УДК 621.513

А.А. Самойленко

Весовой критерий определения информативности аргументов в методах построения моделей с последовательной селекцией переменных

Рассмотрены подходы определения степени информативности аргументов в селекционных алгоритмах переборного типа, показаны их преимущества и недостатки. Разработан новый весовой критерий информативности аргументов.

The paper considers the existing approaches to determine the arguments informativenes degree in the successive selection algorithms. Their advantages and disadvantages are presented, a new weight criterion is developed.

Розглянуто підходи визначення ступеня інформативності аргументів в селекційних алгоритмах переборного типу, показано їх переваги та недоліки. Розроблено новий ваговій критерій інформативності аргументів.

Введение. В селекционных алгоритмах МГУА [1, 2] переборного типа существует необходимость определения степени информативности аргументов, рассмотрены подходы, позволяющие это сделать, а также указаны их преимущества и недостатки.

В работе [3] для определения степени информативности переменных предлагается использовать частотный критерий. Там же обосновано его использование и показаны его достоинства. В ходе экспериментов выявлен ряд недостатков этого критерия, а также условия их проявления. В результате исследования разработан и предложен новый весовой критерий определения степени информативности аргументов, показана его эффективность и существенные преимущества как перед частотным критерием, так и предложенными ранее подходами.

Постановка задачи

Задача идентификации [1, 2] состоит в формировании по данным выборки W некоторого множества $\Im$ моделей различной структуры вида $y_f = f(X, \widehat{\theta}_f)$ и отыскании оптимальной модели по минимуму заданного критерия $CR(\cdot)$

$$f^* = \arg\min_{f \in \Im} CR(y, f(X, \widehat{\theta}_f)). \qquad (1)$$

Здесь оценки параметров $\widehat{\theta}_f$, указывающие на набор ненулевых компонент (структуру) для каждой модели $f \in \Im$, определяются по формуле

$$\widehat{\theta}_f = \arg\min_{\theta_f \in R^{s_f}} Q(y, X, \theta_f), \qquad (2)$$

где $Q(\cdot) \neq CR(\cdot)$ – критерий качества решения задачи параметрической идентификации каждой частной модели, генерируемой в задаче структурной идентификации (1).

Задача структурной идентификации должна состоять в определении наилучшего приближения к неизвестным значениям s_0 (сложность истинной модели) и θ_0 (вектор коэффициентов переменных в истинной модели) при одновременном разделении входных переменных матрицы X на информативные и неинформативные. Под информативностью аргумента понимаем степень зависимости выходной величины от этого аргумента.

Пример 1. Пусть $m = 20$, $n = 250$, $s_0 = 10$. Создадим выборку X, $\dim X = 250 \times 20$, состоящую из векторов $x_j, j = 1,...,20$. Каждому x_j соответствует функция $f_j(t_k)$ (табл. 1). Векторы t_k, $k = 1,...,5$ сгенерированы с помощью датчика случайных равномерно распределенных чисел.

Т а б л и ц а 1. Зависимость переменных x от t_k

x:	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
$f(t)$:	t_1	t_2	t_3	t_1^2	t_1t_2	t_1t_3	t_1t_4	t_2^2	t_2t_3	t_3^2	t_4	t_5	t_1t_5	t_2t_4	t_2t_5	t_3t_4	t_3t_5	t_4^2	t_4t_5	t_5^2

Зададим следующий вектор параметров истинной модели: $\theta_0 = [-3; -3; 5; -1; -1; 3; 1; -2; 1; 1]^T$, а ее структурный вектор в виде такой последовательности нулей и единиц: $d = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$, т.е. первые 10 аргументов – истинные (от них и будет зависеть выходная переменная), остальные назовем избыточными. Используя формулу $y = \overset{o}{X}\theta_0$ и шум ξ (уровень шума в данном примере составляет 2 процента от истинного выхода $\overset{o}{y}$), вычислим выходную величину:

$$y = \overset{o}{X}\theta_0 + \xi = -3x_1 - 3x_2 + 5x_3 - x_4 - \\ - x_5 + 3x_6 + x_7 - 2x_8 + x_9 + x_{10} + \xi. \quad (3)$$

Здесь $\dim y = n \times 1$, $\dim \overset{o}{X} = n \times s_0$, $\dim \theta_0 = s_0 \times 1$, $\dim \xi = n \times 1$, а x_j, $j = 1,...,s_0$ – вектор–столбцы матрицы $\overset{o}{X}$.

