Math-Net.Ru

Общероссийский математический портал

И. И. Астахова, В. Н. Вагин, М. В. Фомина, Проблема формирования обобщенных понятий при работе с неопределенными атрибутами, *Нечеткие системы и мягкие вычисления*, 2017, том 12, выпуск 2, 151–158

DOI: <https://doi.org/10.26456>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением

<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 54.198.64.247

26 апреля 2023 г., 14:34:01

ПРОБЛЕМА ФОРМИРОВАНИЯ ОБОБЩЕННЫХ ПОНЯТИЙ ПРИ РАБОТЕ С НЕОПРЕДЕЛЕННЫМИ АТТРИБУТАМИ¹

Астахова И.И., Вагин В.Н., Фомина М.В.

Национальный исследовательский университет «МЭИ», г. Москва

Поступила 04.09.2017, после переработки 15.12.2017.

Рассматривается проблема обобщения данных с учетом неопределенности в атрибутах. Рассматриваются алгоритмы, способные работать с нечеткими данными. Предлагаются варианты улучшения алгоритма UDT, обрабатывающего неопределенные атрибуты. Приводятся результаты экспериментов по применению алгоритма для работы с неопределенными атрибутами.

Ключевые слова: обобщение, неопределенный числовой атрибут, дерево решений.

Нечеткие системы и мягкие вычисления. 2017. Т. 12, № 2. С. 151–158.
<https://doi.org/10.26456/fssc30>

Введение

Проблема индуктивного формирования понятий является крайне важной при построении современных интеллектуальных систем. Прежде всего, это системы обобщения и системы поддержки принятия решений. Алгоритмы обобщения позволяют получать знания, объясняющие имеющиеся факты, а также формировать классификационные правила. Области применения алгоритмов обобщения являются медицина, банковская деятельность, прогнозирование и многое другое. Но сложность состоит в том, что в обучающих выборках по данным предметным областям могут быть атрибуты, заданные не конкретным значением, а интервалом допустимых значений. Классические алгоритмы обобщения не позволяют работать с этим типом атрибутов, поэтому предлагается исследовать методы обработки таких нечетко заданных атрибутов.

1. Постановка задачи обобщения

Пусть O — множество всех объектов, которые могут быть представлены в некоторой системе знаний, V — множество положительных объектов и W — множество отрицательных объектов. Будем рассматривать случай, когда

$$O = V \cup W, \quad V \cap W = \emptyset.$$

Сформируем множество $K = K^+ \cup K^-$ такое, что $K^+ \subset V$, $K^- \subset W$. Будем называть K обучающей выборкой. На основании обучающей выборки надо

¹Работа выполнена при финансовой поддержке РФФИ (гранты № 18-01-00201-а, № 17-07-00442-а).

построить правило, разделяющее положительные и отрицательные объекты обучающей выборки [1].

Понятие сформировано, если удалось построить решающее правило, которое для любого примера из обучающей выборки указывает, принадлежит ли этот пример понятию или нет. Рассмотрим алгоритмы, в которых решающие правила представлены в виде дерева решений [1].

2. Алгоритмы обобщения, работающие с реальными данными

Наиболее известными методами решения поставленной задачи являются методы, основанные на теории приближенных множеств, методы деревьев решений (алгоритмы ID3, C4.5) [3, 4].

Алгоритм ID3 (Induction of Decision Trees, разработан Р. Куинланом) формирует решающие деревья на основе примеров. Алгоритм начинает работу со всеми обучающими примерами в корневом узле дерева. Для разделения множества примеров корневого узла на основе критерия информативности выбирается один из атрибутов, и для каждого значения, принимаемого этим атрибутом, строится ветвь и создается дочерний узел. Затем все примеры распределяются по дочерним узлам в соответствии со значением атрибута. Алгоритм повторяется рекурсивно до тех пор, пока в узлах не останутся только примеры одного класса, после чего узлы будут объявлены листьями и разбиение будет завершено [3].

Алгоритм C4.5 [4] представляет собой усовершенствованный вариант алгоритма ID3. Стоит отметить следующие улучшения:

1. Возможность работать не только с качественными атрибутами, но также с числовыми.
2. После построения дерева происходит усечение его ветвей.

