Український журнал дистанційного зондування Землі 22 (2019) 22–26

Ukrainian Journal of Remote Sensing

journal homepage: www.ujrs.org.ua

DOI:https://doi.org.10.36023/ujrs.2019.22.154

УДК 528.8:553.9:004.932

Аппроксимация реальных данных нечеткими множествами для задачи классификации

К. Ю. Суханов *

ГУ "Научный центр аэрокосмических исследований Земли ИГН НАН Украины", ул. О. Гончара 55-Б, Киев, 01054, Украина

В статье рассмотрен метод классификации реальных данных с использованием аппарата нечетких множеств и нечеткой логики как гибкого инструмента обучения и распознавания природных объектов на примере нефтегазоперспективных участков Днепровско-Донецкой впадины. Реальными данными в данном подходе названы значения для функции принадлежности, которые получены не в результате субъективных экспертных оценок, а в результате объективных измерений. Предложено аппроксимировать функциями принадлежности нечетких множеств обучающие данные, чтобы на этапе определения неизвестных объектов использовать результаты аппроксимации, которые были получены на этапе обучения. На первом шаге обучения каждому признаку обучающих данных ставиться в соответствие первичное традиционное одномерное множество, функция принадлежности которого может принимать значения только из бинарного набора — 0, если обучающий объект не принадлежит множеству, и 1, если обучающий объект принадлежит множеству. На втором шаге обучения первичное множество отображается на нечеткое множество, а параметры функции принадлежности этого нечеткого множества определяется в результате аппроксимации этой функцией принадлежности традиционного множества. На третьем шаге совокупность одномерных нечетких множеств, которые соответствуют отдельному признаку объекта, отображается на нечеткое множество, которое соответствует всем признакам объекта из набора обучающих данных. Такое множество представляет собой пересечение нечетких множеств отдельных признаков, к которым на последнем шаге применяют операции размытия и концентрирования из теории нечетких множеств. Таким образом, функция принадлежности к нечеткому множеству класса является операцией выбора минимального значения из функций принадлежностей нечетких множеств отдельных признаков объектов, которые возведены в некоторую степень, которая соответствует операции размытия или концентрирования. Задача отнесения исследуемого объекта к тому или иному классу сводится к сравнению значений функций принадлежности многомерного нечеткого множества и выбора класса, у которого функция принадлежности принимает наибольшее значение. Дополнительно после этапа обучения можно определить степень значимости признака объекта, которая является индексом нечеткости, чтобы исключить из анализа несущественные данные (признаки объекта).

Ключевые слова:нечеткая логика, нечеткие множества, классификация, распознавание образов, аппроксимация

Нечеткие множества (НМ) (Zadeh, 1965) являются одним из математических аппаратов, который используют для задач классификации и обнаружения объектов. В 30-х годах прошлого века Лофти Заде для описания интуитивной логики человека предложил разработанную им теорию нечетких множеств, в которой принадлежность элемента множеству определяется функцией принадлежности (ФП), которая может принимать значения на непрерывном интервале [0; 1]. Это расширило традиционную двоичную логику. Традиционно ФП строят как аппроксимацию мнений экспертов, а сами ФП задаются эвристически. В этой статье предлагается аппроксимировать функциями принадлежности реальные данные, которые были получены в результате наблюдений или экспериментов, чтобы обойти ограничения, которые возникают для неполных данных или выборок, которые имеют низкую статистическую достоверность.

Толчком к этой идее послужил метод нечеткой кластеризации С — средних Dunn, (Dunn, 1973), однако в этом методе единственным параметром ФП, который изменяется, есть смещение многомерной функции (центр кластера) в пространстве признаков. В отличие от метода С средних предлагается искать каждый параметр ФП, а в перспективе выбирать функцию принадлежности из заданного набора, которая бы оптимально аппроксимировала данные из набора функций. Автоматическая настройка параметров ФП, в отличие от метода С — средних, используется в нечетких нейронных системах (Fuller, 1995), но для них характерно использование ФП с жестко заданными формами.

На данном этапе исследований в качестве ФП использовалась функция (3), которая является произведением *S*–функции (1) и *Z*–функции (2).

$$S(x;a,b)=\frac{1}{e^{(a-x)/b}+1}, \quad (1)$$

$$Z(x;c,d)=\frac{1}{e^{(x-c)/d}+1}, \quad (2)$$

$$\mu(x;a,b,c,d)=S(x;a,b)\cdot Z(x;c,d), \quad (3)$$

где μ — функция принадлежности; x — аргумент функции принадлежности; a, b, c, d — параметры функции принадлежности.