Пример 2. На основании примера 1 сформируем выборку из 200 аргументов путем дополнения неистинных аргументов к этой выборке. Таким образом, оба примера имеют одинаковые истинные модели, но разное количество аргументов.

Теперь задачей будет восстановление вектора θ_0 путем поиска оптимальной модели $\overset{o}{y} = \overset{o}{X}\theta_0$ на всей выборке $W(X, y)$ по некоторому критерию $CR(\cdot)$. Воспользуемся критерием регулярности AR [2].

В работе [3] показано, что решение задачи с количеством аргументов более 30 с помощью алгоритма МГУА с полным перебором становится невозможным.

Методы определения степени информативности аргументов позволяют отсеять неинформативные и тем самым решить задачу структурной идентификации без выполнения полного перебора.

Таким образом, цель этой статьи – выполнить анализ существующих подходов определения подмножества информативных аргументов и на основании проведенных исследований предложить новый подход, а также показать его преимущества в сравнении с рассмотренными ранее методами.

Существующие подходы к решению проблемы

В работах [2, 4, 5] степень информативности аргументов предлагается оценивать по величине модуля коэффициента парной корреляции аргумента с выходной переменной. Модель строится только для отобранных наиболее информативных аргументов.

Недостаток этого подхода продемонстрируем на приведенном ранее примере. На рис. 1 построен график, где по оси ординат отложено значение модуля коэффициента парной корреляции каждого аргумента, указанного на оси абсцисс, с выходной величиной.

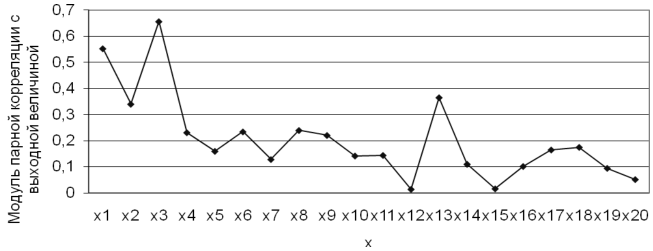


Рис. 1. Значение модуля парной корреляции каждого аргумента с выходной величиной

Теперь очевидно, что подход с использованием парной корреляции имеет ряд недостатков. Количество отбираемых в модель аргументов зависит от порогового значения, т.е. минимального значения корреляции, при котором аргумент может быть включен в модель. При малых его значениях в модель попадает много лишних аргументов, а при высоких – существует большая вероятность исключения истинных аргументов из модели. Задача нахождения наиболее информативных аргументов и оптимальной модели в этом случае дополняется определением оптимального порога коэффициента корреляции.

В работе [6] предлагалось разбивать последовательность аргументов, ранжированных по модулю парного коэффициента корреляции, на группы по 20–25 переменных. С помощью полного перебора для каждой группы находится лучшая модель по заданным критериям. Все аргументы, вошедшие в лучшие модели, объединяются и определяют собой пространство моделирования.

В модификации комбинаторного алгоритма [7] предложен похожий способ, когда из ранжированных по модулю коэффициента корреляции аргументов формируется два подмножества, первое из которых содержит около 25 аргументов, второе – все остальные. По комбинаторному алгоритму *COMBI* МГУА [1] строится оптимальная модель, и аргументы, не вошедшие в эту модель, удаляются из выборки. Далее выборка данных пополняется аргументами из другого подмножества, имеющими наибольшее значение модуля коэффициента корреляции, до 25 переменных и снова строится модель. Алгоритм продолжает работу до тех пор, пока не будут использованы все аргументы входной выборки.

Очевидно, что такой подход не позволит найти модель, сложность которой превышает количество аргументов, отбираемых в первое множество. Еще один недостаток этого подхода состоит в том, что на первом же этапе будут отобраны наиболее коррелируемые с выходной величиной аргументы, и последующее добавление менее коррелируемых переменных не позволит существенно изменить модель. В итоге, после долгих итераций, будет получена модель, мало в чем уступающая построенной с помощью более быстрого подхода, описанного в [2, 4, 5].

