Интернет-журнал «Науковедение» ISSN 2223-5167 http://naukovedenie.ru/

Том 9, №3 (2017) http://naukovedenie.ru/vol9-3.php

URL статьи: http://naukovedenie.ru/PDF/100TVN317.pdf

Статья опубликована 06.07.2017

Ссылка для цитирования этой статьи:
Горбатков С.А., Фархиева С.А. Системный подход к агрегированию экзогенных и эндогенных переменных в нейросетевых моделях банкротств на основе функций Харрингтона // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 9, №3 (2017) http://naukovedenie.ru/PDF/100TVN317.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

УДК 378.675

Горбатков Станислав Анатольевич
ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации»
Филиал в г. Уфа, Россия, Уфа[1]
Профессор кафедры «Математика и информатика»
Доктор технических наук
E-mail: sgorbatkov@mail.ru
РИНЦ: http://elibrary.ru/author_items.asp?id=158740
SCOPUS: https://www.scopus.com/authid/detail.uri?authorId=8646868800

Фархиева Светлана Анатольевна
ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации»
Филиал в г. Уфа, Россия, Уфа
Заведующий кафедрой «Математика и информатика»
Кандидат технических наук
E-mail: ok-xi@yandex.ru
РИНЦ: https://elibrary.ru/author_items.asp?id=567037

Системный подход к агрегированию экзогенных и эндогенных переменных в нейросетевых моделях банкротств на основе функций Харрингтона

Аннотация. В статье на основе системного подхода рассмотрена проблема компрессии (агрегирования) экзогенных и эндогенных переменных в обратных задачах восстановления зависимостей, скрытых в данных. Эти задачи относятся к классу некорректно поставленных по Адамару и поэтому требуют регуляризации решения. В статье впервые предложен концептуальный базис агрегирования факторов в виде обобщенной функции желательности Харрингтона. Основной эмерджентный эффект (идея) данного подхода состоит в том, что образуемый агрегат учитывает нелинейное взаимодействие факторов друг с другом, сохраняя при этом возможность оценки вклада каждого фактора в модели интерпретации данных. Теоретико-практическая значимость предлагаемого подхода и реализующего его метода построения модели состоит в обеспечении работоспособности и хорошего качества модели в условиях высокого уровня неопределенности (наличия триады «НЕ-факторов»: неполноты, неточности и неопределенности в данных). Основная идея апробирована на реальных данных банкротств корпораций строительной отрасли экономики. Впервые выявлен эффект повышения качества нейросетевой модели банкротств за счет совместного влияния агрегирования переменных и учета нелинейного влияния друг на друга агрегируемых показателей в составе образуемого агрегата.

[1] 450015, г. Уфа, ул. Мустая Карима 69/1

Ключевые слова: кредитные системы; нейросетевое моделирование; агрегирование переменных; триада «НЕ-факторов»; обобщенная функция желательности Харрингтона; системный закон управления энтропией комплексной информационной системы; регуляризация обратных задач интерпретации данных

1. Введение. Постановка задачи моделирования и разработка подхода к ее решению

Пусть рассматривается обратная задача восстановления закономерностей, скрытых в данных:

$$y_u = \varphi_u(\vec{x}), u = \overline{1,m};\ \vec{x} = \left(x_0, x_1, \dots, x_j, \dots, x_n\right). \quad (1)$$

Здесь: y_u – восстановленное (расчетное) значение u-ой эндогенной переменной; $\vec{x}$ – вектор экзогенных переменных (факторов); $\varphi_u(\cdot)$ – оператор восстанавливаемой нелинейной зависимости (модель системы) в ансамбле моделей; $x_j \epsilon R^n, y_u \epsilon R^m$ – пространства вещественных чисел.

Решение некорректной обратной задачи (1) затрудняется сложными условиями моделирования, характерными для практики. Это так называемая триада «НЕ-факторов» (неполнота, неопределенность, неточность в данных), отягченная отсутствием априорных сведений о виде закона распределения шумов в данных.

В данной статье предложен общий подход (концептуальный базис) агрегирования переменных при разработке моделей вида (1) в сложных условиях моделирования, в частности нейросетевых моделей, а также оригинальный метод, реализующий этот концептуальный базис. Концепции (см. ниже) в своей методологической основе опираются на общесистемный закон управления энтропией при агрегировании отдельных подсистем в общую систему и рациональном их взаимодействии для достижения общей цели [1].

