УДК 621.396.41

И.А. Ходашинский, М.Б. Бардамова, В.С. Ковалев

Построение нечеткого классификатора алгоритмом гравитационного поиска

Исследуется эффективность применения дискретного и непрерывного алгоритма гравитационного поиска для построения нечеткого классификатора. Приводится описание и псевдокод алгоритма гравитационного поиска. Результаты тестирования сравниваются с результатами работы аналогичных алгоритмов с помощью статистического критерия Манна–Уитни–Уилкоксона.

Ключевые слова: нечеткий классификатор, алгоритм оптимизации, гравитационный алгоритм, отбор признаков.

doi: 10.21293/1818-0442-2017-20-2-84-87

Задача классификации данных решается различными методами машинного обучения: нейронными сетями, дискриминантным анализом, деревьями решений, методом ближайших соседей. Среди всех методов способ классификации, основанный на использовании нечеткой логики, отличается наиболее легкой возможностью интерпретации благодаря использованию базы правил и лингвистических термов [1] и отсутствием допущений, необходимых для статистической классификации [2].

Классификация состоит в разделении объектов в пространстве признаков по классам или категориям на основе ретроспективных наблюдений, для которых известны значения меток классов. Цель классификации состоит в том, чтобы сопоставить входному вектору одну из меток классов. Классификатор, реализующий отображение векторов признаков в метки класса, представляет собой алгоритм классификации данных. Построение нечетких классификаторов предполагает решение следующих основных задач: отбор признаков, генерация базы нечетких правил, оптимизация параметров антецедентов (ЕСЛИ-частей) правил [3].

Так как классификация является задачей обучения с учителем, важно подготовить обучающие данные таким образом, чтобы они позволяли алгоритму наиболее точно и быстро построить обучающую модель. Для этого необходимо выделить комбинацию информативных признаков и исключить из таблицы наблюдения признаки, уменьшающие точность и замедляющие процесс обучения. Принято выделять три процедуры генерации подмножества признаков: полный перебор, эвристики и случайный выбор. Известны две схемы отбора признаков: «фильтры» и «обертки», различающиеся участием («обертки») или неучастием («фильтры») алгоритма классификации в отборе признаков [3, 4].

В данной работе исследуется эффективность применения дискретного алгоритма гравитационного поиска для отбора признаков, а также непрерывного гравитационного алгоритма для оптимизации параметров нечеткой системы.

Нечеткий классификатор

Классификатор может быть описан следующей функцией:

$$c: \mathbf{X} \rightarrow \mathbf{C}, \quad (1)$$

где $\mathbf{X} = \{x_1, \dots, x_{|\mathbf{X}|}\}$ – множество входных признаков; $\mathbf{A} = \{a_1, \dots, a_n\}$ – непустое конечное множество признаков, таких, что $a: \mathbf{X} \rightarrow V_a$ для любого $a \in \mathbf{A}$; V_a – множество допустимых значений признака a ; $\mathbf{C} = \{c_1, c_2, \dots, c_m\}$ – множество меток классов [3]. Значение $c_i = c(\mathbf{a}, \boldsymbol{\theta}, \mathbf{S})$ является меткой для вектора признаков $\mathbf{a}$; $\boldsymbol{\theta}$ – вектор параметров классификатора, $\mathbf{S}$ – вектор участвующих в классификации признаков.

Основой для построения нечеткого классификатора с возможностью отбора информативных признаков является нечеткое правило, которое в общем виде можно представить следующим образом:

ЕСЛИ $s_1 \wedge a_1 = A_{1i}$ И ... И $s_n \wedge a_n = A_{ni}$ ТО $\text{class}_i = c_j$, где A_{li} – нечеткий терм, характеризующий l -й признак в i -м правиле ($i \in [1, R]$); R – число правил; запись $s_j \wedge a_j$ указывает на наличие ($s_j = 1$) или отсутствие ($s_j = 0$) признака в классификаторе. Вектор $\mathbf{S} = (s_1, s_2, \dots, s_n)$ формируется алгоритмом отбора признаков.