В работе [8] предложено вместо парной корреляции использовать множественную. В соответствии с этим подходом вначале по многорядному алгоритму МГУА, обладающему большим быстродействием, отбирается F наилучших моделей, составляющих множество $\Im_F$. Степень информативности каждого аргумента определяется числом его попаданий в лучшие модели. Этот критерий определения степени информативности аргументов назовем частотным, так как он определяет, как часто конкретный аргумент присутствует в построенных моделях из множества $\Im_F$:

$$FC_j = \frac{q_j}{F}\ , \tag{4}$$

где q_j – число моделей из множества $\Im_F$, в которых присутствует j-й аргумент.

Следует отметить, что аргументы, имеющие большее значение критерия *FC*, считаются более информативными.

Основываясь на подходе [8], в [9–11] предложен новый многоэтапный метод переборного типа, в котором информативность аргументов оценивается по критерию *FC* и на каждом этапе отсеиваются наименее информативные. В работе [3] проведен анализ эффективности *FC* критерия совместно с *FSS* (*Forward Successive Selection*), *BSS* (*Backward Successive Selection*) и *CSS* (*Combined Successive Selection*) методами последовательного отсеивания аргументов.

Все три метода разбивают процесс поиска лучшей модели на несколько этапов. На каждом из них с помощью алгоритма *COMBIS* строится множество $\Im_F$ моделей определенной сложности. Затем с помощью частотного критерия оценивается степень информативности аргументов. Наименее информативные аргументы исключаются из выборки. Такие действия выполняются до тех пор, пока количество аргументов в выборке W на каком-то этапе не позволит выполнить полный перебор. Метод *FSS* рассматривает модели малой сложности ($s = 1 ... s_{\max}$), на каждом этапе пытаясь увеличить $s_{\max}$. В методе *BSS* процесс выполняется в обратном порядке, начиная с моделей большей сложности ($s = m - s_{\max} ... m$). Метод *CSS* – комбинация двух предыдущих подходов, в котором выполнятся перебор моделей как малой, так и большой сложности.

Метод *BSS* показал наилучшие результаты при решении задач с большим количеством аргументов. Такой результат обоснован тем, что *BSS* на каждом этапе рассматривает модели высокой сложности, тем самым учитывая не попарную и не частичную корреляцию, а корреляцию всех истинных аргументов с выходной величиной.

Также в [3] использование частотного критерия *FC* было обосновано тем, что значение критерия *AR* для моделей $f'_S \in \Im'_S$, содержащих все истинные аргументы, значительно меньше минимального значения критерия *AR* для моделей $f^{\sim}_S \in \Im^{\sim}_S$, не содержащих ни одного истинного аргумента. Таким образом, частотный критерий имеет смысл, поскольку:

$$\min(AR(f^{\sim})) > \max(AR(f')) \,. \tag{5}$$

Это условие также подтверждает и целесообразность использования метода *BSS*.

Но обратим внимание на результаты некоторых экспериментов, приведенных в табл. 2.

Таблица 2. Сравнение эффективности работы алгоритма *BSS* с алгоритмом *COMBI* при решении задач с числом аргументов m = 20 и m = 200

Алгоритмы	m	F	Число истинных аргументов в лучшей модели	$P_{l,max}$	Количество построенных моделей	Время выполнения, с	AR	Ошибка модели на экзаменационных данных, %
COMBI	20	m	10		1048575	21	2,36	2,8
BSS	20	m_l	10	100000	125 994	3	2,36	2,8
BSS	200	m_l	0	1000	5541	13	5592	>100
BSS	200	m_l	9	100000	508664	948	80,35	23,5
BSS	200	50	10	1000	10427	22	1,89	1,4

Используя свободу выбора $F = 50$, алгоритм находит модель, по критерию регулярности не уступающую модели, найденной с помощью полного перебора. Но при $F = m_l$ и многих других значениях F, а также при рассмотрении малого количества моделей, на первом же этапе отсеиваются истинные аргументы, так как они имеют минимальное значение частотного критерия. Причиной неспособности в некоторых случаях однозначно определить степень информативности аргумента по частотному критерию есть то, что условие (5) не всегда выполнимо.