Согласно этому закону при объединении двух изолированных (не взаимодействующих друг с другом) систем A_1 и A_2 в одну общую систему (A_1,A_2) энтропия объединенной системы Э как мера ее неупорядоченности уменьшается, т.е. будет меньше суммы энтропии исходных изолированных систем A_1 и A_2:

$$Э(A_1, A_2) < [Э(A_1) + Э(A_2)]. \quad (2)$$

Энтропия Э системы, имеющей дискретное множество допустимых состояний равна [2]:

$$Э = -\sum_{k=1}^{n} p_k \log_2\left(\frac{1}{p_k}\right), \quad (3)$$

где: p_k – вероятность k-го состояния системы; n – число возможных (с $p_k \neq 0$) состояний.

Соотношение (2) отражает этот факт, что если системы (или подсистемы) A_1 и A_2 взаимодействуют, то в объединенной системе (A_1,A_2) появляются новые связи, которые неизбежно ограничивают число возможных состояний системы (A_1,A_2) и, соответственно, уменьшают ее энтропию. Тогда из (2) следует:

$$Э(A_1, A_2) < [Э(A_1) + Э(A_2)];\ => \Delta I_S > 0. \quad (4)$$

ΔI_S – это уменьшение энтропии, которое можно интерпретировать как приращение структурной информации (меру «негэнтропии»), характеризующей упорядоченность структуры системы и, соответственно, появление новых знаний о объединенной системе (A_1,A_2).

В [2] приводится наглядный пример, иллюстрирующий закон (4): энтропия порции информации, передаваемой по телекоммуникационному каналу связи в виде связного текста, всегда, меньше чем энтропия приходящаяся, на один символ, умноженная на количество символов в несвязном тексте. Дело в том, что в связном тексте появляются дополнительные лингвистические связи, ограничивающие число возможных состояний системы p_k в формуле (2).

В рассматриваемой математико-информационной системе изолированными «подсистемами» являются частные желательности $\{d_u\}, u = \overline{1,n}$ [3], а также модели-гипотезы банкротств в байесовском ансамбле. Под объединением подсистем на первом уровне понимается агрегирование частных желательностей в одну общую систему – агрегат $D(d_1, \dots, d_n)$, который называется «обобщенной функцией желательности Харрингтона». Второй уровень объединения подсистем – это апостериорная фильтрация моделей байесовского ансамбля по качеству (прогностическим свойствам) и затем осреднение выходных характеристик на отфильтрованном ансамбле (см. ниже).

Замечание. Предлагаемые концепции и метод агрегирования переменных в данной статье иллюстрируются количественными оценками для сложных задач оценки риска банкротств при сопровождении банком своего кредитного портфеля. Однако авторы статьи не видят принципиальных ограничений для применения предлагаемого подхода к агрегированию переменных в других областях знаний: налогового контроля, экологии, медицине, педагогике и др.

2. Актуальность темы исследования

Широко известен факт, что удачный выбор переменных, как экзогенных, так и эндогенных, предопределяет качество разрабатываемой модели, особенно в условиях упомянутой выше триады «НЕ-факторов» [4]. В частности, в [4] разработаны на эвристическом уровне рекомендации по отбору и предварительной обработке исходных данных, которые учитывают изменчивость моделируемых объектов, область допустимых параметров модели, сбалансированность набора данных, репрезентативность объема выборки данных. Даются рекомендации по масштабированию (нормировке) экзогенных и эндогенных переменных, масштабированию случайных сигналов.

Тем не менее, в области экономики до сих пор не решена проблема оптимального выбора системы факторных признаков для задач (1) восстановления зависимостей, скрытых в данных. Известно несколько сот работ по моделированию банкротств. Однако авторы в них либо не описывают формализацию выбора переменных и их компрессию, либо вообще оставляют этот вопрос «за кадром». Обзор этих работ приведен в [1, 5].