Для определения метки класса был использован принцип «команда победителей получает всё», где «команда» – это подмножество правил с одинаковыми метками класса в консеквентной части. Для поступившего на вход классификатора вектора $\mathbf{a}$ вычисляется степень его соответствия «команде» правил:

$$\beta_j(\mathbf{a}) = \sum_{R_{ij}} \prod_{k=1}^n A_{ki}(a_k), j=1, 2, \dots, m. \quad (2)$$

Выходом классификатора является метка класса, определяемая по максимальному суммарному значению, которое набирает «команда»:

$$\text{class} = c_{j^*}, j^* = \arg \max_{1 \leq j \leq m} \beta_j. \quad (3)$$

На множестве обучающих данных Tr определим единичную функцию.

$$\text{delta}(p, \boldsymbol{\theta}, \mathbf{S}) = \begin{cases} 1, & \text{если } c_p = c(\mathbf{x}_p, \boldsymbol{\theta}, \mathbf{S}) \\ 0, & \text{иначе} \end{cases}, p=1, \dots, |Tr|, \quad (4)$$

тогда целевая функция, являющаяся мерой точности классификации, может быть выражена следующим образом:

$$E(\theta, S) = \frac{\sum_{p=1}^{|Tr|} \text{delta}(p, \theta, S)}{|Tr|}. \quad (5)$$

Проблема построения нечеткого классификатора сводится к поиску максимума целевой функции. Для оптимизации вектора антецедентов, состоящего из параметров θ , предлагается использовать непрерывный алгоритм гравитационного поиска (GSA_R), а отбор признаков осуществлять дискретным алгоритмом гравитационного поиска (GSA_D) по схеме обертки [5].

Алгоритм гравитационного поиска

Алгоритм гравитационного поиска является метаэвристическим алгоритмом, основанным на фундаментальных законах тяготения. Популяция, состоящая из входных векторов, представляет собой систему частиц, между которыми действуют силы притяжения [4]. На вход непрерывному алгоритму подается вектор параметров нечеткой системы θ , на основе которого генерируется популяция частиц, и следующие параметры: количество частиц N , количество итераций T , начальное значение гравитационной постоянной G_0 , коэффициент точности поиска α , малая константа ε . Значение гравитационной постоянной рассчитывается на каждой итерации на основе монотонно убывающей функции.

Для каждой i -й частицы из популяции θ на t -й итерации рассчитываются следующие физические характеристики: $m_i(t)$ – масса; $a_i(t)$ – ускорение; $V_i(t)$ – скорость. На последнем шаге происходит обновление позиции частицы путем изменения текущих координат на величину, пропорциональную скорости. Расчеты итерационно повторяются до истечения T , затем на выход подается вектор с наименьшим значением ошибки θ_{best} . Ниже приведен псевдокод алгоритма.

Вход: $T, N, G_0, \alpha, \varepsilon$.

Выход: значение θ_{best} .

Population = $\{\theta_1, \theta_2, \dots, \theta_N\}$;

$t := 1$;

цикл пока ($t \neq T$)

$G[t] := G_0 * (t/T)^{\beta}$;

цикл по i от 0 до N

$m[i] := (E[\theta_i, S] - E[\theta_{worst}, S]) / (E[\theta_{best}, S] - E[\theta_{worst}, S])$;

цикл по j от 0 до S

$R[i, j] := \|\theta_j - \theta_i\|$;

цикл по d от 1 до $|\theta_i|$

$a_i^d := a_i^d + \text{rand} * G[t] * M[j] * (\theta_j^d - \theta_i^d) / (R[i, j] + \varepsilon)$;

$V_i^d[t+1] := \text{rand} * V_i^d[t] + a_i^d[t]$;

$\theta_i^d[t+1] := \theta_i^d[t] + V_i^d[t+1]$;

конец цикла по d

конец цикла по j

конец цикла по i

$t := t + 1$;

конец цикла по t

вывод $\theta_{best} := \text{Search_best}(\text{Population})$.