В реальных базах данных значения некоторых атрибутов могут быть неточными, а иногда и отсутствовать. Такие данные в обучающих выборках затрудняют работу алгоритмов обобщения. Алгоритмы построения деревьев решений, названные выше, требуют расширения возможностей при работе с такими атрибутами. Неточные значения могут возникнуть в результате ошибок измерений, устаревших данных, нескольких повторяющихся измерений, и т.д. В случае, если данные представляются несколькими значениями, для их описания может быть использован интервал. С таким интервалом связывается функция распределения вероятности появления различных значений. Неопределенность в атрибутах может возникать естественным путем, например, если учтена погрешность измерительного прибора и вместо точного значения в базу данных внесен некоторый диапазон. Атрибут, значения которого заданы интервалом, представляет собой неопределенный числовой атрибут (для краткости обозначаем его далее НЧА).

3. Методы обработки неопределенных числовых атрибутов

Для решения задачи построения дерева решений по таблицам, хранящимся в базе данных, которые содержат НЧА, можно использовать метод усреднения (AVG) [7]. Основная идея метода заключается в том, что в обучающем множестве все значения типа «интервал» заменяются на свое среднее значение. Так, например, если признак «доход» представлен интервалом значений от 10 до 40, то после усреднения можно заменить его значением 25. После применения усреднения

можно использовать хорошо изученные алгоритмы построения дерева решений, способные обрабатывать количественные атрибуты, например, алгоритм C4.5.

Следует учесть, что при замене интервала значений для атрибута на среднее неизбежно теряется часть информации. Поэтому был предложен способ обработки НЧА на основе статистических методов [5]. При таком подходе значение интервального атрибута рассматривается как случайная величина, которой сопоставлено некоторое распределение.

Так как каждое значение НЧА является интервалом, то границы интервала назовем критическими точками. Далее, если упорядочить значения критических точек и убрать повторяющиеся, то получим систему непересекающихся интервалов. Пусть один из таких интервалов задан как (a, b) . Каждый интервал может пересекаться с несколькими значениями НЧА многократно. Если значение интервального атрибута в конкретном примере пересекается с интервалом (a, b) , то его вероятность $\int_a^b f(x)dx$. На основании вероятности каждого отдельного экземпляра, пересекающегося с интервалом (a, b) , можно вычислить суммарную вероятность для всех экземпляров набора данных, которые пересеклись с интервалом (a, b) . Это число будет являться критерием выбора наиболее важного интервала неопределенного атрибута при построении дерева решений. Обозначим такой критерий РС как сокращение термина «probabilistic cardinality».

Для нахождения наилучшего интервала разбиения для каждого интервала рассчитывается величина критерия РС и далее выбирается интервал с максимальным значением этого критерия.

Пусть рассматривается интервал $(a, b) = Pa$. Первоначально следует рассчитать его критерий РС для общего числа примеров, т.е. для всего обучающего множества [5]:

$$PC(Pa) = \sum_{j=1}^n P(A_j^{u_n} \in (a, b)) = \sum_{j=1}^n \int_a^b A_j^{u_n} * f(x)dx,$$

где

- Pa — рассматриваемый интервал,
- $A_j^{u_n}$ — интервал значений для j -ого примера атрибута A^{u_n} ,
- n — общее количество рассматриваемых примеров.

Если значение атрибута не пересеклось с рассматриваемым интервалом, то $\int_a^b A_j^{u_n} * f(x)dx = 0$, во всех остальных случаях будет получено некоторое вещественное число.

4. Алгоритм UDT и его модификация

На основе описанного выше подхода был разработан алгоритм UDT (Uncertain Decision Trees) [6]. В соответствии с этим алгоритмом для неопределенного числового атрибута все критические точки сортируются по возрастанию. В результате сортировки и удаления дубликатов будут получены непересекающиеся подмножества, что уже уменьшает количество рассматриваемых интервалов.

Одним из способов оптимизации алгоритма UDT является увеличение его быстродействия. Для этого предлагается применить процедуру вывода из рассмотрения пустых интервалов.