* E-mail adress: k.sukhanov@gmx.com (К. Ю. Суханов). ORCID: 0000-0002-3518-3471

Online ISSN 2313-2132

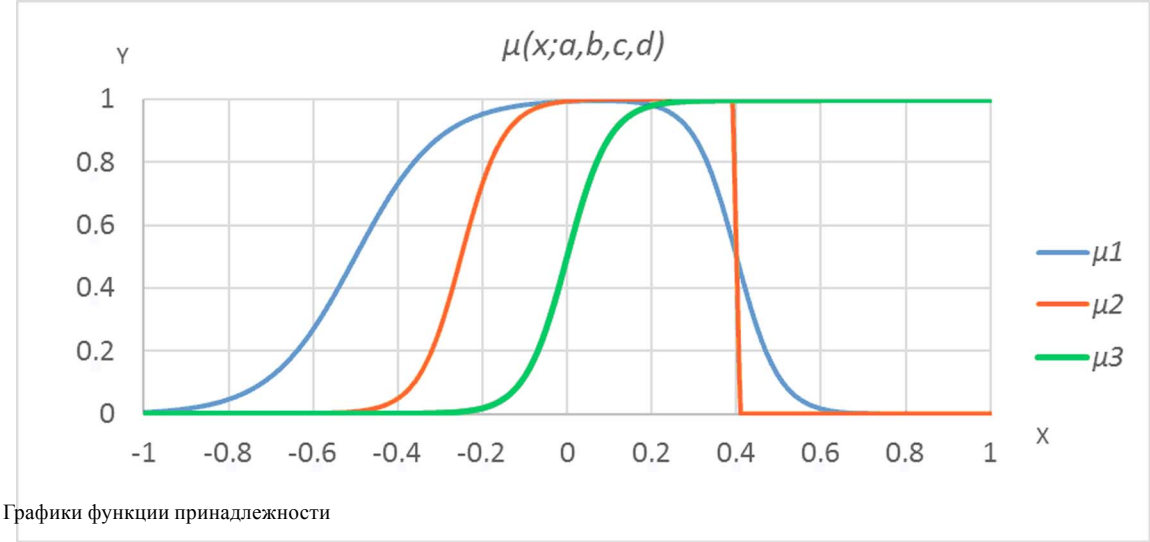

Рис.1. Графики функции принадлежности

μ_1(x)= μ (x;– 0,5; 0,1; – 0,4; 0,05), μ_2(x)= μ (x; – 0,25; 0,05;– 0,4; 0,0001), μ_3(x)= μ (x; 0; 0,05; – 0,8; – 0,0001)

Как видно на рис.1, изменения параметров могут приводить к значительному изменению формы функции: скошенный колокол (μ_1), сигмоидальная функция (μ_3), “обрезанная” сигмоидальная функция (μ_2).

Сама аппроксимация заключается в следующем. В пространстве признаков X множеству объектов поставим в соответствие НМ A такое, что $\left\{ A_i \mid \mu_A(\bar{x}) \right\}$, где μ_A — функция принадлежности A, $\bar{x}$ — вектор признаков объекта. В свою очередь каждому признаку поставим в соответствие НМ Si такое, что $\left\{ S_i \mid \mu_{Si}(x_i) \right\}$, где i — номер признака. Тогда определим A как пересечение множеств признаков $A = \bigcap S_i$. Тогда функция принадлежности к нечеткому множеству A будет (4)

$$\mu(\bar{x}) = \min_i \left\{ \mu_{Si}(\bar{x}_i)^{p_i} \right\}, \qquad (4)$$

где p_i — степень концентрации-размытости множества.

Для каждого признака обучающих данных необходимо построить табличную функцию $L_i(x_i)$, аргументами которой будут значения признака, а значениями самой функции присвоить традиционные логические 0 и 1. 1 — если значение соответствует объекту и 0 — если не соответствует. А общей табличной функцией от вектора признаков будет $L_A(\bar{x})$. Таким образом задача аппроксимации отдельного признака сводится к аппроксимации функции $L_i(x_i)$ функцией $\mu_{Si}(x_i;a,b,c,d)$, используя метод наименьших квадратов для определения параметров a,b,c,d. Следующий этап аппроксимации заключается в определении степеней p_i, при которых среднеквадратичное отклонение ФП μ_A от L_A будет наименьшей. Очевидно, что значения L_A равны значениям L_i, поэтому в дальнейшем их обозначим как L.