Дискретный гравитационный алгоритм аналогичен непрерывному, но принимает на вход бинарный вектор S и на каждой итерации после вычислений преобразует полученные количественные значения в бинарные с помощью функции трансформации [6].

Этапы построения нечеткого классификатора включают применение алгоритмов отбора информативных признаков, использование критерия выбора оптимального по точности и сложности набора признаков, формирования структуры и оптимизации параметров классификатора.

На первом этапе бинарным и непрерывным алгоритмом гравитационного формируются группы информативных признаков. Поскольку присутствующая в гравитационных алгоритмах случайность способна привести к отбору не самых информативных признаков, формируется несколько наборов признаков.

На следующем этапе производится построение нечетких классификаторов на каждой отобранной группе признаков. Нечеткие правила и термы генерируются с помощью оригинального алгоритма формирования структур нечеткого классификатора на основе экстремальных значений таблицы наблюдений [7].

Из построенных классификаторов производится выбор оптимального набора признаков относительно точности и количества используемых признаков с помощью статистического информационного критерия Акаике [8]. Алгебраически критерий выражается через сумму меры ошибки модели и штрафа за число параметров модели. Информационный характер критериев связан с концепцией информации Кульбака–Лейблера. Вычисление критерия производится по формуле, адаптированной для нахождения компромисса между точностью и сложностью (количеством используемых признаков), приведенной ниже:

$$AIC = \ln ER_{Tr} + \frac{2}{|Tr|} (1 + cF_S), \quad (6)$$

где ER_{Tr} – ошибка классификатора на обучающей выборке; F_S – количество признаков, используемых классификатором; c – эмпирический коэффициент штрафа, если необходимо, чтобы критерий учитывал в большей степени точность, нежели сложность, то коэффициент c должен принимать значение меньше 1; если же наоборот, то значение c должно быть больше 1; если предпочтений нет или они равнозначны, то значение коэффициента равно 1. Оптимальный (эффективный) классификатор обладает наименьшим значением AIC .

Данные для тестирования

Эксперимент был проведен по схеме кросс-валидации с соотношением 90% – обучающая выборка, 10% – тестовая на наборах данных KEEL [8]; описание наборов приведено в табл. 1.

Результаты тестирования

В табл. 2 приведены усредненные значения процента правильной классификации на выбранных

наборах данных, полученные с помощью нечетких классификаторов, структура которых сформирована алгоритмами GSA_R и $GSA_R + GSA_D$. Для сравнения результатов в табл. 2 также указаны результаты работы алгоритмов классификации данных, представленных в [9]. Алгоритм «Ant-Miner» (Ant Colony-based Data Miner) основан на системе муравьиной колонии. В методике коэволюционной классификации CORE – «CO-Evolutionary Rule Extractor» – применяются эволюционные алгоритмы для генерации правил классификации в двух взаимодействующих и сотрудничающих популяциях. Алгоритм HIDER – «Hierarchical DEcision Rules» – также основан на применении эволюционных алгоритмов для формирования иерархических решающих правил классификации. Генетический алгоритм извлечения нечетких правил классификации из данных SGERD – «Steady-State Genetic Algorithm for Extracting Fuzzy Classification Rules From Data» –

имеет элитарную селекцию и специфические операторы скрещивания и мутации. В методе TARGET – «Tree Analysis with Randomly Generated and Evolved Trees» – генетический алгоритм применяется для поиска в пространстве деревьев решений, оценка каждого дерева основана на байесовском информационном критерии.

