УДК 519.87

С.С. Волкова, Р.Б. Сергиенко

ОТБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ В НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКЕ РЕГРЕССИИ С ИСПОЛЬЗОВАНИЕМ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ

Рассматривается метод отбора информативных признаков в непараметрической оценке регрессии, основанный на использовании генетических алгоритмов. Идея метода заключается в оптимизации параметров размытия признаков генетическими алгоритмами и в последующем исключении признаков, которым соответствуют наибольшие значения параметров размытия. Проведены исследования метода на задачах различной размерности при различных настройках генетического алгоритма.

Ключевые слова: *непараметрическая оценка регрессии*, *генетический алгоритм*, *отбор информативных признаков.*

Одной из ключевых проблем в решении разнообразных задач анализа данных (оценка регрессии, распознавание образов, кластеризация, прогнозирование) является отбор информативных признаков. Реальные процессы в технических и организационных системах могут описываться десятками и сотнями различных признаков. При этом не всегда все из них являются существенными или значимыми, то есть необходимыми для построения адекватной модели процесса (регрессионной модели, классификационной модели и др.). Кроме того, актуальность отбора информативных признаков становится особенно ощутимой в связи с характерной для большинства алгоритмов анализа данных проблемой «проклятья размерности». Эта проблема заключается в резком падении эффективности алгоритма или резком увеличении требуемого вычислительного ресурса для эффективной работы алгоритма при увеличении размерности (увеличении числа признаков) решаемой задачи анализа данных.

На сегодняшний день предложено большое число методов отбора информативных признаков или снижения размерности [1]: метод главных компонент, модели и методы факторного анализа, многомерное шкалирование и другие. Каждый из разработанных методов обладает своими преимуществами и недостатками, во многих случаях есть ограничения на применение того или иного метода. Поэтому научно-техническое направление, связанное с разработкой новых методов снижения размерности или отбора информативных признаков, остается актуальным.

В настоящей работе рассматривается задача отбора информативных признаков для регрессионных моделей, основанных на непараметрической оценке Надарая – Ватсона [2]. Преимущество такой оценки заключается в отсутствии необходимости подбирать структуру регрессионной модели, что сделало её распространённой и популярной для моделирования разнообразных процессов, особенно в технических системах.

Построение непараметрической оценки регрессии сводится к подбору наилучших значений так называемых параметров размытия для признаков задачи, то есть к оптимизации оценки регрессии по параметрам размытия. При этом данная задача оптимизации характеризуется отсутствием аналитического вида целевой

функции (она задана процедурно) и потенциально высокой размерностью (в зависимости от решаемой задачи), что делает затруднительным или даже невозможным использование многих классических методов оптимизации. Для решения подобного рода задач оптимизации хорошо зарекомендовали себя генетические алгоритмы [3], поэтому и предлагается их использование для настройки параметров размытия непараметрической оценки регрессии. Работа генетических алгоритмов основана на использовании подобия природного эволюционного процесса, приводящего к улучшению и адаптации к окружающей среде живых организмов.

Непараметрическая оценка регрессии обладает ещё и тем свойством, что для малоинформативных признаков оптимальные значения параметров размытия стремятся к большим величинам. Следовательно, поиск оптимальных, или хотя бы субоптимальных, значений параметров размытия позволит выявлять малоинформативные признаки, которые можно рассматривать в качестве кандидатов на исключение из рассматриваемой задачи анализа данных.

Таким образом, данная работа посвящена исследованию метода отбора информативных признаков в непараметрической оценке регрессии на основе использования генетических алгоритмов для оптимизации параметров размытия и последующего выявления малоинформативных признаков.

При этом в рамках исследования были поставлены следующие задачи:

- провести исследования предлагаемого метода на задачах различной размерности;
- провести исследования предлагаемого метода при зашумленности обучающих выборок;
- провести исследования предлагаемого метода при различных настройках генетического алгоритма, так как использование генетических алгоритмов сопряжено с проблемой выбора настроек алгоритма, таких, как тип селекции, тип скрещивания, частота мутации и других [4].