При обучении количество объектов, которые принадлежат множеству, может не совпадать с числом объектов, которые не принадлежат множеству. Чтобы уменьшить влияние случайных отклонений при малых объемах данных критерий аппроксимации (5) представим в виде среднего от двух среднеквадратичных ошибок — отклонений ФП для значений принадлежности множеству (истинных значений, равных 1) и отклонений ФП для значений не принадлежности множеству (ложных значений, равных 0).

$$\sigma_A^2 = \frac{1}{2N_T}\sum_k^{N_T}(1-\mu_k)^2 + \frac{1}{2N_F}\sum_j^{N_F}\mu_j^2 \qquad (5)$$

где σ_A^2 — среднеквадратичная ошибка; N_T — число объектов, которые принадлежат четкому множеству; N_F — число объектов, которые не принадлежат четкому множеству; k — номер объекта, который принадлежит множеству; j — номер объекта, который не принадлежит множеству.

Для ФП признаков тоже не делается никаких допущений о распределении, поэтому одинаковые значения для признака объединяются в одну точку. Следует отметить, что при достаточном объеме данных, когда можно делать заключение о статистической достоверности, объединять точки не следует. Тогда значения функции принадлежности будут аппроксимировать частоту попадания объекта в область четкого множества. Отметим, что условием решения задачи аппроксимации есть поиск минимума значения критерия (5).

Была исследована аппроксимация значений троичной логики, при которой значение признака одновременно присутствовало для объектов принадлежащим и не принадлежащим множеству. В этом случае функции L_i присваивалось значение 0.5 (не определено). Были получены результаты такие же, как и для двоичной логики, но от идеи использования троичной логики пришлось отказаться из-за сложности, которая возникла в связи с формальной оценкой индекса нечеткости. По определению индекс нечеткости есть расстоянием между нечетким и четким множествами, принадлежность к которым задана значениями двоичной логики.

Ниже приведены результаты исследований нефтегазоперспективных участков (объектов) в Днепровско-Донецкой впадине.

Рассмотрим промежуточные результаты вычислений из таблицы 1, в которой представлены значения ФП признаков (μ_A) и ФП вектора признаков (μ_i) до второй аппроксимации операциями концентрации-размытости. Как видно из таблицы значения (μ_A) для обучающих объектов, которые принадлежат множеству, результат неудовлетворительный, т. к. большинство этих значений не превышает 0.5 (неопределенность). Но уже по некоторым

Таблица 1.
Аппроксимация ФП каждого признака

N_{pt}	L	μ_A	μ_i								
			$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$
14	1	0.50	0.79	0.64	0.50	0.59	0.50	0.50	0.50	1.00	0.50
15	1	0.50	0.72	0.64	0.50	0.62	0.50	0.50	0.50	1.00	1.00
16	1	0.50	0.65	0.61	0.50	0.65	0.50	0.50	0.50	1.00	1.00
17	1	0.50	0.57	0.60	0.50	0.66	0.50	0.50	0.50	1.00	1.00
18	1	0.50	0.57	0.56	0.50	0.59	0.50	0.50	0.50	1.00	1.00
21	1	0.46	0.65	0.51	0.50	0.46	0.50	0.50	0.50	0.50	0.50
29	1	0.38	0.48	0.42	0.50	0.38	0.50	0.50	0.50	0.50	0.50
5	0	0.00	0.00	0.00	0.50	0.57	0.50	0.50	0.50	0.00	0.50
6	0	0.00	0.72	0.62	0.50	0.59	0.50	0.50	0.50	0.00	0.50
8	0	0.00	0.48	0.59	0.50	0.62	0.50	0.50	0.50	0.00	0.50
9	0	0.50	0.57	0.57	0.50	0.62	0.50	0.50	0.50	0.50	0.50
10	0	0.43	0.57	0.46	0.50	0.43	0.50	0.50	0.50	0.50	0.50
28	0	0.38	0.48	0.42	0.50	0.38	0.50	0.50	0.50	0.50	0.50

где N_{pt} — номер точки

значениям (μ_i) , которые равны или незначительно отклоняются от 0.5, видно, что эти признаки i являются неинформативными.

После аппроксимации, применяя операции концентрирования-размытия, были получены результаты, представленные в таблице 2 и таблице 3.