Таблица 1

Описание используемых наборов данных

Наименование	Количество		
	записей	признаков	классов
Wine	178	13	3
Glass	214	9	7
Newthyroid	215	5	3
Cleveland	297	13	5
Wisconsin	699	9	2
Pima	768	8	2
Vehicle	846	18	4

Таблица 2

Результаты тестирования нечетких классификаторов, настроенных алгоритмом гравитационного поиска, по сравнению с аналогичными алгоритмами

		Wine	Glass	Newthyroid	Cleveland	Wisconsin	Pima	Vehicle
Ant Miner	Обуч.	99,7	81,5	99,2	60,3	92,6	71,9	59,5
	Тест.	92,1	53,7	90,8	57,5	90,4	66,3	53,1
Core	Обуч.	99,1	54,3	92,7	56,3	94,7	72,7	36,5
	Тест.	94,9	45,7	90,8	53,6	92,4	73,1	36,4
Hider	Обуч.	97,2	90,1	96,0	82,0	97,3	77,8	84,2
	Тест.	82,6	64,4	90,3	55,9	96,1	73,2	63,1
Sgerd	Обуч.	91,8	53,8	90,2	46,6	93,0	73,7	51,5
	Тест.	87,1	48,3	88,4	44,2	92,7	73,7	51,2
Target	Обуч.	85,2	45,1	88,1	55,8	96,1	73,4	51,6
	Тест.	82,2	44,1	86,8	53,0	95,8	73,0	49,8
GSA_R	Обуч.	99,3	63,4	98,3	63,4	96,0	76,9	53,4
	Тест.	97,1	57,5	98,1	62,6	96,3	74,0	50,0
$GSA_D + GSA_R$	Обуч.	99,7	66,0	98,2	58,1	97,3	75,4	50,4
	Тест.	99,4	70,0	99,0	58,3	97,2	77,9	51,1
Количество оставшихся признаков		7 из 13	4 из 9	3 из 5	2 из 13	4 из 9	2 из 8	7 из 18

Об эффективности классификатора, настроенного алгоритмом гравитационного поиска, можно судить по количеству отброшенных признаков и точности классификации на обучающей и тестовой выборке.

Для сравнения эффективности построенных нечетких классификаторов с аналогами использовался статистический критерий Манна–Уитни–Уилкоксона, позволяющий подтвердить или опровергнуть наличие сдвига в медианах случайных величин (уровень значимости считается равным 0,05). На первом этапе сравнения данный критерий использовался для определения наличия или отсутствия превосходства точности классификаторов на тестовых данных, полученных алгоритмами оптимизации. В табл. 3 приведены значения вероятностей, полученные с помощью критерия в процессе сравнения точности классификаторов на выбранных наборах данных.

Исходя из вероятностей по критерию Манна–Уитни–Уилкоксона, можно сделать вывод о статистической неразличимости точности классификато-

ров, настроенных непрерывным алгоритмом гравитационного поиска и приведенными аналогами.

Таблица 3

Значения вероятностей по критерию Манна–Уитни–Уилкоксона

	AM	Core	Hider	Sgerd	Target
GSA_R	0,48	0,31	0,8	0,31	0,31
$GSA_R + GSA_D$	0,37	0,31	0,61	0,31	0,31

На втором этапе критерий Манна–Уитни–Уилкоксона использовался для проверки эффективности дискретного алгоритма гравитационного поиска при решении задачи отбора признаков. В качестве первой выборки значений было выбрано исходное количество признаков, на которых получены значения точности алгоритмов из [9]. Вторая выборка состоит из количества признаков, отобранных дискретным алгоритмом гравитационного поиска, на которых затем были построены нечеткие классификаторы. Выборки и значение вероятности по критерию Манна–Уитни–Уилкоксона приведены в табл. 4.

Таблица 4
Количество признаков для сравнения
и результирующее значение вероятности по критерию
Манна–Уитни–Уилкоксона

	wn	gl	nth	cld	wsn	pm	vhl	p
Всего	13	9	5	13	9	8	18	0,0047
GSA _D	7	4	3	2	4	2	7	

Основываясь на данных статистических расчетах, можно заключить, что алгоритм гравитационного поиска является эффективным, поскольку достигает значений точности, сопоставимых с результатами аналогичных алгоритмов при значительно меньшем количестве признаков.