1. Непараметрическая оценка регрессии

Рассмотрим подробнее непараметрическую оценку регрессии Надарая – Ватсона.

Пусть $(x_1, x_2, \dots, x_n)$ – вектор значений признаков, y – значение регрессии. Предположим, что имеется обучающая выборка значений признаков и соответствующих значений регрессии длиной N. Тогда непараметрическая оценка регрессии для вектора признаков $(x_1^*, x_2^*, \dots, x_n^*)$ выглядит следующим образом [2]:

$$\hat{y}(\overline{x}^*) = \sum_{i=1}^{N} y_i \cdot \prod_{j=1}^{n} \Phi\left(\frac{x_j^i - x_j^*}{c_j}\right) \Bigg/ \sum_{i=1}^{N} \prod_{j=1}^{n} \Phi\left(\frac{x_j^i - x_j^*}{c_j}\right), \qquad (1)$$

где c_j – параметры размытия, $\Phi(\dots)$ – колоколообразная функция. Один из распространенных видов колоколообразной функции следующий [2]:

$$\Phi(t) = \begin{cases} 1 - |t|, & \text{если } |t| < 1, \\ 0, & \text{иначе.} \end{cases}$$

При построении непараметрической оценки регрессии вводится критерий качества оценки W, который обычно определяется как среднеквадратическая ошибка полученных оценок от истинных значений регрессии по тестовой выборке объёма N_t:

$$W = \frac{1}{N_t}\sum_{i=1}^{N_1}(\hat{y}_i - y_i)^2 . \qquad (2)$$

Задача построения непараметрической оценки регрессии сводится к подбору наилучших значений параметров размытия c_j, то есть к минимизации критерия качества оценки W по параметрам размытия c_j.

Обратим внимание, что для малоинформативных признаков оптимальный параметр размытия c_j будет иметь тенденцию к увеличению. Действительно, при устремлении c_j к бесконечности ($c_j \to \infty$) аргумент функции $\Phi(t)$ стремится к нулю, при этом $\Phi(0)=1$. Из формулы (1) видно, в такой ситуации оценка регрессии полностью перестаёт зависеть от значений признака, параметр размытия для которого стремится к бесконечности.

2. Генетический алгоритм для оптимизации параметров размытия и отбора информативных признаков

Генетический алгоритм (ГА) относится к классу стохастических алгоритмов оптимизации [5]. Преимущество генетических алгоритмов перед другими методами оптимизации в способности эффективно решать многомерные, многоэкстремальные задачи; при зашумлённости целевой функции, её неявном (например, алгоритмическом) задании; при дискретности переменных.

Название генетического алгоритма объясняется тем, что в основе него лежит имитация процессов, происходящих в природе среди особей какой-либо популяции. Индивид или особь представляет собой решение (вектор значений параметров), закодированное произвольным образом, например в бинарную строку-хромосому. Совокупность решений в фиксированный момент времени составляет популяцию. Каждый индивид обладает пригодностью, привязанной к значению целевой функции. Индивиды текущей популяции конкурируют друг с другом за передачу своей генетической информации (создание потомков) в следующую популяцию. Отобранные с помощью селекции индивиды из текущей популяции проходят этапы создания новых решений-потомков – рекомбинации и мутации. Селекция, рекомбинация и мутация относятся к основным операторам генетического алгоритма. Распространённые типы селекции в ГА: пропорциональная, турнирная, ранговая; распространённые типы рекомбинации (скрещивания): одноточечное, двухточеное, равномерное. Также возможна различная частота мутации. Видно, что существует большое число различных комбинаций настроек генетического алгоритма.

Одно из основных проблем в использовании генетических алгоритмов заключается в том, результат работы алгоритма сильно зависит от выбора комбинации его настроек. Наилучшей универсальной комбинации настроек не существует [4]. Главной причиной этому является то, что в процессе работы генетический алгоритм реализует две стратегии. Первая стратегия – исследование, ее целью является поиск новых областей решений. Применение этой стратегии наиболее обосновано на начальных этапах поиска. В генетическом алгоритме эту стратегию реализует оператор мутации. Вторая стратегия – использование, применяется для улучшения существующего решения, этому следовало бы уделять больше внимания на заключительных этапах работы алгоритма оптимизации. В генетическом алгоритме эту функцию выполняет оператор скрещивания. Вследствие этого можно считать обоснованной идею уменьшения влияния оператора мутации в течение работы генетического алгоритма, но стандартный генетический алгоритм использует обе стратегии в постоянных (для одного запуска) пропорциях.