Таблица 2.
Аппроксимация ФП вектора признаков

N_{pt}	L	μ_A	μ_i								
			$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$
14	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
17	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
18	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
21	1	0.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	1.00
29	1	0.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	1.00
5	0	0.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.00	1.00
6	0	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00
8	0	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00
9	0	0.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	1.00
10	0	0.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	1.00
28	0	0.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	1.00

Рассмотрим таблицу 3 результатов аппроксимации. Абсолютные значения функции $log_2(p_i)$ соответствуют числу операций концентрирования или размытия множества. Отрицательные значения соответствуют размыванию множества, а положительные — концентрированию. Эти значения можно рассматривать в качестве показателя информативности или неинформативности признака. Значения операций размывания множества соответствуют степени неинформативности признака, что может служить основанием для его игнорирования при исследованиях после процедуры обучения. В то же время, значения концентрации показывает хорошую степень информативности признака.

Кроме степени концентрации-размытости множества об информативности признака можно судить по индексу нечеткости (ИН) I_f, который в нашем случае соответствует среднеквадратичной ошибке аппроксимации ФП. Из таблицы 4 наименьший ИН у 8-го признака, а максимально возможные значения у 3, 5, 6 и 7.

Для проверки метода был использован тест, в котором из обучающих данных исключается одна точка, а затем проходит обучение по оставшейся совокупности. После обучения,

Таблица 3.
Параметры ФП признаков

i	a	b	c	d	p_i	$Log_2(p_i)$
1	–5.812	2.921	–1.152	0.003358	6.104×10^{-5}	–14
2	–8.353	11.11	–1.819	0.004268	6.104×10^{-5}	–14
3	28.85	19.34	58.59	14.62	6.104×10^{-5}	–14
4	238	–17.72	–222	8.57	6.104×10^{-5}	–14
5	90	1.013	100	1.024	6.104×10^{-5}	–14
6	36.6	169.5	274	109.7	3.052×10^{-5}	–15
7	80	1	100	1	6.104×10^{-5}	–14
8	4	0.07071	19.87	0.4573	2	1
9	0.9996	0.4654	–14.2	–1.974	6.104×10^{-5}	–14

Таблица 4.
Индексы нечеткости для признаков

i	1	2	3	4	5	6	7	8	9
I_f	0.44	0.47	0.5	0.49	0.5	0.5	0.5	0.27	0.41

полученные значения параметров аппроксимации проверяются на исключенной точке, которая должна принадлежать или не принадлежать множеству объектов. Эта операция была последовательна применена для каждой точки обучающих данных, а результаты теста представлены в таблице 5.

В ней для сравнения представлены значения ФП из результатов обучения μ_A и μ_i для 8-го признака (μ_8), взятые из таблицы 2, и аналогичные значения из исключающего теста μ_A (иск) и μ_8(иск). Значения 0.50^*, которые помечены звездочкой, являются значениями неконцентрированного множества, т.е. после концентрации они должны быть равны $0.50^2 = 0.25$. Зафиксировано только одно несовпадение для всех точек (7.69%), поэтому можно говорить о положительном результате.

Таблица 5.
Исключающий тест

N_{pt}	L	μ_A (иск)	μ_A	μ_8 (иск)	μ_8	N_{pt}	L	μ_A (иск)	μ_A	μ_8 (иск)	μ_8
14	1	0.00	1.00	1.00	1.00	5	0	0.00	0.00	0.00	0.00
15	1	1.00	1.00	1.00	1.00	6	0	0.00	0.00	0.00	0.00
16	1	1.00	1.00	1.00	1.00	8	0	0.00	0.00	0.00	0.00
17	1	1.00	1.00	1.00	1.00	9	0	0.50^*	0.25	0.50^*	0.25
18	1	1.00	1.00	1.00	1.00	10	0	0.50^*	0.25	0.50^*	0.25
21	1	0.00	0.25	0.00	0.25	28	0	0.50^*	0.25	0.50^*	0.25
29	1	0.00	0.25	0.50^*	0.25						

Литература

Zadeh L. A. Fuzzy sets. *Information and Control*. 1965. Vol. 8. No. 3. P. 338–353.

Dunn, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*. 1973. Vol. 3. No. 3. P. 32–57.

Robert Fuller. Neural Fuzzy Systems. Åbo Akademis tryckeri, Åbo,ESF ASeries A:443 [ISBN 951-650-624-0, ISSN 0358-5654]. 1995.

References

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8 (3), 338–353.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3 (3), 32–57.

Robert Fuller. (1995). Neural Fuzzy Systems.Åbo Akademis tryckeri, Åbo, ESF A Series A:443 [ISBN 951-650-624-0, ISSN 0358-5654].

