Исследование алгоритма построения нечетких деревьев решений на основе интеграции энтропии и теории нечетких множеств

С.Б. Бегенова[1], Т.В. Авдеенко[1]

[1]Новосибирский государственный технический университет, пр. К. Маркса 20, Новосибирск, Россия, 630073

Аннотация. В данной статье рассмотрен алгоритм построения нечетких деревьев решений для нечетких и четких данных, на основе интеграции энтропии и теории нечетких множеств. Алгоритм предполагает нечеткое представление числовых атрибутов и дальнейшее построение дерева. В данной статье представлен метод, который включает в себя особенности двух вышеупомянутых подходов: графическое представление системы правил в виде дерева и нечеткое представление данных. Подход характеризуется такими преимуществами как высокая познаваемость деревьев решений и способность справляться с неточной и неопределенной информацией. Полученный метод обучения подходит для классификации задач, как с численными, так и с символьными атрибутами. В статье будут приведены иллюстрации решений и численные результаты.

1. Введение

В настоящее время, в эпоху массового хранения большого количества данных, извлечение знаний представляет собой узкое место в области инженерии знаний. Компьютерные программы, извлекающие знания из данных, успешно пытаются решить эту проблему. Среди данных программ очень популярны системы, занимающиеся построением деревьев решений для решения задач по принятию решений и классификации. Полученные знания в виде деревьев решений и процедур вывода высоко ценятся за понятность и наглядность представления данных. Подобная оценка, в свое время, вызвала интерес со стороны ученых, что привело к ряду методологических и эмпирических достижений. Однако, изначально деревья принятия решений были популяризированы Куинланом и его алгоритмом ID3.

Одним из расширений классического построения деревьев решений является нечеткий подход. Нечеткое представление становится все более популярным в решении проблем неопределенности, шума и неточных данных. Оно успешно применяется к проблемам во многих промышленных сферах. Большинство исследований по применению данной репрезентативной основы для существующих методологий сосредоточено, в основном, на новых областях, таких как нейронные сети и генетические алгоритмы. На данный момент популярность набирает нечеткий подход, который интегрирует понятия нечетких множеств и энтропии.

В данной статье представлен метод, который включает в себя особенности двух вышеупомянутых подходов: графическое представление системы правил в виде дерева и нечеткое представление данных. В разделе 2 описывается принцип работы деревьев решений, их достоинства и недостатки, алгоритмы их построения. Во разделе 3 показывается принцип построения нечетких деревьев решений, вводятся понятия нечеткой логики, приводится сам с алгоритм построения с соответствующими формулами. В разделе 4 описываются проведенные исследования, строятся выводы и в последнем разделе приводится заключение.

2. Деревья решений

Дерево решений – это общепринятая формализация для отображения переходов значений атрибутов в классы в виде карты, которая состоит из узлов атрибутов или, так называемых, тестов, которые могут иметь два и более поддеревьев, листов или узлов решений, которые помечены классом, указывающим на решение. Главное преимущество такого подхода заключается в визуализации решения. Одним из самых часто используемых алгоритмов построения деревьев решений является ID3 метод, формализованный Куинланом в 1986 [1].

Деревья решений создают эффективные и работоспособные модели для проведения машинного обучения. Приведем следующие характеристики деревьев решений:

- они легко интерпретируемы и наглядны;
- полученная, путем построения дерева решений, модель может быть выражена как графически, так и с помощью текстовых правил;
- конкурентоспособны в сравнении с более дорогостоящими подходами;
- деревья решений масштабируемы;
- они могут обрабатывать дискретные и непрерывные данные;
- деревья решений могут применяться к различным объемам наборов данных, включая множества с большой выборкой;

В процессе построения дерева, образец представлен набором признаков, которые выражены некоторым описательным языком. Образцы, признаки которых известны, называются примерами. Целью построения такого дерева является решение задачи классификации или регрессии.

ID3 и CART являются двумя наиболее важными дискриминационными алгоритмами обучения, работающими с помощью рекурсивного разбиения. Их основные идеи примерно одинаковы: разбиение входящей выборки на подмножества и представление разбиений в качестве дерева. Важным свойством этих алгоритмов является то, что они одновременно пытаются минимизировать размер дерева с оптимизацией некоторой меры качества. Впоследствии они используют один и тот же логический вывод.

3. Нечеткие деревья решений

Введем понятие нечеткого множества.

