа данных) [4, 5].

Важно отметить, что для многих задач машинного обучения многозначный ФК должен быть предварительно преобразован в однозначный [3-5], который описывает бинарное соответствие «объекты-свойства», т.е. совокупность оценок истинности $\|b_{xy}\| \in \{\text{Истина}, \text{Ложь}\}$ базовых семантических суждений (БСС) об изучаемой предметной области (ПрО) вида $b_{xy} =$ «объект x обладает свойством y ».

Трансформация многозначного ФК в однозначный сводится к замене исходных табличных значений на истинностные константы **Истина** и **Ложь**:

- Константой **Истина** заменяются все значения кроме **None**.
- Константой **Ложь** заменяются значения **None** - специальные результаты измерения любого свойства объекта изучаемой ПрО, когда у соответствующей процедуры измерения обнаруживается либо «семантическое несоответствие» с исследуемым объектом, либо значение измеряемого свойства лежит вне динамического диапазона средства измерений [6, 7].

С общих позиций, значения, равные информационной константе **None**, появляются в многозначном ФК в результате выполнения когнитивного приема, известного как концептуальное шкалирование свойств [8, 9]. Другое обобщающее обстоятельство связано с тем, что исходный многозначный ФК отражает обсуждавшиеся в [3] фундаментальные реалии накопления эмпирической информации, предопределяющие ее неполноту и противоречивость.

Поэтому преобразование многозначного ФК в однозначный в общем случае включает два этапа:

- Свойства исходного многозначного ФК подвергаются концептуальному шкалированию, и видоизмененный многозначный контекст трансформируется в однозначный с использованием истинностных констант той или иной многозначной логики, выбираемой для моделирования неполноты и противоречивости данных многозначного ФК. Возникающие для каждого БСС независимые оценки истинности совмещаются по правилам избранной многозначной логики. В итоговом однозначном ФК соответствие «объекты-свойства» оказывается «мягким» (нечетким, нестрогим и т.п. – конкретизация определяется терминологией примененной многозначной логики – см., например, [10, 11]);
- В однозначном ФК «мягкое» бинарное отношение «объекты-свойства» замещается его сечением уровня α , где истинностное значение α используемой многозначной логики указывает субъективно устанавливаемый порог доверия к исходным данным.

Коллизия этих действий состоит в том, что стандартная процедура α -сечения не учитывает зависимости существования одних свойств объектов от существования других их свойств – ограничений сосуществования свойств (ОСС), - которые обычно имеются в исследуемой ПрО и, во всяком случае, возникают в результате концептуального шкалирования свойств многозначного ФК [12].

Ясно, что в основу метода корректной трансформации многозначного ФК в однозначный должна быть положена адекватная модель ОСС. В этом смысле предложенная ранее авторами двухуровневая модель ОСС [13] отражает лишь одну из простых, пусть и распространенных на практике ситуаций.

В данной статье представлена значительно более общая модель ОСС, которая обладает свойствами рекурсивности (фрактальности) и основана на представлении о логико-понятийном генезисе ОСС, сопровождаемом возникновением бинарных отношений несовместимости и обусловленности на множестве измеряемых свойств. Дополнительно очерчиваются использующие эту модель метод и алгоритм корректного α -сечения «мягкого» соответствия «объекты-свойства».

2. Рекурсивная модель ограничений сосуществования свойств

2.1. Постановка задачи

Исследуем генезис ОСС исходя из следующих достоверных тезисов:

- Решение измерять у объектов некоторое свойство z есть следствие выдвижения гипотезы о присутствии в изучаемой ПрО объектов, обладающих таким свойством [7]. Другими словами, априори вводится предположение, что изучаемую ПрО характеризует понятие с содержанием $\{z\}$ (строго говоря, с содержанием, которое включает z ; используемое упрощение непринципиально для дальнейшего анализа и существенно сокращает обозначения) и некоторым гипотетическим объемом $\{z\}'$ (здесь «'» - оператор Галуа бинарного соответствия, в данном случае соответствия «объекты-свойства»).
- Концептуальное шкалирование свойства z формально представляет собой субъективное конструирование концептуальной шкалы - невырожденного, содержащего не менее двух элементов покрытия динамического диапазона D_z процедуры измерения этого свойства (или домена значений этого свойства) [8, 12], - и по определению способно только добавлять в многозначный ФК новые измеряемые свойства.