Заключение

В работе была исследована эффективность применения алгоритма гравитационного поиска для решения задачи классификации данных. На пяти из семи наборов данных непрерывный гравитационный алгоритм показывает лучшие результаты после усе­чения признаков дискретным алгоритмом. Эффективность классификаторов, настроенных алгоритмом гравитационного поиска, подтверждена с помощью статистического критерия Манна–Уитни–Уилкоксона. Алгоритм способен значительно сокращать размерность пространства признаков и демонстрирует хорошие способности к обучению и прогнозированию.

Работа выполнена в рамках базовой части государственного задания Министерства образования и науки Российской Федерации на 2017–2019 гг. Номер 8.9628.2017/БЧ.

Литература

1. Ходашинский И.А. Методы построения трехкритериальных парето-оптимальных нечетких классификаторов / И.А. Ходашинский, И.В. Горбунов // Искусственный интеллект и принятие решений. – 2015. – № 2. – С. 75–87.
2. Scherer R. Multiple Fuzzy Classification Systems // Studies in Fuzziness and Soft Computing. – Vol. 288. – Berlin: Springer-Verlag, 2012. – 132 p.
3. Ходашинский И.А. Построение нечеткого классификатора на основе методов гармонического поиска / И.А. Ходашинский, М.А. Мех // Программирование. – 2017. – № 1. – С. 54–65.
4. Dash M. Feature Selection for Classification / M. Dash, H. Liu // Intelligent Data Analysis. – 1997. – Vol. 1. – P. 131–156.
5. Rashedi E. GSA: A Gravitational Search Algorithm / E. Rashedi, H. Nezamabadi-pour, S. Saryazdi // Information Sciences. – 2009. – Vol. 179. – P. 2232–2248.

6. Mirjalili S. S-shaped versus V-shaped transfer functions for binary Particle Swarm Optimization / S. Mirjalili, A. Lewis // Swarm Evolut. Comput. – 2013. – Vol. 9. – P. 1–14.

7. Алгоритмы структурной идентификации компактных и точных нечетких систем / И.А. Ходашинский, И.В. Горбунов, К.С. Сарин, С.Р. Субханкулова // Информационные и математические технологии в науке и управлении. – 2016. – № 1. – С. 82–93.

8. Yen J. Application of Statistical Information Criteria for Optimal Fuzzy Model Construction / J. Yen, L. Wang // IEEE Trans. Fuzzy Systems. – 1998. – Vol. 6, No. 3. – P. 362–372.

9. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework / J. Alcalá-Fdez, A. Fernandez, J. Luengo et al. // Journal of Multiple-Valued Logic and Soft Computing. – 2011. – Vol. 17. – P. 255–287.

Ходашинский Илья Александрович

Д-р техн. наук, профессор каф. комплексной информационной безопасности электронно-вычислительных систем (КИБЭВС) ТУСУРа
Тел.: +8 (382-2) 41-34-26
Эл. почта: hodashn@rambler.ru

Бардамова Марина Борисовна

Студентка каф. КИБЭВС ТУСУРа
Тел.: +7-903-951-61-13
Эл. почта: 722bmb@gmail.com

Ковалев Виталий Сергеевич

Студент каф. КИБЭВС ТУСУРа
Тел.: +7-952-888-80-53
Эл. почта: Vitaly_979@mail.ru

Hodashinsky I.A., Bardamova M.B., Kovalev V.S.

Identification of fuzzy classifiers based on gravitational search algorithm

In this paper the authors investigate how gravitational search algorithm can be applied to identify fuzzy classifiers. The description of the algorithm and the pseudo code of gravitational search algorithm are shown. The results of the test are compared with the results of similar algorithms using statistical test Mann-Whitney-Wilcoxon.

Keywords: fuzzy system, classification, gravitational search algorithm, features selections.