Если рассматриваемый интервал не пересекает ни один интервал из сформированной для обучающего множества системы интервалов, то он называется пустым интервалом.

Все пустые интервалы можно исключить из рассмотрения, так как на коэффициент прироста информации они не влияют. Был разработан алгоритм UDT_OPT, включающий процедуру удаления пустых интервалов из рассмотрения.

5. Экспериментальные результаты

Было выполнено программное моделирование изложенных выше методов и алгоритмов. Эксперименты проводились на тестовых наборах данных из репозитория UCI калифорнийского университета информации и компьютерных наук [2]. Для сравнительного анализа проводилось два типа экспериментов.

На первом эксперименте наша цель — оценить точность классификации алгоритмов (AVG, UDT и UDT_OPT) на примере одного набора «Iris», в зависимости от максимального интервального отклонения².

Рис. 1: Результаты первого эксперимента

Второй эксперимент проводился на трех наборах данных «Iris», «Glass» и «Vehicle». Цель данного эксперимента — сравнить время выполнения алгоритмов (AVG, UDT и UDT_OPT).

Из проведенных экспериментов можно сделать вывод, что применение метода UDT_OPT значительно увеличивает точность классификации по сравнению с методом AVG и при этом уменьшает время, требуемое на выполнение по сравнению с методом UDT.

²Обозначение w1 означает, что максимальное интервальное отклонение от точного числа равно 1%, для остальных значений аналогично. Рассматриваются отклонения, равные: 5%, 10%, 20%, 40%.

Рис. 2: Результаты второго эксперимента

Заключение

В работе рассмотрена проблема обобщения данных, в которых используются нечеткие атрибуты, значениями которых могут быть множества. Таким множеством может быть конечный набор значений, либо числовой интервал. Рассмотрены алгоритмы обобщения, использующие разные стратегии обработки нечетких атрибутов и предложена модификация для уменьшения времени обработки таких атрибутов. Приведены результаты программного эксперимента.

Список литературы

- [1] Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В. Достоверный и правдоподобный вывод в интеллектуальных системах / Под ред. В.Н. Вагина, Д.А. Поспелова. М.: Физматлит, 2008. 712 с.
- [2] Merz C.J., Murphy P.M. UCI Repository of Machine Learning Datasets [Electronic resource] // Information and Computer Science University of California, 1998. URL: <http://archive.ics.uci.edu/ml/>.
- [3] Quinlan J.R. Induction of Decision Trees Machine Learning 1. 1986. Pp. 81–106.
- [4] Quinlan J.R. Improved Use of Continuous Attributes in C 4.5 // Journal of Artificial Intelligence Research. 1996. Vol. 4. Pp. 77–90. <http://dx.doi.org/10.1613/jair.279>
- [5] Qin B., Xia Y., Prabhakar S., Tu Y. A rule-based classification algorithm for uncertain data // The Workshop on Management and Mining Of Uncertain Data (MOUND). 2009.
- [6] Tsang S., Kao B., Yip K.Y., Ho W.S., Lee S.D. Decision Trees for Uncertain Data // IEEE 25th International Conference on Data Engineering (ICDE '09). Shanghai, China, 2009. Pp. 441–444. <https://doi.org/10.1109/ICDE.2009.26>

- [7] Xia Y., Xi B. Conceptual clustering categorical data with uncertainty // Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007). Patras, Greece, 2007. Pp. 329–336. <https://doi.org/10.1109/ICTAI.2007.135>

Образец цитирования

Астахова И.И., Вагин В.Н., Фомина М.В. Проблема формирования обобщенных понятий при работе с неопределенными атрибутами // Нечеткие системы и мягкие вычисления. 2017. Т. 12, № 2. С. 151–158. <https://doi.org/10.26456/fssc30>

Сведения об авторах

1. **Астахова Ирина Игоревна**

старший преподаватель кафедры вычислительной техники Национального исследовательского университета МЭИ.

Россия, 111250, г. Москва, Красноказарменная улица, д. 14, НИУ «МЭИ».