Таким образом, априорное формирование в ФК набора измеряемых свойств можно рассматривать как его последовательное расширение либо при «непосредственном» добавлении новых свойств и соответствующих априорных гипотез, либо путем концептуального шкалирования уже присутствующих в ФК свойств.

При этом и в том и другом случае резонно ожидать изменений в уже сложившихся априорных понятиях об изучаемой ПрО. Например, при концептуальном шкалировании свойства z понятие, описываемое парой «объем-содержание» $(\{z\}', \{z\})$, неизбежно претерпит некую метаморфозу в силу реструктуризации его содержания.

В логике известны лишь две возможности трансформации понятий, связанных с расширением набора рассматриваемых признаков (свойств) исследуемых объектов: деление и ограничение [14]. Поэтому при исследовании процесса формирования набора измеряемых свойств ФК путем указанных выше механизмов его расширения, уместна постановка следующих вопросов:

- Как реализуется при этом деление и ограничение априорно выдвинутых гипотетических понятий об изучаемой ПрО?
- Какие образуются при этом зависимости существования одних гипотетических свойств объектов изучаемой ПрО от существования других их свойств?

2.2. Деление гипотетического понятия

Деление понятия связывается с перечислением всех непересекающихся частей его объема по некоторому одному основанию – признаку (свойству), который видоизменяется в рамках делимого объема [14].

Пусть делимым является гипотетическое понятие $(\{z\}', \{z\})$, и роль основания деления играет свойство z . Видоизменение свойства в рамках делимого объема $\{z\}'$ естественно трактовать как характерное изменение значения свойства z при переходе между некоторыми дизъюнктивными частями диапазона D_z или (в терминах концептуального шкалирования) между дизъюнктивными элементами покрытия этого диапазона, количество которых, вообще говоря, произвольно, но, разумеется, не меньше двух:

$$D_z = \cup_{i=1}^n D_z(z_i),$$

где $z_i, i = 1, \dots, n, n \geq 2$ – новые измеряемые свойства, которые в результате выполнения процедуры концептуального шкалирования замещают в ФК свойство z ; $D_z(z_i)$ – элемент покрытия, некоторая непустая часть динамического диапазона процедуры измерения свойства z , которая составляет домен свойства $z_i, D_z(z_i) \cap D_z(z_j) = \emptyset$ при $i \neq j$.

Поскольку измеренное значение свойства z в рассматриваемом случае будет принадлежать одному и только одному домену $D_z(z_i)$, то новые измеряемые свойства $z_i, i = 1, \dots, n$, вводимые в ФК при делении гипотетического понятия $(\{z\}', \{z\})$ будут несовместимыми.

Наконец, следует отметить, что в априорных гипотезах об изучаемой ПрО делимое понятие $(\{z\}', \{z\})$ замещается набором понятий $(\{z_i\}', \{z_i\}), i = 1, \dots, n$, для которых справедливо $\{z_i\}' \cap (\{z_j\}') = \emptyset$ при $i \neq j$ и $\cup_{i=1}^n \{z_i\}' = \{z\}'$.

2.3. Ограничение гипотетического понятия

Ограничение понятия означает введение нового понятия, объем которого составляет часть объема исходного понятия, а содержание отличается дополнительным признаком (свойством), присущим лишь той части мыслимых в исходном понятии предметов (объектов), которая образует объем нового вводимого понятия [14].

В наших обозначениях ограничением гипотетического понятия $(\{z\}', \{z\})$ будет понятие $(\{z, z_1\}', \{z, z_1\})$, где z_1 – новое измеряемое свойство, добавляемое в ФК, и $\{z, z_1\}' \subseteq \{z\}'$.