Проведенные эксперименты показали, что при использовании критерия *FC* на выборках с большим количеством аргументов задача структурной идентификации решается не так однозначно, как хотелось бы. Некоторые истинные аргументы имеют худшее значение критерия *FC*, нежели неистинные, из-за чего возникает возможность потери истинного аргумента. Причем модуль коэффициента корреляции между аргументами не превышает 0,26.

Описание предлагаемого подхода к решению проблем частотного критерия

Недостаток критерия *FC* состоит в неспособности учитывать качество $CR(f)$ моделей, входящих во множество $\Im_F$, на которых оценивается аргумент. Как показали эксперименты, неинформативные аргументы имеют близкие к нулю коэффициенты в моделях, в результате чего они слабо влияют на значение выходной переменной. Следовательно, при оценке степени информативности аргумента необходимо учитывать также степень его влияния на выходную величину в построенной модели. Этого можно достичь путем совместного учета нормы вектора аргумента в выборке и его коэффициента в рассматриваемой модели.

Исходя из этих рассуждений, предлагается использовать следующий весовой критерий *WC* оценки степени информативности аргумента на множестве $\Im_F$:

$$WC(x_j, \theta_{jf}, CR(f)) = \frac{\|x_j\|}{n \cdot F} \sum_{k=1}^{F} \frac{\theta_{jf_k}^2}{CR(f_k)}, \quad (6)$$

где $\|x_j\|$ – норма вектора аргумента в выборке W; F – свобода выбора; θ_{jfk} – коэффициент аргумента x_j в модели f_k.

На нормированных данных квадрат нормы вектора и количество наблюдений n в формуле (6) можно не использовать.

В отличие от критерия *FC*, учитывающего только присутствие аргумента в лучшей модели, весовой критерий *WC* учитывает его вес в каждой модели и качество $CR(f)$ модели, в которую он входит.

Результаты проведенных экспериментов содержатся в табл. 3. Они свидетельствуют о том, что даже на ненормированной выборке весовой критерий способен выделить истинные аргументы независимо от величины свободы выбора.

В табл. 4 приведены значения *WC* оставшихся 20 аргументов на последнем селекционном этапе алгоритма *BSS* при решении задачи с 200 аргументами.

Очевидно, что истинные аргументы имеют существенно большие значения весового критерия в сравнении с неистинными, что позволяет четко выделить группу истинных аргументов.

Дополнительные эксперименты также показали, что критерий *WC* не так зависит от свободы выбора F, как критерий *FC*.

Эти результаты позволяют говорить об эффективности критерия *WC* в решении задачи структурной идентификации.

Т а б л и ц а 3. Сравнение значений весового критерия WC истинных x' и неистинных $x^{\sim}$ аргументов на нормированной и ненормированной выборках

Сложность моделей	m	F	$\max(WC(x'))$	$\min(WC(x'))$	$\max(WC(x^{\sim}))$	$\min(WC(x^{\sim}))$
Ненормированные данные						
199...200	200	200	0,81122	0,02712	0,00014	0,00001
199...200	200	190	0,81116	0,02711	0,00014	0,00001
199...200	200	100	0,75426	0,02519	0,00009	0,00001
Нормированные данные						
199...200	200	200	309,91233	28,25671	0,06265	0,0002
199...200	200	190	307,84483	28,04561	0,06204	0,00004
199...200	200	100	164,37579	14,99214	0,03337	0,00001

Т а б л и ц а 4. Сравнение значений критерия WC аргументов последнего этапа алгоритма *BSS* при решении задачи с числом аргументов $m = 200$

Неистинные аргументы										
x:	x_{96}	x_{22}	x_{134}	x_{136}	x_{87}	x_{49}	x_{50}	x_{125}	x_{133}	x_{47}
WC:	0,00036	0,00038	0,00040	0,00044	0,00049	0,00049	0,00050	0,00051	0,00053	0,00059
Истинные аргументы										
x:	x_4	x_6	x_3	x_9	x_8	x_7	x_5	x_1	x_0	x_2
WC:	0,37	0,41	0,43	0,44	0,47	1,96	3,26	3,64	4,32	11,34

Результаты экспериментов

Учитывая, что приведенные ранее результаты экспериментов получены на выборках с малым шумовым показателем $\sigma = 2\,\%$, рассмотрим зависимость значений $\min(AR(f^{\sim}))$ и $\max(AR(f'))$, а также значений весового критерия от уровня шума (табл. 5).