E-mail: ast_irina@mail.ru

2. **Вагин Вадим Николаевич**

профессор кафедры прикладной математики Национального исследовательского университета МЭИ.

Россия, 111250, г. Москва, Красноказарменная улица, д. 14, НИУ «МЭИ».

E-mail: vagin@appmat.ru

3. **Фомина Марина Владимировна**

доцент кафедры вычислительной техники Национального исследовательского университета МЭИ.

Россия, 111250, г. Москва, Красноказарменная улица, д. 14, НИУ «МЭИ».

E-mail: m_fomina2000@mail.ru

THE PROBLEM OF THE FORMATION OF GENERALIZED CONCEPTS WHEN WORKING WITH UNCERTAIN ATTRIBUTES

Astakhova Irina Igorevna

Senior lecturer at Computer Engineering department,
National Research University «Moscow Power Engineering Institute»
Russia, 111250, Moscow, Krasnokazarmennaya st., 14.
E-mail: ast_irina@mail.ru

Vagin Vadim Nikolaevich

Professor at Computer Engineering department,
National Research University «Moscow Power Engineering Institute»
Russia, 111250, Moscow, Krasnokazarmennaya st., 14.
E-mail: vagin@apmat.ru

Fomina Marina Vladimirovna

Associate professor at Computer Engineering department,
National Research University «Moscow Power Engineering Institute»
Russia, 111250, Moscow, Krasnokazarmennaya st., 14.
E-mail: m_fomina2000@mail.ru

Received 04.09.2017, revised 15.12.2017.

This paper considers the problem of data generalization, taking into account uncertain information in attributes. Algorithms capable of operating with uncertain data are considered. We suggest methods that improve UDT algorithm. The results of experiments on the application of the proposed algorithms are given.

Keywords: generalization problem, undefined numeric attribute, decision tree.

Nechetkie Sistemy i Myagkie Vychisleniya [Fuzzy Systems and Soft Computing], 2017, vol. 12(2), pp. 151–158. <https://doi.org/10.26456/fssc30>

References

- [1] Vagin V.N., Golovina E.Yu., Zagoryanskaya A.A., Fomina M.V. *Dostoverniy i Pravdopodobnyy Vывод v Intellektual'nykh Sistemakh* [A Reliable and Plausible Derivation in Intellectual Systems]. Eds. by V.N. Vagin, D.A. Pospelov. Fizmatlit Publ., Moscow, 2008. 712 p. (in Russian)
- [2] Merz C.J., Murphy P.M. *UCI Repository of Machine Learning Datasets*. Information and Computer Science University of California, 1998. URL: <http://archive.ics.uci.edu/ml/>.
- [3] Quinlan J.R. *Induction of Decision Trees Machine Learning 1*. 1986. Pp. 81–106.

- [4] Quinlan J.R. Improved Use of Continuous Attributes in C 4.5. *Journal of Artificial Intelligence Research*, 1996, vol. 4, pp. 77–90. <http://dx.doi.org/10.1613/jair.279>
- [5] Qin B., Xia Y., Prabhakar S., Tu Y. A rule-based classification algorithm for uncertain data. *The Workshop on Management and Mining Of Uncertain Data (MOUND)*. 2009.
- [6] Tsang S., Kao B., Yip K.Y., Ho W.S., Lee S.D. Decision Trees for Uncertain Data. *IEEE 25th International Conference on Data Engineering (ICDE '09)*. Shanghai, China, 2009. Pp. 441–444. <https://doi.org/10.1109/ICDE.2009.26>
- [7] Xia Y., Xi B. Conceptual clustering categorical data with uncertainty. *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*. Patras, Greece, 2007. Pp. 329–336. <https://doi.org/10.1109/ICTAI.2007.135>

Citation

Astakhova I.I., Vagin V.N., Fomina M.V. The problem of the formation of generalized concepts when working with uncertain attributes. *Nechetkie Sistemy i Myagkie Vychisleniya* [Fuzzy Systems and Soft Computing], 2017, vol. 12(2), pp. 151–158. (in Russian). <https://doi.org/10.26456/fssc30>