Тогда согласно определению операции ограничения обнаружение у объекта изучаемой ПрО свойства z_1 достоверно свидетельствует о принадлежности этого объекта объему $\{z\}'$. Но поскольку все объекты в $\{z\}'$ обладают свойством z , следовательно, наличие у исследуемого

объекта свойства z_1 обуславливает наличие у него свойства z . Отсюда также вытекает, что всякое измеряемое свойство самообусловлено.

Очевидно, что понятие $(\{z\}', \{z\})$ может иметь, вообще говоря, произвольное число различных ограничений вида $(\{z, z_i\}', \{z, z_i\})$, где $z_i, i = 1, \dots, n, n \geq 1$ – новые измеримые свойства, вводимые в ФК с условием $\{z, z_i\}' \cap (\{z, z_j\}' = \emptyset$ при $i \neq j$ и $\{z, z_i\}' \subseteq \{z\}'$.

Ограничиваемое понятие $(\{z\}', \{z\})$ сохраняется как часть априорных предположений об изучаемой ПрО и становится родовым (обобщающим) для всех его ограничений вида $(\{z, z_i\}', \{z, z_i\})$, которые дополняют гипотетические представления об изучаемой ПрО.

2.4. Рекурсивное расширение гипотетических предположений и ОСС

Первичное решение измерять у исследуемых объектов некоторое свойство z задает исходную априорную гипотезу об изучаемой ПрО (разумеется, количество подобных протосвойств и соответствующих гипотез произвольно). Деление или ограничение гипотетического понятия $(\{z\}', \{z\})$ приводит к расширению набора априорных гипотез об изучаемой ПрО, которое может быть рекурсивно продолжено.

Действительно, каждое новое понятие, порожаемое в результате выполнения этих логических операций, может быть в свою очередь подвергнуто делению или ограничению по описанным выше правилам и только по ним. Таким образом, рекурсивная процедура расширения гипотетических предположений об изучаемой ПрО включает произвольный выбор и затем конкретную реализацию одного из двух имеющихся способов трансформации существующего априорного понятия об изучаемой ПрО.

При каждом выполнении этой рекурсивной процедуры в ФК вводятся новые измеряемые свойства. Прделанный анализ позволяет утверждать, что ограничения сосуществования этих свойств у любого объекта изучаемой ПрО исчерпываются двумя и только двумя бинарными отношениями: обусловленностью («имея свойство x , объект обязательно обладает и свойством y , но обратное неопределенно») и несовместимостью («имея свойство x , объект заведомо лишен свойства y , и наоборот»).

2.5. Модель ограничений сосуществования свойств

В основу разработанной модели положено очерченное выше представление о логико-понятийном генезисе ОСС. При таком подходе присутствие ОСС понимается как следствие существования развитых гипотетических предположений об изучаемой ПрО. Эти гипотетические понятия структурированы в виде, вообще говоря, нескольких самоподобных, рекурсивных иерархий. Поэтому предлагаемая модель ОСС имеет аналогичную структуру с отмеченными фрактальными свойствами.

Стержневым элементом модели является группа сопряженных свойств (ГСС):

- Измеряемое свойство есть одиночная ГСС, и единственный член группы самообусловлен.
- Набор измеряемых свойств, введенный в ФК в результате деления гипотетического понятия, есть множественная ГСС, члены которой несовместимы.
- Набор измеряемых свойств, введенный в ФК в результате гнездования нескольких ограничений гипотетического понятия, есть множественная ГСС, члены которой обусловлены «снизу до верху».

В целом модель ОСС может быть изображена как граф, в котором вершины суть измеряемые свойства, дуги фиксируют обусловленность свойств, а ребра – несовместимость. Граф состоит из нескольких несвязных компонент, каждая из которых соответствует развитию гипотетических представлений об изучаемой ПрО, начиная с предположения о существовании некоторого протосвойства.