По результатам экспериментов (табл. 5) можно утверждать, что при повышении шума некоторые модели, не содержащие всех истинных аргументов, могут оказаться лучшими по критерию регулярности, чем некоторые модели, содержащие все истинные аргументы. Это и есть причиной ухудшения частотного и весового критериев. Тем не менее, следует заметить, что весовой критерий – более устойчив к шумовым показателям в сравнении с частотным критерием.

Т а б л и ц а 5. Зависимость значений $\min(AR(f^{\sim}))$, $\max(AR(f'))$ и весового критерия WC от s_0 и уровня шума σ

Сложность моделей	s_0	σ	$\min(AR(f^{\sim}))$	$\max(AR(f'))$	$\min(WC(x'))$	$\max(WC(x^{\sim}))$
1...20	10	2	24,25	4,44	83,17	0,14
		20	386,95	391,45	3,98	0,04
		40	2230,77	2344,87	0,79	0,24
199...200	10	2	7,81	3,46	28,25	0,06
		20	789,84	806,08	0,27	0,03
		40	1040,97	1153,55	0,08	0,04
199...200	100	2	39,66	47,24	0,76	0,01
		20	3481,61	4555,06	0,0017	0,019
		40	9568,06	12072,94	0,00003	0,016
199...200	180	2	122,41	121,72	0,04	0,003
		20	5895,41	6258,33	0,00001	0,001
		40	37439,08	38283,75	0,00001	0,0017

На рис. 2–3 показаны результаты экспериментов по исследованию зависимости частотного FC и весового WC критериев от свободы выбора F.

В случаях когда $\min(AR(f^{\sim})) > \max(AR(f'))$ и количество аргументов в рассматриваемых моделях приближается к m, частотный критерий сильно зависит от F (рис. 3). Наиболее приемлемое значение FC принимает при свободе выбора F, равной количеству моделей f' среди всех рассматриваемых, которое может быть вычислено по формуле

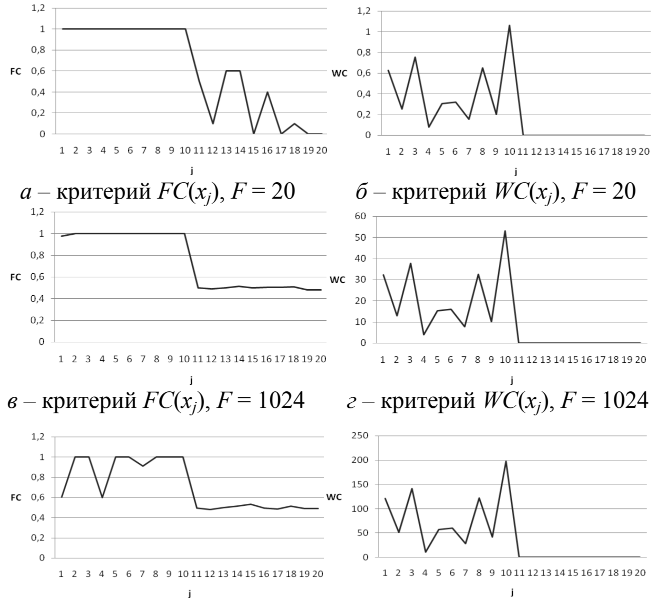
$$F' = P' = \sum_{s=s_{\min}}^{s_{\max}} C_{m-s_0}^{s-s_0}\ . \qquad (7)$$

Проведем эксперименты, когда $\min(AR(f^{\sim})) < \max(AR(f'))$ (рис. 3).

Как следует из рис. 2 и 3, весовой критерий ведет себя более устойчиво по отношению к частотному критерию при разных значениях свободы выбора. Даже в случае, когда F превышает количество рассматриваемых моделей (рис. 3,*д*,*е*), весовой критерий WC все еще хорошо работает, чего нельзя сказать о частотном критерии FC.

