УДК 519.237+517.518.82

А.Б. Иващенко
Донецкий национальный технический университет, г. Донецк
кафедра компьютерных систем мониторинга

МЕТОДЫ ОТБОРА ПРИЗНАКОВ И ОПТИМИЗАЦИИ СТРУКТУРЫ МОДЕЛЕЙ

Аннотация

Иващенко А.Б. Методы отбора признаков и оптимизации структуры моделей. *Выполнен анализ методов отбора признаков и формирования моделей оптимальной сложности, произведен сравнительный анализ модификации алгоритма SAF с подходами SFS, SFFS и ABFC. Проведен ряд вычислительных экспериментов, выявлены слабые и сильные стороны методов, сформулированы идеи для развития алгоритма SAF.*

Ключевые слова: *синтез аппроксимирующих функций, методы отбора признаков, полиномиальные модели, оптимальная сложность.*

Постановка проблемы. При анализе данных и восстановлении модели зависимости, нередко возникает ряд вопросов, касающихся выбора оптимального набора информативных признаков для описания переменной-отклика и поиска модели с оптимальной структурой.

Проблема выбора подмножества признаков особо остро стоит при анализе объемных многомерных данных, описывающих поведение сложных процессов и систем. Удачное решение этой проблемы способствует как снижению размерности модели, так и повышению ее адекватности.

Анализ литературы. Первые работы по разработке методов отбора признаков были выполнены еще в 70-х годах как в советском союзе, так и за рубежом. С тех пор интерес к этому направлению только увеличивался, особенно в наши дни, когда внимание к интеллектуальному анализу данных и желание постигнуть законы сложных реальных систем особенно повышено. Множество существующих подходов можно разделить на три основные группы: методы исчерпывающего и сокращенного перебора; методы последовательного отбора и методы случайного поиска признаков. Обобщенные особенности этих подходов представлены в табл.1. Большая часть этих методов освещена в работах [1-6].

Первичный обзор существующих методов позволил выделить методы последовательного отбора признаков как наиболее перспективные для дальнейшего исследования (имея в виду удовлетворительную степень достигаемой точности и относительно невысокую вычислительную сложность методов).

Таблица 1 – Характеристики методов отбора признаков

Характеристика\Метод	**Исчерпывающие**	**Рандомизированные**	**Последовательные**
Точность	Всегда находят оптимальное решение (в рамках заданного пространства признаков)	Хорошая при тщательно подобранных контролирующих параметрах	Хорошая, если нет необходимости в возврате в поиске
Сложность	Экспоненциальная $O(2^N)$	В большинстве случаев низкая (при удачных параметрах)	Квадратичная сложность $O(N^2)$
Преимущества	Высокая точность	Разработаны с целью избежать локального минимума	Интуитивно простые и быстрые
Недостатки	Высокая сложность	Сложности в подборе хороших параметров	Нет возврата и удаления «устаревшего» признака после добавления новых

Цель данной работы является сравнение модифицированного алгоритма методики синтеза аппроксимирующих функций (SAF), предложенного в работе [7] с часто используемыми методами отбора признаков и выявление его преимуществ и недостатков.

Постановка и ход исследования. С целью проведения сравнительных экспериментов были отобраны основные и наиболее популярные методы последовательного отбора признаков (Sequential Features Selection): классический метод последовательного прямого отбора признаков (Sequential Forward Selection, SFS), метод последовательного плавающего отбора (Sequential Floating Forward Selection, SFFS) Adaptive Basis Function Construction,ABFC), модифицированный метод Эглайса синтеза аппроксимирующей функции (Synthesis of Approximating Function, SAF). В качестве экспериментальных зависимостей использовались такие трансцендентные функции, как: y=sin(x), y=cos(x), y=tg(x), y=arctg(x),

y=sinh(x), y=cosh(x), y=th(x), y=exp(x), y=ln(1+x), y=1/(1-x) и др. Для каждой функции были сформированы файлы, содержащие равномерные выборки разной длины (N=90, 180, 360, 720, 1440). При конструировании моделей с помощью каждого метода учитывались следующие параметры: p – сложность модели (количество функций в модели); SSR – сумма квадратов ошибок; MeanSSR – среднеквадратичная ошибка, AIC и AICc – информационный критерий Акаике и его скорректированное значение.