В рамках изолированной компоненты граф полносвязен, а его вершины распределены между ГСС, которые связаны между собой в рамках иерархической рекурсивной структуры, формирующейся при делении и ограничении соответствующего протосвойства. В рамках изолированной компоненты несовместимость свойств симметрична, антирефлексивна и нетранзитивна, а обусловленность асимметрична, рефлексивна и транзитивна.

3. Метод и алгоритм корректной аппроксимации формального контекста

Введение новой модели ОСС, имеющей «фрактальную генетику» не влечет пересмотр эвристики, предложенной еще в [15] для коррекции стандартного α -сечения «мягкого» однозначного ФК:

- выбор порога доверия к исходным данным произволен;
- соответствующий порогу и, в общем случае, недопустимый состав свойств каждого объекта целевого однозначного ФК следует последовательно сокращать за счет отсекаемых на каждом шаге свойства, нарушающего ОСС;
- механизм отсекаемых состоит в локальном ужесточении порога доверия в пределах каждой ГСС модели ОСС, когда основным критерием выбора измеряемого свойства для отсекаемых служит минимальное ужесточение выбранного порога доверия [11, 13].

Разумеется, алгоритм коррекции стандартного α -сечения «мягкого» соответствия «объекты-свойства» при новой модели ОСС существенно изменяется. Исключается специфика, которая была связана с использованием предложенной ранее двухуровневой модели ОСС [13], приобретаются рекурсивные составляющие для обработки фрактальных черт новой иерархической модели ОСС.

4. Заключение

Предложенная модель ОСС обобщает формальные схемы, ранее построенные в работах авторов, и по-прежнему базируется на элементарных ограничениях сосуществования любых сущностей – несовместимости и обусловленности (нейтральность, безразличие традиционно исключается из ограничений, но неявно почти всегда учитывается; наша модель в этом смысле так же не является исключением).

Применительно к анализу данных и машинному обучению такая узкая номенклатура элементарных ограничений сосуществования свойств (как и сам термин «properties existence constraints») впервые явно предложена в [16], где она выявлена в результате переопределения семантических ограничений теории реляционных баз данных [17]. При этом в работе [16], а затем в статье [18] ОСС рассматриваются не как гипотетическое, а достоверное исходное знание об изучаемой ПрО, которое используется при построении однозначных ФК. Подобный подход в принципе известен в распознавании образов (см., например, [19]).

Обоснование, предложенное в нашей работе, не только прямо адресуется к фундаментальным законам классической логики, но и связывает ОСС с гипотетическими предположениями, которые необходимо вводятся при традиционном формировании обучающих выборок как результатов измерения объектов изучаемой ПрО.

Весьма важным оказалось и то, что существенное усложнение и фрактальная специфика новой модели ОСС не дезавуировала ранее разработанный эвристический подход к построению непротиворечивых однозначных ФК в задачах машинного обучения.

5. Благодарности

Статья подготовлена по материалам научных исследований в рамках субсидированного государственного задания Институту проблем управления сложными системами РАН на НИР по теме «Интерсубъективное управление инновационным развитием социотехнических объектов с применением онтологических моделей ситуаций».

6. Литература

- [1] Барсегян, А.А. Анализ данных и процессов / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. – СПб.: БХВ-Петербург, 2009. – 512 с.
- [2] Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Институт математики СО РАН, 1999. – 270 с.
- [3] Смирнов, С.В. Двухсоставность феномена информации и анализ данных (с примерами из когнитивного анализа) / С.В. Смирнов // Информационные технологии и нанотехнологии (ИТНТ-2017). - Самара: Новая техника, 2017. – С. 1846-1849.