В случаях, когда $\min(AR(f^{\sim}))$ значительно меньше, чем $\max(AR(f'))$, и сложность рассматриваемых моделей максимально приближена к m (табл. 5), весовой критерий WC все еще работает хорошо (рис. 3,*б*,*г*,*е*), чего нельзя сказать о частотном критерии (рис. 3,*а*,*в*,*д*). Эта характеристика очень важна для алгоритма *BSS* при отсеивании аргументов на выборках с высоким значением m.

Способность весового критерия хорошо выделять информативные аргументы не только при малых количествах рассматриваемых моделей (P=201, s=199…200, m=200),

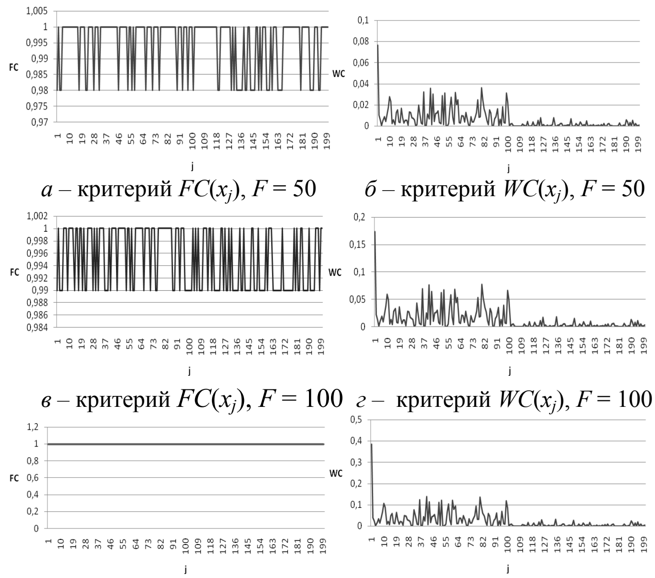


а – критерий $FC(x_j)$, $F = 20$ *б* – критерий $WC(x_j)$, $F = 20$

в – критерий $FC(x_j)$, $F = 1024$ *г* – критерий $WC(x_j)$, $F = 1024$

д – критерий $FC(x_j)$, $F = 4000$ *е* – критерий $WC(x_j)$, $F = 4000$

Рис. 2. Зависимость критериев *FC* и *WC* от свободы выбора *F* на выборке с уровнем шума σ = 20% и на моделях сложности: 1…20; $m = 20$, $s_0 = 10$



а – критерий $FC(x_j)$, $F = 50$ *б* – критерий $WC(x_j)$, $F = 50$

в – критерий $FC(x_j)$, $F = 100$ *г* – критерий $WC(x_j)$, $F = 100$

д – критерий $FC(x_j)$, $F = 500$ *е* – критерий $WC(x_j)$, $F = 500$

Рис. 3. Зависимость критериев *FC* и *WC* от свободы выбора *F* на выборке с уровнем шума σ = 40% и на моделях сложности: 199…200; $m = 200$, $s_0 = 100$

но и при одной ($P = 1$, $s = 600$, $m = 600$) позволяет решать задачи на выборках, которые имеют 600 и более аргументов. В табл. 6 приведены результаты работы алгоритмов на выборках с разными значениями m, s_0, σ. Здесь *ISA* – алгоритм [4], *KSA* – алгоритм [7], s^* – сложность (число аргументов) выбранной оптимальной модели.

Таблица 6. Сравнение результатов работы алгоритмов с применением весового критерия *WC*