Учитывая известный факт, что указанные функции представимы в виде полиномиальных разложений (рядов Маклорена), заранее предполагалось, что именно соответствующие им формулы будут отражены в восстанавливаемых моделях. Сравнение структуры восстанавливаемых моделей с «эталонной» формулой и позволит определить наиболее адекватный метод. Поэтому, кроме прочего, учитывались степени включенных в модель функций и значения коэффициентов при них. Алгоритмы SFS, SFFS и SAF предполагают задание пользователем параметра d (максимальной степени полинома), поэтому варьируя этот параметр, было построено по несколько вариантов моделей для каждого из методов. Алгоритм ABFC позволяет производить автоматические расчеты без участия пользователя, хотя варьирование параметром rd (глубина рекурсии наращивания степеней) в ручном режиме позволило подобрать более успешную модель.

В табл. 2, 3 продемонстрирован фрагмент результатов вычислений, а именно расчет характеристик моделей для функции y=sin(x) при объеме выборки N=90. Из таблиц видно, что модели, восстанавливаемые с помощью SAF, гораздо более точно соответствуют эталонным разложениям, чем модели, произведенные с помощью остальных методов, причем как структурно (по составу включенных в модель одночленов), так и по точности соответствия стоящих при них коэффициентов.

На рис.1 изображен график зависимости критерия AIC от длины выборки (объема исходных данных) для моделей разной сложности (построенных с помощью SAF).

Таблица 2 – Характеристики восстановленных моделей

Метод	p	Включенные степени	SSR	MeanSSR	AIC (AICc)
Объем выборки N=90					
SAF (d=5)	6	0-5	0,001672902	1,86E-05	-968,37 (-967,358)
SAF (d=10)	6	0,1,3,5,7,8	0,000259545	2,88E-06	-1136,08 (-1135,063)
SAF (d=15)	8	0,1,3,5,7,9,11,12	1,14487E-06	1,27E-08	-1620,2 (-1618,425)
SFS (d=5)	6	0-5	0,001672902	1,85878E-05	-968,37 (-967,358)

SFS (d=10)	11	0-10	1,13232E-09	1,25814E-11	-2236,892 (-2233,508)
SFS (d=15)	15	0-11,13-15	7,37156E-17	8,19062E-19	-3718,151 (-3711,665)
SFFS (d=5)	6	0-5	0,001673	1,86E-05	-968,37 (-967,358)
SFFS (d=10)	11	0-10	1,13232E-09	1,25814E-11	-2236,892 (-2233,508)
SFFS (d=15)	14	0-5,7,9-15	5,4599E-18	6,0665E-20	-3954,4 (-3948,8)
ABFC (auto)	2	0,1	17,66212488	0,196245832	-142,555 (-142,417)
ABFC (rd=3)	4	0-3	0,396644721	0,004407164	-480,207 (-479,737)
ABFC (rd=5)	6	0-5	0,001672902	1,85878E-05	-968,37 (-967,358)

Таблица 3 – Структура восстановленных моделей

Эталонная функция $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$	
SAF (d=5)	$y=+0.86679x+0.2827x^2-0.40083x^3+0.088578x^4-0.0056436x^5+0.013425$
SAF (d=10)	$y=+0.99405x-0.16458x^3+0.008127x^5-0.00022046x^7+0.000016563x^8+0.0024383$
SAF (d=15)	$y=+0.99999x-0.16666x^3+0.008332x^5-0.00019839x^7+0.0000027768x^9--0.000000030114x^{11}+0.0000000015653x^{12}+0.0000036662$
SFS (d=5)	$y=+0.866794194618676x+0.282702688910505x^2-0.400831838140753x^3+ +0.0885782506351679x^4-0.0056436173430752 3x^5+0.0134252607778799$
SFS (d=10)	$y=+0.999742635773834x+0.00148376950634983x^2-0.170124488100599x^3++0.00419660966900534x^4+0.00535393932712308x^5+0.00130599367078889x^6--0.00055552208860912x^7+5.82727514515377E-5x^8-1.93339817218185E-6x^9--7.75465683334575E-9x^{10}+8.01365682625226E-6$
SFS (d=15)	$y=+0.999999896292832x+1.10069961498083E-6x^2-0.166671213575435x^3+ +9.80940880712325E-6x^4+0.00832062460517778x^5+1.06810207942766E-5x^6--0.00020449482206753 2x^7+2.39343174683028E-$