Пусть имеется некоторое обычное (универсальное, или универсум) множество X элементов x. Нечеткое множество A определяется как упорядоченное множество пар вида $< x, \mu_A(x) >$, где $x \in X$– является элементом некоторого универсального множества X (универсума), $\mu_A(x)$ – функция принадлежности $\mu_A(x): X \rightarrow [0,1]$. При этом $\mu_A(x) = 1$ для некоторого x означает, что элемент x определенно принадлежит нечеткому множеству A, а значение $\mu_A(x) = 0$ означает, что элемент x определенно не принадлежит нечеткому множеству A [2, 3].

Формально конечное нечеткое множество записывается в виде

$A = \{< x_1, \mu_A(x_1) >, < x_2, \mu_A(x_2) >, \dots, < x_n, \mu_A(x_n) >\}$

В терминах деревьев решений, в случае, когда мы хотим классифицировать объект в дереве решений, мы начинаем с корня и спускаемся вниз по дереву, проверяя соответствие атрибута данного объекта текущему узлу. В терминах нечеткой логики, это будет означать принадлежность нечеткого атрибута заданному множеству(узлу). В четких деревьях решений объект может принадлежать только одной ветке, а в нечетком подходе, рассматриваются все ветки, для которых степень принадлежности не будет равна нулю.

Для построения нечеткого дерева решений предлагается следующая процедура [4]:

1. Определение нечеткой базы данных, т.е. нечеткого разбиения для области значений непрерывных атрибутов.
2. Замена непрерывных атрибутов обучающей выборки лингвистическими переменными (термами) нечетких множеств с максимальным совпадением входных значений [5, 6].
3. Расчет энтропии и прирост информации для каждого атрибута, для того, чтобы разделить обучающую выборку на подмножества и определить все узлы дерева до тех пор, пока не будут использованы все атрибуты или до тех пор, пока все объекты обучающей выборки не будут классифицированы.

На рисунке 1 представлен пример фаззификации непрерывных данных

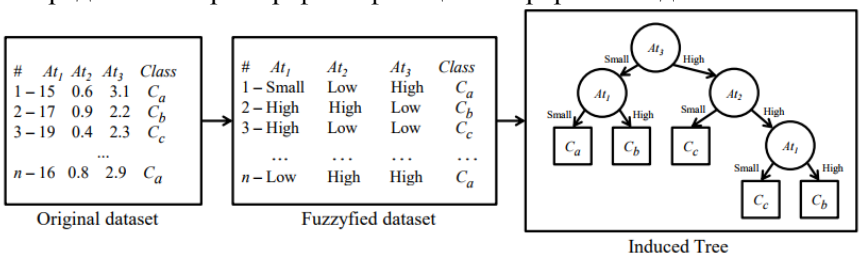

Рисунок 1. Алгоритм построения нечеткого дерева решений.

В первом блоке рис. 1 представлен набор данных с n объектами, 3 атрибутами - At_1, At_2, At_3 и классифицирующим атрибутом. Фаззифицированная версия данного набора данных представлена во втором блоке. Фаззифицированная тестовая выборка используется для построения дерева решений, представленного в последнем блоке рисунка 1.

Формулы энтропии и прироста информации остаются такими же, как для классической версии алгоритма C4.5. Введем следующие обозначения:

$U = \{u_1, u_2, \dots, u_s\}$ – множество выборок данных;

$C = \{c_1, c_2, \dots, c_n\}$– набор атрибутов;

$D = \{d\}$ – одноэлементное множество с атрибутом решения или атрибутом класса. Пусть данный атрибут имеет m различных значений, тогда s_i – количество образцов множества U в классе d_i.

Прирост информации I относительно подмножества S_i равен

$$I(s_{1j}, s_{2j}, \dots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2 p_{ij},$$

$$p_{ij} = \frac{s_{ij}}{|S_j|},$$

где $|S_j|$– количество примеров в подмножестве S.

Энтропия $E(c_i)$ равна

$$E(c_i) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + \dots + s_{mj}}{s} I(s_{1j}, s_{2j}, \dots, s_{mj}),$$

соответственно критерием для выбора атрибута является прирост информации

$$Gain(c_i) = I(s_{1j}, s_{2j}, \dots, s_{mj}) - E(c_i).$$

Разница между обычным алгоритмом C4.5 и фаззифицированной версией алгоритма C4.5 состоит в том, что атрибуты объектов имеют степени принадлежности к тому или иному узлу, и вполне возможна такая ситуация, когда атрибут с определенными вероятностями принадлежит нескольким узлам.