m	s_0	σ	Алгоритмы	s^*	Число истинных аргументов в лучшей модели	Время выполнения, *с*	*AR*, %	Ошибка модели на экзаменационных данных, %
200	10	20	*KSA*	14	8	3756	9,2	8,2
			ISA	17	8	21	13,8	8,2
			MULTI	78	10	70	5,4	9,1
			FSS	14	9	4	9,4	8,2
			BSS	11	10	24	8,0	7,5
			CSS	11	10	27	8,0	7,5
		40	*ISA*	13	9	21	32,6	11,2
			MULTI (*F*=10)	90	9	101	17,3	13,0
			MULTI (*F*=*m*)	81	9	1669	17,5	12,9
			FSS	15	9	4	27,6	13,0
			BSS	20	10	24	24,8	11,4
			CSS	21	9	27	24,4	12,7
200	100	20	*ISA*	19	19	21	397,4	37,4
			FSS	15	15	4	54,8	18,3
			BSS	127	95	24	13,4	7,5
			CSS	102	92	26	14,9	7,5
		40	*ISA*	19	19	21	286,3	36,6
			FSS	15	15	4	94,7	19,8
			BSS	81	70	12	38,8	14,2
			CSS	65	60	14	40,2	14,5
200	180	20	*BSS*	163	159	12	17,0	6,6
		40	*BSS*	116	111	12	41,0	12,9
400	20	20	*BSS*	42	20	14	11,2	5,7
600	20	20	*BSS*	20	20	16	7,5	6,2
800	20	20	*BSS*	21	19	7	7,5	5,7

Как следует из табл. 6, алгоритм *KSA* строит модель, практически равную по качеству модели, построенной алгоритмом *ISA*, но за время, значительно превышающее время работы алгоритма *ISA*. Все алгоритмы *FSS*, *CSS*, *BSS* совместно с весовым критерием *WC* показали лучшие результаты в сравнении с алгоритмами *KSA* и *ISA* как по временным, так и по качественным показателям. Как видим, алгоритмы *CSS* и *BSS* способны эффективно находить модели, сложность которых превышает 20–25 аргументов, в отличие от остальных рассмотренных алгоритмов. К тому же в этом случае алгоритм *FSS* получает грубую, но более качественную модель, нежели алгоритм *ISA*.

Результаты экспериментов (табл. 6) показывают эффективность алгоритма *BSS* с использованием весового критерия *WC*. Алгоритм *BSS* совместно с *WC* показал наилучшие результаты в сравнении с другими алгоритмами, принимающими участие в экспериментах. Модели, построенные алгоритмом *BSS*, не уступают по качеству моделям, которые могли быть построены алгоритмом полного перебора. Эффективность критерия *WC* позволила алгоритму *BSS* принимать верное решение в отборе наиболее информативных аргументов, рассматривая на некоторых этапах только одну модель, и для решения задачи с $m = 800$ перебрать только 643 модели.

На рис. 4 показаны минимальные значения критерия регулярности моделей, построенных алгоритмом *BSS* при малых значениях $P_{\max}$, в сравнении с полным перебором. На рисунке не отображены минимальные значения критерия регулярности моделей сложности, меньшей 10, так как их значения в сотни раз превышают отображенные на рисунке и лежат в диапазоне [19,66; 853,41]. Следует отметить, что все минимальные значения критерия *AR* при s = 1…9 одинаковы для *COMBI* и *BSS* даже для случаев, когда *BSS* на каждом этапе строит всего одну модель. Это свидетельствует об эффективности весового критерия *WC*, использованного при этом.

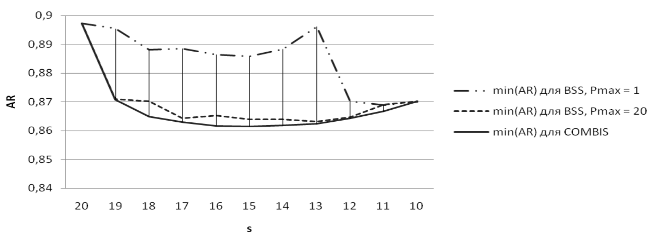


Рис. 4. Диапазон изменения значений критерия регулярности для моделей определенной сложности, построенных алгоритмом *BSS* при малых значениях $P_{\max}$ и $m = 20$, $s_0 = 10$

Анализируя проведенный эксперимент (рис. 4), можно увидеть, что алгоритм *BSS*, рассматривая на каждом этапе малое количество моделей, способен построить модель со всеми истинными аргументами, но немного грубее (с меньшим количеством неистинных аргументов), чем алгоритм полного перебора. Этот эксперимент еще раз доказывает, что алгоритм *BSS* вместе с критерием *WC* способен находить истинную модель, строя для этого небольшое количество моделей.