	$6x^8$+2.10748423527453E-6x^9+ +1.15661619165532E-7x^{10}- 3.65185373917668E-8x^{11}+3.31995383338057E-10x^{13}- -2.33616330645295E-11x^{14}+5.23116599230388E-13x^{15}+1.40153694760947E-9
SFFS (d=5)	y=+0.866794194618676x+0.282702688910505x^2- 0.400831838140753x^3+ +0.0885782506351679x^4- 0.00564361734307523x^5+0.0134252607778799
SFFS (d=10)	y=+0.999742635773834x+0.00148376950634983x^2- 0.170124488100599x^3+ +0.00419660966900534x^4+0.00535393932712308x^5+0.00130599367078889x^6- -0.00055552208860912x^7+5.82727514515377E-5x^8- 1.93339817218185E-6x^9- -7.75465683334575E-9x^{10}+8.01365682625226E-6
SFFS (d=15)	y=+1.00000000448279x-5.02495992784807E-8x^2- 0.166666501213163x^3- -2.37139498466278E-7x^4+0.00833347682887116x^5- 0.000198449864152707x^7+ +2.77835821738866E-6x^9- 1.72073666435838E-8x^{10}-1.82482997779759E-8x^{11}- -1.67596350926884E-9x^{12}+4.26656223372516E-10x^{13}- 2.61246774829944E-11x^{14}+ +5.47586017539848E-13x^{15}+2.4161025957703E-10
ABFC (auto)	y=-0.30387759984419x+0.944052305982044
ABFC (rd=3)	y=+1.8971425691834x- 0.878353840492117x^2+0.0933780921852978x^3-0.189732711124462
ABFC (rd=5)	y=+0.866794194618676x+0.282702688910505x^2- 0.400831838140753x^3+ +0.0885782506351679x^4- 0.00564361734307523x^5+0.0134252607778799

Анализ результатов исследований. Полученные результаты позволяют достаточно высоко оценить эффективность рассматриваемого метода, как в плане точности модели, так и в смысле способности алгоритма улавливать структуру зависимости в данных, что особенно наглядно представлено на примерах восстановления полиномиальных разложений трансцендентных функций. На рис. 2-7 выборочно представлены модели некоторых функций, восстановленных с помощью алгоритма SAF. Не сложно заметить, что структура аппроксимирующих уравнений восстановленных моделей очень

близка к соответствующим рядам Тейлора-Маклорена, описывающим данные функции.