- [4] Ganter, B. Formal Concept Analysis. Mathematical foundations / B. Ganter, R. Wille. – Berlin-Heidelberg: Springer-Verlag, 1999. – 290 p.
- [5] Formal Concept Analysis Homepage [Electronic resource]. – Access mode: <http://www.upriss.org.uk/fca/fca.html> (11.11.2017).
- [6] Рубашкин, В.Ш. Представление и анализ смысла в интеллектуальных информационных системах / В.Ш. Рубашкин. – М.: Наука, 1989. – 192 с.
- [7] Смирнов, С.В. Онтологический анализ предметных областей моделирования / С.В. Смирнов // Известия Самарского научного центра РАН. – 2001. – Т. 3, №1. – С. 62-70.
- [8] Ganter, B. Conceptual scaling / B. Ganter, R. Wille // Applications of Combinatorics and Graph Theory to the Biological and Social Sciences. – New York Springer-Verlag, 1989. – P. 139-167.
- [9] Belohlavek, R. Scaling, Granulation, and Fuzzy Attributes in Formal Concept Analysis / R. Belohlavek, J. Konecny // The IEEE International Conference on Fuzzy Systems (London, UK, July 23-26, 2007). – P. 918-923.
- [10] Yang, K.M. Fuzzy Concept Mining based on Formal Concept Analysis / K.M. Yang, E.H. Kim, S.H. Hwang, S.H. Choi // Int. J. of Computers. – 2008. – Vol. 2(3). – P. 279-290.
- [11] Смирнов, С.В. Нестрогий анализ формальных понятий / С.В. Смирнов // Знания – Онтологии – Теории: Материалы Всероссийской конф. с международным участием. Т. 2. – Новосибирск: Институт математики СО РАН, 2015. – С. 142-150.
- [12] Самойлов, Д.Е. Субъективные аспекты формирования и обработки данных в анализе формальных понятий / Д.Е. Самойлов, С.В. Смирнов // Информационные технологии и нанотехнологии (ИТНТ-2016). – Самара: Изд-во СНЦ РАН, 2016. – С. 1038-1043.
- [13] Семенова, В.А. Модели и методы интеллектуального анализа неполных данных для построения формальных онтологий / В.А. Семенова, С.В. Смирнов // Информационные технологии и нанотехнологии (ИТНТ-2015). – Самара: Изд-во СНЦ РАН, 2015. – С. 192-196.
- [14] Ивин, А.А. Словарь по логике / А.А. Ивин, А.Л. Никифоров. – М.: Гуманит. изд. центр ВЛАДОС, 1997. – 384 с.
- [15] Офицеров, В.П. Метод альфа-сечения нестрогих формальных контекстов в анализе формальных понятий / В.П. Офицеров, В.С. Смирнов, С.В. Смирнов // Проблемы управления и моделирования в сложных системах: Труды XVI международной конф. – Самара: СНЦ РАН, 2014. – С. 228-244.
- [16] Lammari, N. Building and maintaining ontologies: a set of algorithms / N. Lammari, E. Metais // Data & Knowledge Engineering. – 2004. – Vol. 48(2). – P. 155-176.
- [17] Codd, E.F. The Relational Model for Database Management: Version 2 / E.F. Codd. – Addison-Wesley, 1990. – 538 p.
- [18] Пронина, В.А. Использование отношений между атрибутами для построения онтологии предметной области / В.А. Пронина, Л.Б. Шипилина // Проблемы управления. – 2009. – №1. – С. 27-32.
- [19] Закревский, А.Д. Логика распознавания / А.Д. Закревский – Мн.: Наука и техника, 1988. – 118 с.

Fractality of the object's properties existence constraints in machine learning

D.E. Samoylov^{1,2}, V.A. Semenova², S.V. Smirnov²

¹Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086

²Institute for the Control of Complex Systems RAS, Sadovaya street 61, Samara, Russia 443020

Abstract. Properties existence constraints for objects of the subject domain are necessary knowledge for constructing formal contexts in machine learning. The article presents a new multilevel (fractal) model of these constraints. The model combines the point of view on the genesis of constraints and the elementary binary relations of incompatibility and conditionality. An algorithm for “crisp” approximation of the “soft” formal context is described.

Keywords: machine learning, formal context, properties existence constraints, conceptual scaling, fractality.