Заключение. Выполнен анализ существующих подходов отбора наиболее информативных переменных. Предложен новый весовой *WC* критерий оценки информативности аргументов и приведены его преимущества в сравнении с частотным критерием. Этот критерий показал хорошие результаты на тестовых задачах в сравнении с рассмотренными в работе подходами.

Критерий *WC* значительно повышает качество отбора информативных аргументов как на сильно зашумленных данных (около 40 процентов), так и на выборках с большим количеством переменных, когда условие (5), необходимое для критерия *FC*, не выполняется. В отличие от частотного критерия, критерий *WC* не зависит от свободы выбора и способен хорошо работать даже при $F = 1$, что позволяет максимально ограничивать количество перебираемых моделей, тем самым существенно повышать скорость нахождения лучшей модели. Такие качества весового критерия совместно с алгоритмом *BSS* позволяют значительно расширить диапазон решаемых задач.

Проведенные эксперименты подтвердили эффективность этого критерия и его существенные преимущества перед рассмотренными подходами.

1. *Степашко В.С.* Комбинаторный алгоритм МГУА с оптимальной схемой перебора моделей // Автоматика. – 1981. – № 3. – С. 31–36.
2. *Ивахненко А.Г.*, *Степашко В.С.* Помехоустойчивость моделирования. – Киев: Наук. думка, 1985. – 216 с.
3. *Самойленко О.А.*, *Степашко В.С.* Аналіз ефективності застосування частотного критерію в алгоритмі послідовного відсіювання неінформативних аргументів // Індуктивне моделювання складних систем: Зб. наук. праць. – К.: МННЦІТтаС НАНУ, 2012. – С. 191–209.
4. *Problems* of Further Development of GMDH Algorithms. P. 2 / A.G. Ivakhnenko, G.A. Ivakhnenko, E.A. Savchenko et al. // Pat. Recog. and Image Analysis. – 2002. – **12**, N 1. – P. 6–18.
5. *Ивахненко А.Г.*, *Ивахненко Г.А.*, *Савченко Е.А.* Концепция последовательного алгоритмического приближения к точному решению интерполяционных задач искусственного интеллекта // Кибернетика и вычислительная техника. – 2000. – **127**. – С. 47–58.
6. *Ivakhnenko A.G.*, *Ivakhnenko G.A.*, *Savchenko E.A.* GMDH Algorithm for Optimal Model Choice by the External Error Criterion with the Extension of Definition by Model Bias and its Applications to the Committees and Neural Networks // Pat. Recog. and Image Analysis. – 2002. – **12**, N 4. – P. 347–353.

Окончание на стр. 46

7. *Кошулько О.А.*, *Кошулько Г.А.* Багатоетапний комбінаторний алгоритм МГУА для моделювання об'єктів з великою розмірністю // Матеріали міжнар. конф. «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (*ISDMCI*'2009), Євпаторія – Херсон: Вид-во ХНТУ, 2009. – Т. 2. – С. 331–332.
8. *Степашко В.С.*, *Копа Ю.В.* Опыт применения системы АСТРИД для моделирования экономических процессов по статистическим данным // Кибернетика и вычислительная техника. – 1998. – **117**. – С. 24–31.
9. *Samoilenko O.A.*, *Stepashko V.S.* Combinatorial GMDH algorithm with successive selection of arguments // Proc. of the II Int. Workshop on Inductive Modelling IWIM–2007, 19–23 Sept. 2007, Prague, Czech Republic. – Prague: Czech Techn. Univ., 2007. – P. 139–143.
10. *Самойленко О.А.*, *Степашко В.С.* Оптимізація структури моделей методом послідовного відсіювання аргументів / Матеріали міжнар. конф. «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту», Євпаторія, 19–23 травня 2008 р.: У 3 т. – Херсон: Вид-во ХНТУ, 2008. – Т. 3; Ч. 2. – С. 83–86.
11. *Samoilenko O.*, *Stepashko V.* A method of Successive Elimination of Spurious Arguments for Effective Solution of the Search-Based Modelling Tasks // Proc. of the II Int. Conf. on Inductive Modelling ICIM–2008, 15–19 Sept. 2008, Kyiv, Ukraine. – Kyiv: IRTC ITS NANU, 2008. – P. 36–39.

E-mail: *soa0pga@gmail.com*