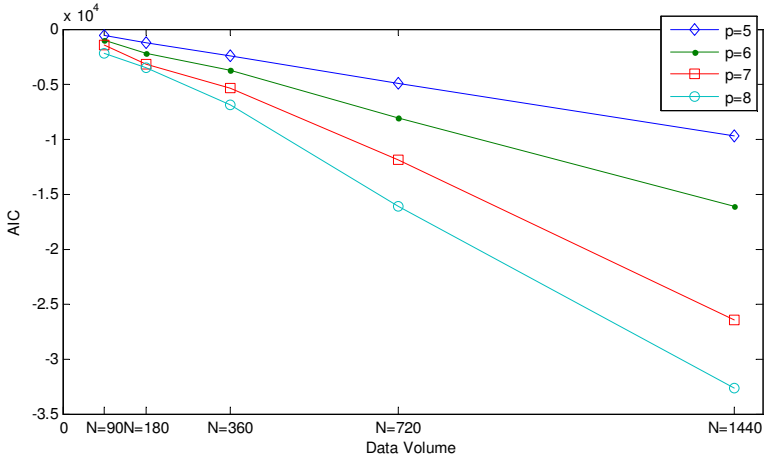

Рисунок 1 – Зависимость AIC модели от объема исходной выборки

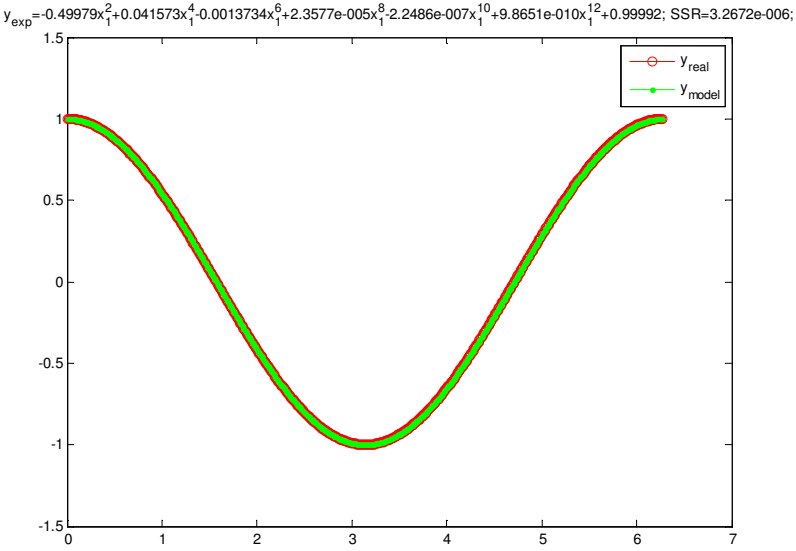

Рисунок 2 – Аппроксимация функции y=cos(x)

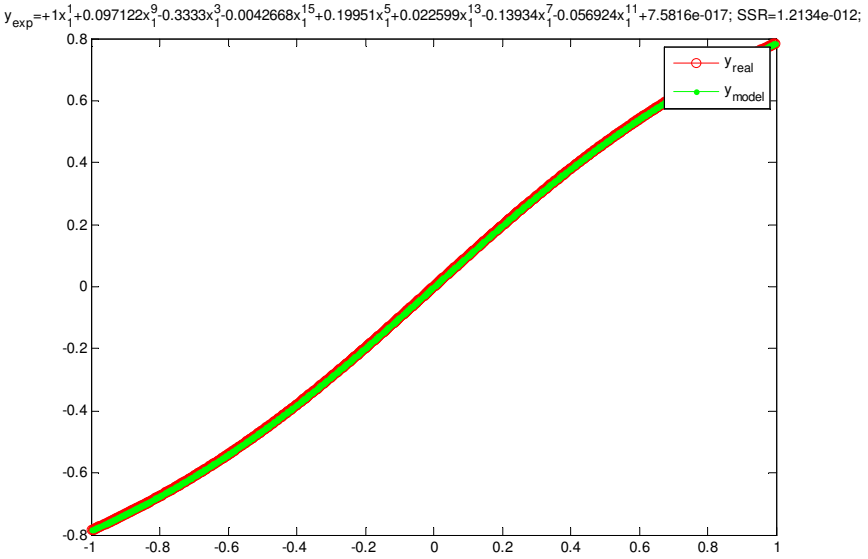

Рисунок 3 – Аппроксимация функции y=arctg(x)

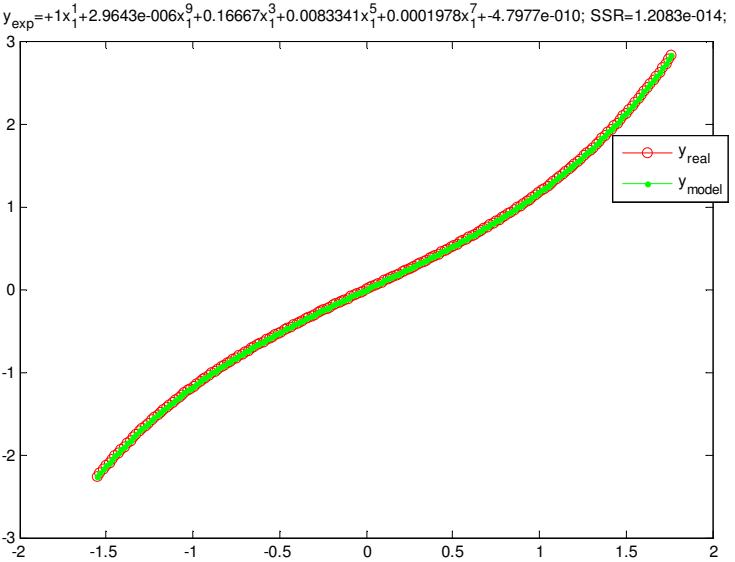

Рисунок 4 – Аппроксимация функции y=sh(x)