В области нейросетевого моделирования следует особо отметить работы J. Rissanen [6] и российского ученого С.А. Шумского [7], где получен теоретически обоснованный фундаментальный результат, базирующийся на принципе минимальной длины описания (Minimum Description Length) в выбранной модели-гипотезе из байесовского ансамбля моделей h, который выражается через числовую меру Куллбака-Лейблера:

$$|P(D|h) - P(D|h_0)| = \sum_D P(D|h_0) \log \frac{P(D|h_0)}{P(D|h)} = \sum_D P(D|h_0)[L(D|h) - L(D|h_0)] \geq 0. \qquad (5)$$

Здесь эмпирический риск $L(D|h) = -\log P(D|h)$ аддитивен, т.е. выражается суммой по числу примеров в данных D и пропорционален эмпирической ошибке. Усредненный по бесконечному набору выборок $(D \to \infty)$ ожидаемый риск $\sum_D P(D|h_0)\, L(D|h)$ соответствует

ошибке обобщения E разрабатываемой модели-гипотезы h. К сожалению, для конечного числа примеров в данных D ожидаемый риск – ненаблюдаемая величина!

В указанных работах [6, 7] введена новая измеримая величина – регуляризированный риск, которая ведет себя аналогично ненаблюдаемому риску:

$$L(D,h) = -\log P(D,h) = -\log P(D|h) - \log P(h), \quad (6)$$

где: $P(h)$ – априорная вероятность самой модели h в байесовском ансамбле моделей-гипотез. При этом максимизируется функция правдоподобия $L(D,h)$ в (6), т.е. совместная вероятность данных D и гипотезы-модели h в ансамбле.

Эмпирический риск $L(D,h) = -\log P(D,h)$ согласно подходу [6, 7] по (6) можно трактовать как длину оптимального кодирования данных с помощью модели-гипотезы h, например в нейросети с помощью параметров (синаптических весов) W. $L(h) = -\log P(h)$ – это длина кодирования самой этой гипотезы-модели.

Таким образом, регуляризируемый ожидаемый риск представляет собой суммарную длину описания данных D и гипотезы-модели h:

$$L(D,h) = L(D|h) + L(h). \quad (7)$$

Это означает, что суммарная длина описания (6) и определяет прогностическую силу модели h, ограничивая сверху ожидаемый риск.

Подход [6, 7] имеет одно существенное ограничение: он базируется на допущении об априорном знании вида закона распределения шумов в данных. В нашей статье и цитируемой монографии авторов [1], с целью приближения моделей банкротств к практике, мы отказались от этого допущения. В [1] использован байесовский подход к оптимальному выбору системы экзогенных переменных (n порядка 16 … 22). В ансамбле метагипотез $\{H_q\}, q = 1,2,3,4$ сравнивались 4 системы показателей: А.О. Недосекина, Г.З. Рахимкуловой, О.П. Зайцевой, Э. Альтмана, после чего использован случайный выбор всех экзогенных переменных внутри выбранной оптимальной системы H_m^* методом последовательного включения факторов по критерию качества распознавания банкротства – взвешенной сумме правильно идентифицированных объектов и ошибок I и II рода:

$$K_{H_q} = \left(N_{H_q}^* \cdot r_1 - N_{H_q}^{(1)} \cdot r_2 - N_{H_q}^{(2)} \cdot r_3\right) \to max, \quad (8)$$

где: $N_{H_q}^*$ – число верно идентифицированных объектов на тестовом множестве; r_1, r_2, r_3 – весовые множители (коэффициенты Фишберна). В итоге была выбрана система показателей А.О. Недосекина [8]. Число верно идентифицированных объектов 91,66% на тестовом множестве из 36 точек; число ошибок I рода равно 1, число ошибок II рода равно 2).

Однако работа [1] не дает теоретически обоснованного решения поставленной проблемы: сравнивались лишь небольшое число систем показателей – 4 широко распространенные «системы показателей» (см. выше), а оптимальное решение задачи сокращения пространства переменных

$$R = [Y_u \in R^m] \times [X_j \in R^n] \quad (9)$$

может лежать вне этого множества R либо состоять из части (подмножества) R.

Отсюда вытекает актуальность поставленной в статье проблемы компрессии переменных в моделях сложных систем [9].

3. Идея предлагаемых концепций компрессии переменных

Концепция 1. При агрегировании эндогенных и экзогенных переменных в моделях (1) восстановления закономерностей, скрытых в данных, при неизвестном априори виде закона распределения шумов, наличии триады «НЕ-факторов» и большой размерности пространства R переменных следует учитывать взаимное влияние показателей друг на друга. Этот учет должен быть нелинейным для приближения моделей к практике.