На рисунке 2 представлены два дерева решений, которые были построены с помощью вышеупомянутых алгоритмов.

В качестве пример был взят классический набор данных – ирисы Фишера, которые имеют 4 атрибута: длина и ширина чашелистника, длина и ширина лепестка и три результирующих класса – setosa, versicoloru virginica.

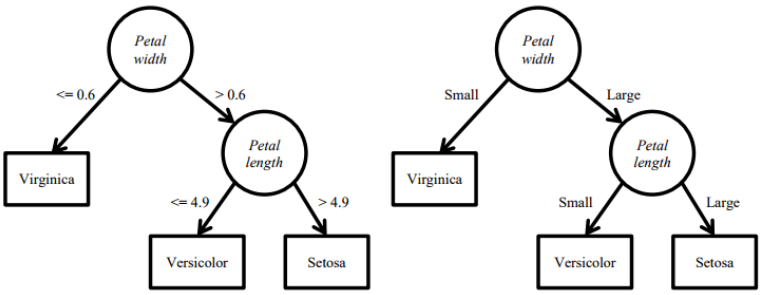

Рисунок 2. Классическое (слева) и нечеткое (слева) деревья решений.

Результатом использования нечеткого дерева будут, например, правила:

1. Если ширина лепестков маленькая, то класс = virginica (с вероятностью 0.75)
2. Если ширина лепестков большая и длина лепестков маленькая, то класс = радужный (с вероятностью 0.34)

4. Исследования

В качестве объекта исследований, так же как и в предыдущем примере, был использован набор данных – ирисы Фишера.

Для построения нечеткого дерева решений на первом этапе необходимо провести процедуру фаззификации.

При проведении процедуры фаззификации множество определения нечетких атрибутов делится на нечеткие подмножества. Значению нечеткого атрибута ставится в соответствие терм, и данное соответствие находится с помощью функции принадлежности. Разбиение множества определения на нечеткие подмножества можно производить равномерно, то есть разбивать множество определения на одинаковые интервалы. Однако, в большинстве реальных наборов данных, полученных из окружающей среды, предпочтительнее проводить разбиение с учетом особенностей исходной выборки. К примеру, может получится так, что большинство объектов выборки лежит в первой трети множества определения и, в таком случае, равномерное разбиение не даст нужного эффекта.

Результаты фаззификации приведены на рисунках 3, 4, 5 и 6.

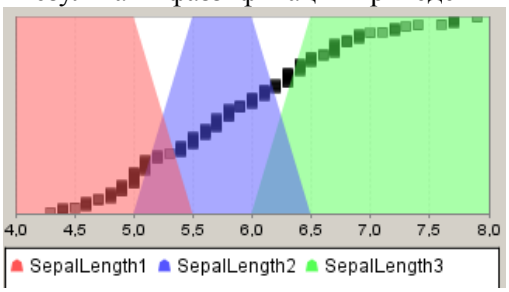

Рисунок 3. Фаззификация атрибута Sepal Length.

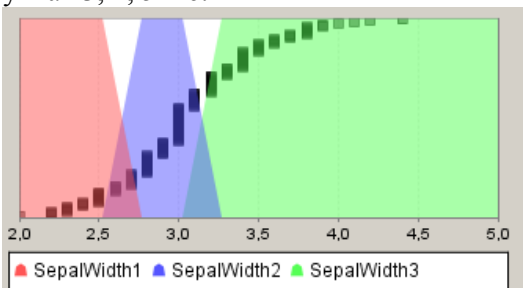

Рисунок 4. Фаззификация атрибута Sepal Width.

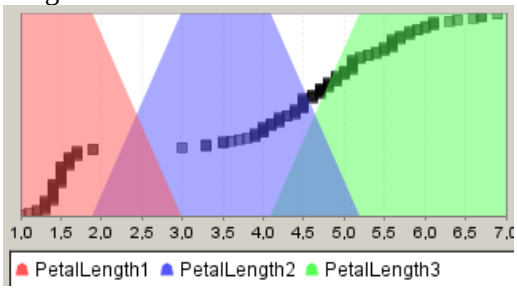

Рисунок 5. Фаззификация атрибута Pepal Length.