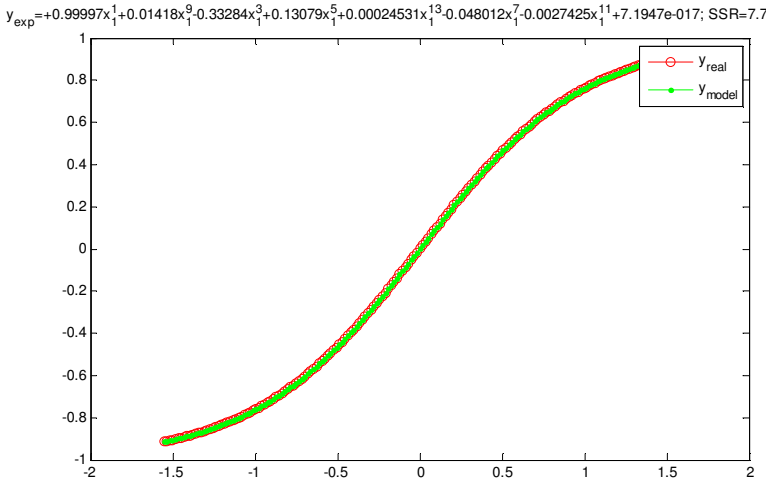

Рисунок 5 – Аппроксимация функции y=th(x)

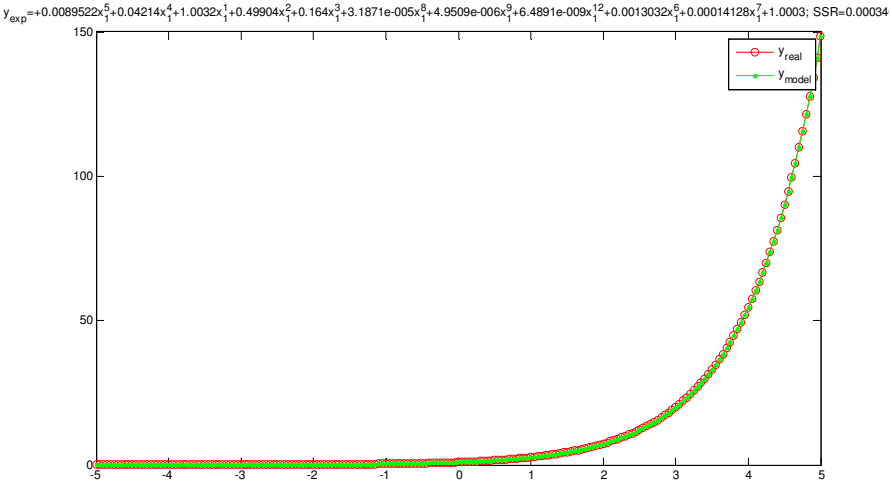

Рисунок 6 – Аппроксимация функции y=exp(x)

В ходе эксплуатации алгоритмов, не смотря на сравнительно высокую точность SAF и весьма обнадеживающие результаты в плане обладания свойством высокой обобщающей способности восстанавливаемой модели, были отмечены некоторые слабые места, в частности, в сравнении с ABFC. Привлекательной особенностью подхода ABFC является отсутствие необходимости задания каких-либо дополнительных параметров для настройки модели – в силу особенностей алгоритма банк функций становится

практически бесконечным, поэтому задание максимальной степени для одночлена или всей аппроксимирующей функции не требуется. Кроме этого, примечательным является использование информационного критерия Акаике как для оценивания моделей, так и в качестве критерия остановки процесса конструирования модели и определения ее оптимальной сложности. Также полезным может оказаться идея использования внешнего критерия для оценки построенных моделей и идея построения более устойчивой модели на основании ансамбля нескольких моделей (хотя часто в ущерб ее точности).