Концепция 2. Модель восстановления закономерностей вида (1) в терминах агрегированных переменных

$$Y_{ag}^{(q)} = f\left(\vec{X}_{ag}^{(k)}, W\right); k = 1,2, \dots, l; q = 1,2, \dots, Q \quad (10)$$

должна быть регуляризирована в силу наличия неопределенности (триады «НЕ-факторов»). Здесь k – номер группы экзогенных переменных; q – номер модели в байесовском ансамбле моделей. В качестве механизма регуляризации возможен следующий двухэтапный алгоритм:

1. Предрегуляризация – сокращение пространства R за счет получения агрегатов в (10) в виде обобщенных функций желательности эндогенных $Y_{ag}^{(q)}$ и экзогенных $\vec{X}_{ag}^{(k)}$ переменных для каждой q-ой модели ансамбля в R^m и R^n соответственно. При этом агрегат Y_{ag} – это единый для всех кластеров $k = \overline{1,l}$ экзогенных переменных, выделенных для удобства интерпретации результатов моделирования.

2. Осреднение выходных характеристик моделей на байесовском ансамбле $\{h_q\}$ моделей вида (10), что регуляризирует модель и повышает достоверность оценок. При этом достоверность оценки на всем отфильтрованном ансамбле Q^* оказывается лучше, чем в лучшей q-ой модели из ансамбля согласно концепции [6, 7].

4. Метод агрегирования переменных для реализации концепций 1 и 2

Рассмотрим теперь конкретно двухступенчатый метод агрегирования переменных (ДМАП), реализующий предложенные концепции 1 и 2:

а. Создаются компактные выражения для единого агрегата X_{ag} экзогенных переменных в виде геометрической средней – обобщенной функции желательности Харрингтона в каждой q-ой модели ансамбля:

$$D = \sqrt[n]{d_1 \cdot d_2 \cdot \dots \cdot d_j \cdot \dots \cdot d_n} \in [0; 1]. \quad (11)$$

Здесь учет нелинейности взаимного влияния экзогенных показателей друг на друга обеспечивается операциями перемножения частных функций желательности $\{d_j\}, j = \overline{1,n}$ друг на друга и последующим извлечением корня n-ой степени из полученного произведения.

Подчеркнем, что компрессия вида (11), учитывая нелинейное «взаимодействие» факторов $\{d_j\}, j = \overline{1,n}$, не уничтожает информацию об индивидуальном вкладе каждой функции желательности d_j в обобщенную функцию желательности D (см. ниже в разделе 5 статьи).

б. Трансформация физических величин $\{X_{ij}\}$ в функции желательности $\{d_{ij}\}$ производится экспертно как для эндогенных, так и для экзогенных переменных:

$$d_{i.j} = exp[-exp(-y')]; i = \overline{1,N}; j = \overline{1,n}. \quad (12)$$

где: i – номер наблюдения; j – номер показателя (компоненты векторов $\vec{X}$ либо $\vec{Y}$); y' – кодированная шкала по оси абсцисс, характеризующая чувствительность функции желательности.

Отметим, что шкала измерения каждого показателя x_{ij} – равномерная, а шкала для $\{d_{i.j}\}$ в силу нелинейного преобразования (12) – нелинейная по оси ординат (рис. 1). Трансформацию (12) удобно проводить экспертно с использованием стандартной таблицы 1 – аналога «пенташкалы» в методах нечетких множеств [1].

Заметим, что алгоритм (12) – это «экспертное вмешательство» в построение модели (10) по ДМАП, полезное для реализации профессионального опыта аналитика в предметной области.

Второе экспертное вмешательство, связанное с управлением адекватностью модели (10) – это выбор класса моделей-гипотез $\{h_q\} \in H$ в байесовским ансамбле.

Таблица 1

Экспертно задаваемая связь лингвистических оценок со значением частных функций желательности (таблица заимствована из [3])

Желательность (лингвистическая оценка)	Отметка на шкале желательности d_u
Очень хорошо	1,00-0,8
Хорошо	0,8-0,63
Удовлетворительно	0,63-0,37
Плохо	0,37-0,2
Очень плохо	0,2-0

Замечание. Группировку факторов в кластеры (*k=1, 2, ... , l*) удобно проводить экспертно, руководствуясь экономическими соображениями. В статье выделено четыре кластера: R – «рентабельность»; L – «ликвидность»; A – «деловая активность»; F – «платежеспособность».