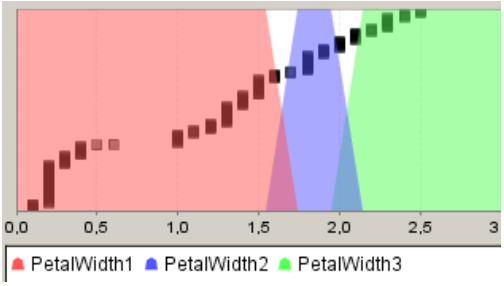

Рисунок 6. Фаззификация атрибута Petal Width.

Приведем результаты классификации, полученные с помощью нечетких деревьев решений.

В контексте проведенных исследований введем следующие обозначения:

Correct- число корректно проклассифицированных объектов выборки;

Incorrect – число некорректно проклассифицированных объектов выборки;

WithoutClass – число объектов без класса.

PercentCorrect – процент корректно проклассифицированных объектов выборки, который вычисляется следующим образом:

$$PercentCorrect = 100 - \frac{(Incorrect + WithoutDecision) * 100}{Correct}$$

Для исследования гипотезы о том, что при уменьшении объема выборки точность классификации нечетких деревьев решений лучше, чем при классификации с помощью классических, была построена зависимость точности классификации от количества экземпляров в наборе данных.

При проведении данного исследования, деревья строились для 3 случайно отобранных N экземпляров выборки и в таблице приведены полученные усредненные значения (сумма значений атрибутов / 3).

Таблица 1. Сравнение результатов классификации, полученных с с использованием нечетких и классических деревьев решений.

Количество экземпляров в наборе данных (N)	Нечеткие деревья решений	Классические деревья решений
120	Correct = 115 Incorrect = 5 WithoutClass = 0 PercentCorrect= 95.65	Correct = 117 Incorrect = 3 WithoutClass = 0 PercentCorrect= 97.43
90	Correct = 88 Incorrect = 2 WithoutClass = 0 PercentCorrect= 97.72	Correct = 85 Incorrect = 5 WithoutClass = 0 PercentCorrect= 94.11
60	Correct = 58 Incorrect = 2 WithoutClass = 0 PercentCorrect= 95.55	Correct = 57 Incorrect = 3 WithoutClass = 0 PercentCorrect= 94.73

По данным, представленным в таблице 1, видно, что при уменьшении выборки с 150 до 90, точность классификации с использованием нечетких деревьев решений на три процента выше чем при классификации с классическими деревьями решений, а при уменьшении выборки до 60 точность выше на 0.82 процента.

В таблице 2 представлена зависимость точности классификации данных от количества термов.

По данным из таблицы видно, что оптимальное количество термов для тестируемого набора данных – 5. Такое количество дало больший процент корректно классифицированных данных в сравнении с 3 термами, а увеличение до 7 термов повышение точности классификации не дало.

Таблица 2. Зависимость точности классификации данных от количества термов.

Количество термов	Результаты классификации
3	Correct = 142 Incorrect = 8 WithoutClass = 0 PercentCorrect= 94.36
5	Correct = 143 Incorrect = 7 WithoutClass = 0 PercentCorrect = 95.33
7	Correct = 143 Incorrect = 7 WithoutClass = 0 PercentCorrect= 95.33

В таблице 3 представлена зависимость точности классификации данных от значения прироста информации. В данном методе, прирост информации будет точкой останова алгоритма, то есть при достижении заданного значения, дальнейшее разбиение дерева прекращается. По данным видно, что чем ниже прирост информации, тем точнее и «глубже» будет построено дерево.

Таблица 3. Зависимость точности классификации данных от прироста информации.

Прирост информации (information gain)	Результаты классификации
0.02	14 листьев Correct = 142 Incorrect = 8 WithoutClass = 0 PercentCorrect= 94.67
0.1	9 листьев Correct = 142 Incorrect = 8 WithoutClass = 0 PercentCorrect= 94.67
0.2	5 листьев Correct = 139 Incorrect = 11 WithoutClass = 0 PercentCorrect= 92.67
0.4	3 листа Correct = 119 Incorrect = 31 WithoutClass = 0 PercentCorrect= 79.33

5. Заключение

Деревья решений успешно применяются для решения задач регрессии и классификации. Они популярны в области машинного обучения, так как деревья решений строят графические модели, наряду с текстовыми правилами, которые легко интерпретируются конечными пользователями. С другой стороны, нечеткие системы, могут решать задачи классификации с входными неточными и зашумленными данными.