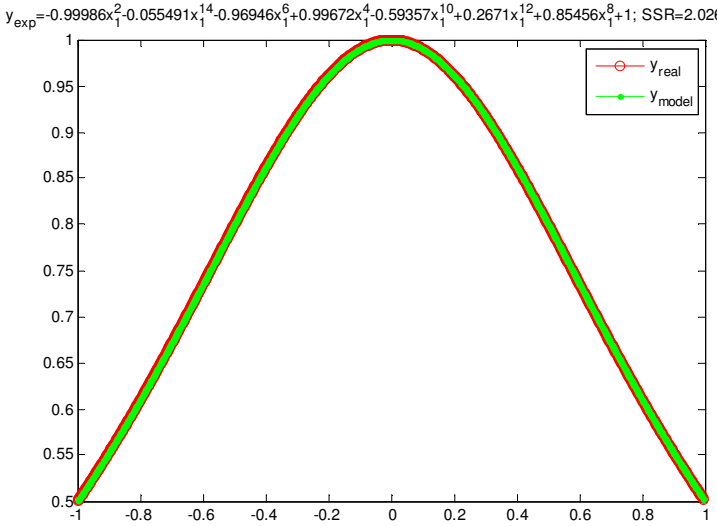

Рисунок 7 – Аппроксимация функции $y=1/(1+x^2)$

Большее число настраиваемых пользователем параметров в алгоритме SAF в сравнении с другими методами является с одной стороны одной из причин эффективности метода, но с другой стороны – основным предметом для дальнейшей оптимизации подхода. Так, в частности, для автоматического определение числа перспективных функций на этапе отбора функций возможно применение критериев останова, используемых в ABFC. Кроме этого, идея использования критерия Акаике для оценки моделей в ABFC была также отмечена положительно и предложена для использования и в алгоритме SAF на этапе элиминации в качестве дополнительного критерия определения оптимальной сложности модели.

Выводы. Выполнен сравнительный анализ методов с точки зрения структуры алгоритмов: стратегии поиска, критерии оценки и выбора моделей, останова и др. Проведены вычислительные эксперименты в части восстановления полиномиальных разложений некоторых трансцендентных функций. Подтверждена высокая эффективность алгоритма SAF в сравнении с другими алгоритмами в плане построения компромиссных моделей (выдерживающих

оптимальный баланс между точностью и сложностью модели). Отмечена привлекательная особенность алгоритма – способность достаточно тонко улавливать структуру зависимостей в данных, что также говорит об его преимуществах и подтверждает целесообразность его дальнейшего использования и развития. С целью повышения автоматизации алгоритма SAF представляется целесообразным использовать критерий Акаике как в качестве критерия-индикатора достаточности отобранных функций на прямом этапе отбора функций, так и для определения точки излома на диаграмме элиминации (при определении оптимальной сложности модели).

Список литературы

1. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

2. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.

3. Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. М.: ВЦ РАН, 2010. 60с. (http://www.machinelearning.ru/wiki/images/5/52/Strijov-Krymova10Model-Selection.pdf)

4. G. Jekabsons. Adaptive Basis Function Construction: an approach for adaptive building of sparse polynomial regression models. Machine Learning, Yagang Zhang (ed.), In-Tech, ISBN: 978-9533070339, 2010, С. 127-156. (http://www.cs.rtu.lv/jekabsons/Files/Jek_ML2010.pdf)

5. P. Somol; J. Novovičová and P. Pudil. "Efficient Feature Subset Selection and Subset Size Optimization". Pattern Recognition Recent Advances, INTECH, ISBN 978-953-7619-90-9. С.75-97. (http://www.intechopen. com/books/pattern-recognition-recent-advances/efficient-feature-subset-selection-and-subset-size-optimization)

6. Ricardo Gutierrez-Osuna. Sequential Feature Selection. Lecture notes. (http://research.cs.tamu.edu/prism/lectures/pr/pr_l11.pdf)

7. Иващенко А.Б., Беловодский В.Н. Некоторые вариации метода Эглайса синтеза аппроксимирующих функций // Научные труды ДонНТУ. Серия: «Проблемы моделирования и автоматизации проектирования динамических систем» (МАП-2011). Выпуск 10 (197) – Донецк: ДонНТУ, – 2011. – С 84-100.