В данной работе генетический алгоритм используется для оптимизации (минимизации) критерия качества оценки непараметрической оценки регрессии (2) по параметрам размытия, далее определяются максимальные значения параметра размытия, соответствующие наименее информативных признакам.

При исследовании генетического алгоритма на рассматриваемой задаче отбора информативных признаков в непараметрической оценке регрессии использовались три типа селекции (пропорциональная, турнирная с турниром 3, ранговая), три типа скрещивания (одноточечное, двухточечное, равномерное), а также различные варианты мутации. В работе рассматривались следующие варианты адаптивной мутации, взятые из [6]:

$$p_t = \frac{1}{240} + \frac{0{,}11375}{2^t},\ p_t = \left(2 + \frac{m-2}{T-1}\cdot t\right)^{-1},$$

где t – текущее поколение, m – число генов в хромосоме, T – максимальное число поколений, p_t – эмпирическая вероятность (частота) мутации в поколении t. Кроме адаптивной мутации в работе рассматривались разные виды постоянной мутации: очень слабая ($p = 1/(9m)$), слабая ($p = 1/(3m)$), средняя ($p = 1/m$), сильная ($p = 3/m$), очень сильная ($p = 9/m$). Именно исследованию сравнительной эффективности различных видов мутации уделено особое внимание в данной работе.

3. Результаты численных исследований

Для исследования предлагаемого метода отбора информативных признаков были взяты четыре тестовые функции различной размерности:

1) $y(\overline{x}) = 0{,}01\cdot x_1 + 7\cdot x_2 + 5\cdot x_3$;
2) $y(\overline{x}) = 0{,}01\cdot x_1 + 7\cdot x_2 + 5\cdot x_3 + 12\cdot x_4 + 8\cdot x_5$;
3) $y(\overline{x}) = 0{,}01\cdot x_1 + 7\cdot x_2 + 5\cdot x_3 + 12\cdot x_4 + 8\cdot x_5 + 15\cdot x_6 + 3\cdot x_7$;
4) $y(\overline{x}) = 0{,}01\cdot x_1 + 7\cdot x_2 + 5\cdot x_3 + 12\cdot x_4 + 8\cdot x_5 + 15\cdot x_6 + 3\cdot x_7 + 9\cdot x_8 + 13{,}5\cdot x_9$.

Видно, что во всех указанных четырёх функциях есть переменная (признак) с малым весовым коэффициентом, то есть являющаяся малоинформативной. Поэтому задача – выявить именно эти признаки с использованием предлагаемого подхода.

Обучающая выборка объёмом 100 для каждой задачи генерировалась случайным образом из интервала [0; 3] с равномерным законом распределения для каждой переменной. Проводились исследования без наложения помехи и с наложением помехи в 10 % на значения обучающей выборки. Интервал варьирования для параметров размытия [0,001; 10]. Ресурс алгоритма – 50 индивидов на 50 поколений. Генетический алгоритм запускался по 20 раз для каждой комбинации настроек (3 типа селекции × 3 типа скрещивания × 7 типов мутации = 63 комбинации настроек) с усреднением значений параметров размытия для каждой переменной. В каждом запуске алгоритма определяется наименее значимый признак, затем вычисляется среднеквадратичная ошибка непараметрической модели, полученная удалением найденного малоинформативного признака. Для сравнения также указаны среднеквадратичные ошибки, полученные изъятием каждого из признаков, а также при включении всех признаков в регрессионную модель. В таблицах приведены результаты численных исследований с усредненными показателями для различных типов мутации в генетическом алгоритме. Жирным шрифтом обозначены наименьшие значения ошибки.