Комбинация нечетких деревьев и нечеткой логики позволяет построить наглядные графические модели для качественных и количественных данных [7, 8, 9]. Использование такого типа деревьев решений дает нам несколько вариантов решения с различными вероятностями принадлежности тому или иному классу.

Кроме того, в ходе проведенных исследований было выявлено преимущество классификации с использованием нечетких деревьев решений в отношении классических, путем сравнения процента корректно склассифицированных объектов. Также была выявлена прямая зависимость точности классификации от значения прироста информации (прирост является критерием останова дальнейшего построения дерева).

6. Благодарности

Работа поддержана грантом Министерства образования и науки РФ в рамках проектной части государственного задания, проект № 2.2327.2017/4.6 «Интеграция моделей представления знаний на основе интеллектуального анализа больших данных для поддержки принятия решений в области программной инженерии».

7. Литература

[1] Quinlan, J.R. Induction of decision trees / J.R. Quinlan // Kluwer Academic Publishers, 1986. – Vol.1(1). – P. 81-106.

[2] Авдеенко, Т.В. Метод определения релевантности прецедентов на основе нечетких лингвистических правил / Т.В. Авдеенко, Е.С. Макарова // Научный вестник Новосибирского государственного технического университета. – 2016. – Т. 62, № 1. – С. 17-34.

[3] Avdeenko, T.V. Acquisition of knowledge in the form of fuzzy rules for cases classification / T.V Avdeenko, E.S. Makarova // Lecture Notes in Computer Science. Data Mining and Big Data. – 2017. – Vol. 10387. – P. 536-544.

[4] Cintra, M.E. Fuzzy DT- a fuzzy decision tree algorithm based on C4.5 / M.E. Cintra, M.C. Monard, H.A. Camargo // CBSF – Brazilian Congress on Fuzzy Systems. – 2012. – P. 199-211.

[5] Janikow, C.Z. Fuzzy Decision Trees: Issues and Methods / C.Z. Janikow // IEEE Transactions of Man, Systems, Cybernetics. – 1998. – Vol 28(1). – P. 1-14.

[6] Faifer, M. Bottom-up Partitioning in Fuzzy Decision Trees / M. Faifer, C.Z. Janikow // Proceedings of the 19th International Conference of the North American Fuzzy Information Society. – IEEE, 2000. – P. 326-330.

[7] Tokumaru, M. Kansei Impression analysis using fuzzy c4.5 decision tree / M. Tokumaru, N. Muranaka // Int. con. on Kansei engineering and emotion research. – 2010. – P. 1333-1342.

[8] Janikow, C.Z. Fid 4.1: an overview / C.Z. Janikow // Proc. of the North American Fuzzy Information Processing Society. NAFIPS. – 2004. – P. 877-881.

[9] Cintra, M.E. The use of fuzzy decision trees for coee rust warning in Brazilian crop / M.E. Cintra, C.A.A. Meira, M.C. Monard, H.A. Camargo, L.H. Rodrigues // Int. Conf. Int. Sys. Design & Applications, 2011. – Vol. 1. – P. 1347-1352.

The research of fuzzy decision trees building using entropy and the theory of fuzzy sets

S.B. Begenova[1], T.V. Avdeenko[1]

[1]Novosibirsk State Technical University, K. Marksa str. 20, Novosibirsk, Russia, 630073

Abstract. Decision trees are widely used in the field of machine learning and artificial intelligence. Such popularity is due to the fact that with the help of decision trees graphic models, text rules can be built and they are easily understood by the end user. Because of the inaccuracy of observations, uncertainties, the data, collected in the environment, often takes an unclear form. Therefore, fuzzy decision trees becoming popular in the field of machine learning. This article presents a method that includes the features of the two above-mentioned approaches: a graphical representation of the rules system in the form of a tree and a fuzzy representation of the data. The approach uses such advantages as high comprehensibility of decision trees and the ability to cope with inaccurate and uncertain information in fuzzy representation. The received learning method is suitable for classifying problems with both numerical and symbolic features. In the article solution illustrations and numerical results are given.

Keywords: decision tree, machine learning, artificial intelligence, fuzzy data.