АППРОКСИМАЦИЯ РЕАЛЬНЫХ ДАННЫХ НЕЧЕТКИМИ МНОЖЕСТВАМИ ДЛЯ ЗАДАЧИ КЛАССИФИКАЦИИ

К. Ю. Суханов

ДУ “Науковий центр аерокосмічних досліджень Землі ІГН НАН України”. E-mail: k.sukhanov@gmx.com, вул. О. Гончара 55-Б, Київ, 01054, Україна. ORCID: 0000-0002-3518-3471

У статті розглянуто метод класифікації реальних даних з використанням апарату нечітких множин та нечіткої логіки як гнучкого інструменту навчання і розпізнання природних об’єктів на прикладі нафтогазоперспективних ділянок Дніпровсько-Донецької западини. Реальними даними в даному підході названі значення для функції приналежності, які отримані не в результаті суб’єктивних експертних оцінок, а в результаті об’єктивних вимірів. Запропоновано апроксимувати функціями належності нечітких множин навчальні дані, щоби на етапі визначення невідомих об’єктів використовувати результати апроксимації, які було отримано на етапі навчання. На першому кроці навчання кожній ознаці навчальних даних ставитися у відповідність первинна традиційна одномірна множина, функція належності якої може приймати значення тільки з бінарного набору — 0, якщо навчальний об’єкт не належить множині, і 1, якщо навчальний об’єкт належить множині. На

другому кроці навчання первинна множина відображається на нечітку множину, а параметри функції приналежності цієї нечіткої множини визначаються в результаті апроксимації цією функцією приналежності традиційної множини. На третьому кроці сукупність одновимірних нечітких множин, які відповідають окремій ознаці об'єкта, відображається на нечітку множину, яка відповідає всім ознакам об'єкта з набору навчальних даних. Така множина є перетином нечітких множин окремих ознак, до яких на останньому кроці застосовують операції розмиття і концентрування з теорії нечітких множин. Таким чином, функція належності до нечіткій множині класу є операцією вибору мінімального значення з функцій належності нечітких множин окремих ознак об'єктів, які зведені в певну ступінь, яка відповідає операції розмиття або концентрування. Завдання віднесення досліджуваного об'єкта до того чи іншого класу зводиться до порівняння значень функцій належності багатовимірного нечіткої множини і вибору класу, у якого функція належності приймає найбільше значення. Додатково після етапу навчання можна визначити ступінь значущості ознаки об'єкта, яка є індексом нечіткості, щоби вилучити з аналізу несуттєві дані (ознаки об'єкта).
Ключові слова: нечіткая логика, нечіткі множини, класифікація, розпізнавання образів, апроксимація

APPROXIMATION OF REAL DATA BY FUZZY SETS FOR THE CLASSIFICATION PROBLEM

K. Sukhanov

Scientific Centre for Aerospace Research of the Earth, National Academy of Sciences of Ukraine, E-mail: k.sukhanov@gmx.com, O. Gonchar st. 55-B, 01054 Kyiv, Ukraine. ORCID: 0000-0002-3518-3471

The article deals with the method of classification of real data using the apparatus of fuzzy sets and fuzzy logic as a flexible tool for learning and recognition of natural objects on the example of oil and gas prospecting sections of the Dnieper-Donetsk basin. The real data in this approach are the values for the membership function that are obtained not through subjective expert judgment but from objective measurements. It is suggested to approximate the fuzzy set membership functions by using training data to use the approximation results obtained during the learning phase at the stage of identifying unknown objects. In the first step of learning, each traditional future of a learning data is matched by a primary traditional one-dimensional set whose membership function can only take values from a binary set — 0 if the learning object does not belong to the set, and 1 if the learning object belongs to the set. In the second step, the primary set is mapped to a fuzzy set, and the parameters of the membership function of this fuzzy set are determined by approximating this function of the traditional set membership. In the third step, the set of one-dimensional fuzzy sets that correspond to a single feature of the object is mapped to a fuzzy set that corresponds to all the features of the object in the training data set. Such a set is the intersection of fuzzy sets of individual features, to which the blurring and concentration operations of fuzzy set theory are applied in the last step. Thus, the function of belonging to a fuzzy set of a class is the operation of choosing a minimum value from the functions of fuzzy sets of individual features of objects, which are reduced to a certain degree corresponding to the operation of blurring or concentration. The task of assigning the object under study to a particular class is to compare the values of the membership functions of a multidimensional fuzzy set and to select the class in which the membership function takes the highest value. Additionally, after the training stage, it is possible to determine the degree of significance of an object future, which is an indistinctness index, to remove non-essential data (object futures) from the analysis.
Keywords: fuzzy sets, fuzzy logic, classification, pattern recognation, approximation

Стаття надійшла до редакції 12.09.2019