Таблица 1

Результаты исследования на задаче 1 без помехи

Мутация	Очень слабая	Слабая	Средняя	Сильная	Очень сильная	Адапт. 1	Адапт. 2
Параметр размытия 1	7,917	8,211	7,646	5,884	4,039	7,681	4,419
Параметр размытия 2	0,468	0,435	0,398	0,453	0,522	0,433	0,547
Параметр размытия 3	0,574	0,574	0,577	0,616	0,696	0,611	0,742
Среднекв. ошибка без признака 1	**0,590**	0,641	0,594	0,771	1,030	0,727	1,611
Среднекв. ошибка без признака 2	36,141	38,028	39,183	39,587	40,157	38,678	40,747
Среднекв. ошибка без признака 3	20,094	19,266	18,931	20,722	20,322	19,381	20,761
Среднекв. ошибка со всеми приз.	**0,590**	0,645	0,606	0,781	1,145	0,729	1,726

Таблица 2

Результаты исследования на задаче 1 с помехой в 10 %

Мутация	Очень слабая	Слабая	Средняя	Сильная	Очень сильная	Адапт. 1	Адапт. 2
Параметр размытия 1	8,384	6,916	7,296	6,070	4,328	7,678	4,212
Параметр размытия 2	0,503	0,416	0,456	0,460	0,600	0,490	0,582
Параметр размытия 3	0,614	0,618	0,584	0,630	0,704	0,610	0,750
Среднекв. ошибка без признака 1	0,585	0,571	0,614	0,571	1,301	0,620	1,634
Среднекв. ошибка без признака 2	39,113	37,775	40,098	37,668	40,423	37,799	41,063
Среднекв. ошибка без признака 3	19,499	20,064	20,605	19,348	19,420	19,267	20,669
Среднекв. ошибка со всеми приз.	0,589	**0,565**	0,621	0,583	1,378	0,623	1,796

Таблица 3

Результаты исследования на задаче 2 без помехи

Мутация	Очень слабая	Слабая	Средняя	Сильная	Очень сильная	Адапт. 1	Адапт. 2
Параметр размытия 1	6,748	6,300	6,647	5,211	4,465	6,292	4,845
Параметр размытия 2	1,238	1,134	1,262	1,231	1,400	1,200	1,678
Параметр размытия 3	1,631	1,611	1,581	1,639	1,910	1,589	1,947
Параметр размытия 4	0,813	0,876	0,815	0,831	0,896	0,758	0,818
Параметр размытия 5	1,251	1,168	1,101	1,182	1,171	1,123	1,316
Среднекв. ошибка без признака 1	13,665	14,909	**13,443**	15,695	22,691	13,728	24,311
Среднекв. ошибка без признака 2	49,069	48,225	48,627	49,668	54,462	49,204	52,254
Среднекв. ошибка без признака 3	29,990	28,757	29,017	31,165	35,063	30,887	35,853
Среднекв. ошибка без признака 4	123,096	118,200	122,023	123,140	122,636	125,228	130,616
Среднекв. ошибка без признака 5	61,366	58,390	63,574	60,747	65,352	61,233	64,402
Среднекв. ошибка со всеми приз.	13,649	14,811	13,613	15,815	23,283	13,678	24,978

Таблица 4

Результаты исследования на задаче 2 с помехой в 10 %

Мутация	Очень слабая	Слабая	Средняя	Сильная	Очень сильная	Адапт. 1	Адапт. 2
Параметр размытия 1	6,943	6,955	6,492	4,991	4,337	6,394	4,779
Параметр размытия 2	1,176	1,228	1,234	1,222	1,364	1,226	1,571
Параметр размытия 3	1,632	1,539	1,613	1,648	2,194	1,649	2,156

Окончание табл. 4

Параметр размытия 4	0,939	0,801	0,826	0,842	0,895	0,905	0,841
Параметр размытия 5	1,152	1,169	1,086	1,148	1,233	1,145	1,308
Среднекв. ошибка без признака 1	14,170	16,838	**12,714**	16,606	24,536	15,763	27,137
Среднекв. ошибка без признака 2	49,265	52,213	43,658	50,065	54,139	50,228	54,636
Среднекв. ошибка без признака 3	30,262	32,125	27,876	31,278	36,319	29,323	40,413
Среднекв. ошибка без признака 4	117,429	133,350	116,187	122,692	138,229	129,352	127,646
Среднекв. ошибка без признака 5	61,134	64,071	60,620	64,293	69,090	63,991	70,642
Среднекв. ошибка со всеми приз.	14,135	16,893	12,756	16,639	25,441	15,837	27,520

Таблица 5

Результаты исследования на задаче 3 без помехи

Мутация	Очень слабая	Слабая	Средняя	Сильная	Очень сильная	Адапт. 1	Адапт. 2
Параметр размытия 1	7,001	7,731	6,579	5,502	5,425	6,668	5,383
Параметр размытия 2	1,677	1,654	1,878	1,685	2,073	1,741	2,630
Параметр размытия 3	2,097	2,536	2,229	2,587	2,826	2,274	2,750
Параметр размытия 4	1,116	1,118	1,088	1,125	1,199	1,137	1,443
Параметр размытия 5	1,590	1,464	1,565	1,525	2,096	1,796	1,790
Параметр размытия 6	1,002	0,906	1,046	1,047	1,044	1,086	1,001
Параметр размытия 7	4,106	3,686	3,490	3,308	3,016	3,905	2,992
Среднекв. ошибка без признака 1	41,869	44,736	43,840	43,448	65,093	**39,380**	64,899
Среднекв. ошибка без признака 2	71,004	76,450	67,988	70,447	89,409	70,434	86,685
Среднекв. ошибка без признака 3	54,900	55,748	57,026	55,216	74,770	51,218	72,460
Среднекв. ошибка без признака 4	154,621	153,418	152,824	147,041	172,396	147,010	151,414
Среднекв. ошибка без признака 5	78,118	86,167	86,499	83,906	95,859	75,783	101,116
Среднекв. ошибка без признака 6	205,639	204,568	213,707	202,851	222,672	207,912	226,348
Среднекв. ошибка без признака 7	43,478	47,592	46,942	46,334	66,083	41,082	65,500
Среднекв. ошибка со всеми приз.	41,872	45,130	43,848	43,542	65,896	39,484	65,604

Таблица 6

Результаты исследования на задаче 3 с помехой в 10 %

Мутация	Очень слабая	Слабая	Средняя	Сильная	Очень сильная	Адапт. 1	Адапт. 2
Параметр размытия 1	7,563	7,254	6,726	5,796	5,528	6,678	5,478
Параметр размытия 2	1,624	1,750	1,761	1,758	2,264	1,428	2,538
Параметр размытия 3	2,447	2,435	2,271	2,509	2,651	2,659	2,284
Параметр размытия 4	1,121	1,189	1,171	1,127	1,188	1,154	1,296
Параметр размытия 5	1,573	1,462	1,622	1,536	1,925	1,610	1,955
Параметр размытия 6	1,004	0,990	1,053	1,053	1,014	1,019	1,035
Параметр размытия 7	4,435	3,462	3,553	3,375	3,179	4,084	3,165
Среднекв. ошибка без признака 1	46,831	42,866	**37,953**	42,931	62,187	44,582	74,501
Среднекв. ошибка без признака 2	79,046	67,534	64,229	71,031	83,442	76,157	94,404
Среднекв. ошибка без признака 3	57,588	54,121	50,088	53,683	70,767	53,755	81,467
Среднекв. ошибка без признака 4	148,348	146,173	143,587	142,853	163,054	148,487	164,002
Среднекв. ошибка без признака 5	89,129	83,019	76,377	82,172	97,027	83,528	102,881
Среднекв. ошибка без признака 6	214,745	211,390	199,436	212,752	219,040	212,389	235,815
Среднекв. ошибка без признака 7	48,767	45,668	40,705	45,824	63,813	46,860	74,690
Среднекв. ошибка со всеми приз.	46,754	42,891	38,088	43,471	62,561	44,642	74,771

Таблица 7

Результаты исследования на задаче 4 без помехи

Мутация	Очень слабая	Слабая	Средняя	Сильная	Очень сильная	Адапт. 1	Адапт. 2
Параметр размытия 1	7,471	7,199	6,585	5,945	5,816	6,608	5,754
Параметр размытия 2	2,874	2,282	2,703	2,321	3,025	2,471	2,761
Параметр размытия 3	3,335	3,640	2,998	3,477	3,145	2,969	3,088
Параметр размытия 4	1,415	1,445	1,492	1,496	1,812	1,578	1,824
Параметр размытия 5	1,935	2,043	1,996	2,361	2,568	2,147	2,703
Параметр размытия 6	1,291	1,316	1,229	1,261	1,180	1,248	1,228
Параметр размытия 7	4,899	4,281	3,851	4,141	3,812	3,734	3,767
Параметр размытия 8	2,205	2,142	1,846	2,011	2,105	1,930	2,251
Параметр размытия 9	1,239	1,270	1,431	1,392	1,500	1,348	1,531
Среднекв. ошибка без признака 1	105,073	93,725	100,095	119,858	161,659	94,305	163,303
Среднекв. ошибка без признака 2	124,998	120,887	118,082	145,867	179,104	116,862	181,159
Среднекв. ошибка без признака 3	112,750	99,911	112,855	128,840	169,306	103,724	170,603
Среднекв. ошибка без признака 4	204,091	188,454	205,834	210,839	240,435	184,640	252,814
Среднекв. ошибка без признака 5	141,475	126,050	139,358	149,211	185,662	128,417	191,206
Среднекв. ошибка без признака 6	267,195	245,714	264,541	285,697	318,998	241,896	311,846
Среднекв. ошибка без признака 7	105,978	95,0553	102,996	121,077	158,801	97,490	162,247
Среднекв. ошибка без признака 8	149,247	133,191	149,692	159,272	196,230	137,032	204,620
Среднекв. ошибка без признака 9	248,216	218,108	219,263	241,911	267,340	212,029	278,311
Среднекв. ошибка со всеми приз.	104,652	**93,623**	100,293	120,930	163,065	94,335	164,284

Таблица 8

Результаты исследования на задаче 4 с помехой в 10 %

Мутация	Очень слабая	Слабая	Средняя	Сильная	Очень сильная	Адапт. 1	Адапт. 2
Параметр размытия 1	7,804	7,846	7,411	6,451	5,638	7,276	5,808
Параметр размытия 2	1,982	2,965	2,286	2,564	3,146	2,281	2,515
Параметр размытия 3	3,200	3,372	3,627	3,600	2,958	3,521	2,914
Параметр размытия 4	1,565	1,506	1,454	1,221	1,828	1,418	1,873
Параметр размытия 5	2,495	2,161	2,300	2,215	2,759	2,278	2,882
Параметр размытия 6	1,185	1,200	1,343	1,274	1,196	1,230	1,238
Параметр размытия 7	4,799	5,014	4,979	4,174	3,687	4,263	3,439
Параметр размытия 8	2,035	1,916	1,768	1,953	2,226	2,099	2,310
Параметр размытия 9	1,309	1,363	1,236	1,207	1,251	1,447	1,540
Среднекв. ошибка без признака 1	108,802	104,548	**103,278**	110,128	172,413	110,136	161,865
Среднекв. ошибка без признака 2	140,405	121,808	128,702	131,260	189,477	135,557	182,102
Среднекв. ошибка без признака 3	116,349	113,316	109,455	115,295	177,679	118,704	167,849
Среднекв. ошибка без признака 4	199,495	198,846	201,369	213,642	244,206	203,395	241,838
Среднекв. ошибка без признака 5	135,114	133,140	138,789	139,074	194,955	143,403	186,799
Среднекв. ошибка без признака 6	270,854	267,295	250,053	286,834	332,546	261,031	320,799
Среднекв. ошибка без признака 7	110,467	106,276	104,923	111,871	168,673	112,656	157,546
Среднекв. ошибка без признака 8	155,108	154,755	160,347	149,426	203,050	150,585	199,607
Среднекв. ошибка без признака 9	227,572	234,789	229,838	237,740	287,421	221,972	274,727
Среднекв. ошибка со всеми приз.	109,133	104,825	103,524	110,663	173,356	110,698	162,774

Из табл. 1 – 8 видно, что алгоритм действительно определяет максимальное значение параметра размытия для наименее информативного признака (признак 1

во всех задачах). Более того, зачастую среднеквадратичная ошибка непараметрической оценки регрессии, получаемой в результате исключения наименее информативного признака, оказывается меньше ошибки при использовании всех признаков. Накладывание помехи на значения признаков элементов обучающей выборки не приводит к ухудшению работоспособности алгоритма. Следует отметить, что на разных задачах показывают наибольшую эффективность различные типы мутации, что подтверждает актуальность проблемы выбора настроек генетического алгоритма.

Заключение

Таким образом, разработана процедура отбора информативных признаков в непараметрической оценке регрессии на основе использования генетических алгоритмов для оптимизации параметров размытия и дальнейшего исключение малоинформативных признаков, соответствующих наибольшим значениям параметра размытия.

Проведены исследования разработанного метода на задачах различной размерности (3, 5, 7 и 9) , без помехи и с помехой в 10 %, на различных комбинациях настроек генетического алгоритма. Особое внимание уделено исследованию сравнительной эффективности различных типов мутации в генетическом алгоритме.

Можно сделать следующие выводы по результатам проведенных численных исследований:

1) Метод определяет наименее информативный признак на задачах различной размерности.

2) Для метода не является существенным наличие помех в значениях признаков элементов из обучающей выборки.

3) На различных задачах могут быть эффективными различные настройки генетического алгоритма, в том числе различные типы мутации, что делает актуальным проблему выбора наилучших настроек генетического алгоритма.

Разработанный метод может быть использован при построении регрессионных моделей реальных процессов, для которых является существенной задача отбора информативных признаков.

ЛИТЕРАТУРА

1. *Айвазян С.А.и др.* Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.
2. *Медведев А.В.* Непараметрические системы адаптации. Новосибирск: Наука, 1983. 174 с.
3. *Goldberg D.E.* Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley, 1989.
4. *Сергиенко Р.Б.* Исследование эффективности коэволюционного генетического алгоритма условной оптимизации // Вестник Сибирского государственного аэрокосмического университета имени академика М.Ф. Решетнёва. 2009. № 3 (24). С. 31–36.
5. *Семенкин Е.С.*, *Семенкина О.Э.*, *Коробейников С.П.* Оптимизация технических систем: учеб. пособие. Красноярск: СИБУП, 1996. 284 с.
6. *Daridi F.*, *Kharma N.*, *Salik J.* Parameterless genetic algorithms: review and innovation // IEEE Canadian Review. Summer 2004. No. 47. P. 19–23.

Волкова Светлана Сергеевна
Сергиенко Роман Борисович
Сибирский государственный аэрокосмический университет
им. акад. М.Ф. Решетнёва (г. Красноярск)
E-mail: sv-vol@yandex.ru; romaserg@list.ru

Поступила в редакцию 14 мая 2012 г.

Volkova Svetlana S., *Sergienko Roman B.* (Reshetnev Siberian State Aerospace University).
Informative attributes selection in nonparametric regression estimation by making use of genetic algorithms.

Keywords: nonparametric estimated regression, genetic algorithms, Informative attributes selection.

A method of informative attributes selection in nonparametric regression estimation based on genetic algorithms is considered. The idea of the method consists in optimization of attributes fuzzy parameters using genetic algorithms and elimination of attribute with maximum value of fuzzy parameter.

Investigation of the method for problems with different dimension (3, 5, 7, and 9), without noise and with 10% noise, for different setting of genetic algorithm parameters was performed. Special attention was paid to investigation of comparative efficiency for different mutation types at genetic algorithm.

It is possible to draw following conclusions based on numerical experiments:

1) The method defines the least informative attribute.

2) Noise is not essential for efficiency of the method.

3) Different settings of genetic algorithm parameters for different problems can be effective. So the problem of genetic algorithm parameters setting is actual.