В каждый кластер входило по 4 показателя. Таким образом, исходное число факторов равнялось 16 (система показателей А.О. Недосекина [8]). В итоге предварительной фильтрации факторов по критерию их информативности было оставлено 8 факторов (см. таблицу 2). Затем из них были образованы 4 агрегата в виде функций Харрингтона D_R, D_L, D_A и D_F.

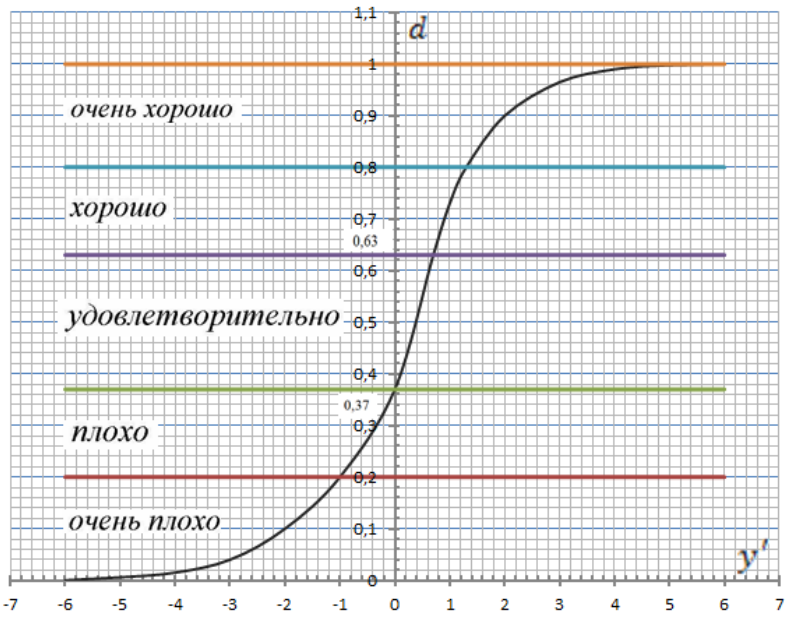

Рисунок 1. *Функция желательности (рисунок заимствован из [3])*

Таблица 2

Частные функции желательности по группам факторов и функции Харрингтона (таблица получена авторами статьи)

№ п/п	Частные функции желательности $\{d_j\}, j = \overline{1,n}$								Обобщенные функции желательности					Вероятность банкротства, P	Y(x)
	dR3	dR5	dL2	dA2	dA4	dA5	dA6	dF3	D_R	D_L	D_A	D_F	D		
1	0,5	0,38	0,63	1	1	1	1	1	0,436	0,63	1	1	0,7239	1	6
2	0,45	0,65	0,27	0,28	0,35	0,9	0,88	1	0,5408	0,27	0,5278	1	0,5268	0	-6
3	0,65	0,3	0,1	0,9	0,1	0,72	0,9	0,9	0,4416	0,1	0,4914	0,9	0,4211	1	6
4	0,55	0,38	0,1	0,2	0,83	0,23	1	1	0,4572	0,1	0,442	1	0,377	0	-6
5	0,6	0,37	0,1	1	0,86	1	0,5	0,72	0,4712	0,1	0,809	0,72	0,407	1	6
6	0,4	0,4	0,73	1	1	0,85	0,93	1	0,4	0,73	0,9423	1	0,7242	0	-6
7	0,42	0,375	0,1	1	0,8	0,5	1	1	0,397	0,1	0,7952	1	0,4215	1	6
8	0,38	0,39	0,25	1	1	1	0,9	0,9	0,385	0,25	0,974	0,9	0,5389	0	-6
9	0,01	0,37	0,64	1	1	0,8	0,45	0,85	0,0608	0,64	0,7746	0,85	0,4	1	6
10	0,58	0,39	0,29	1	1	1	1	1	0,4756	0,29	1	1	0,6094	0	-6
11	0,39	0,375	0,71	0,55	0,81	1	0,1	1	0,3824	0,71	0,4594	1	0,626	1	6
12	0,4	0,37	0,3	0,71	0,95	1	0,1	0,72	0,3847	0,3	0,5096	0,72	0,4536	0	-6
13	0,41	0,378	0,25	1	0,5	0,4	1	1	0,3937	0,25	0,6687	1	0,5065	1	6
14	0,6	0,375	0,1	1	0,53	0,37	0,85	1	0,4743	0,1	0,639	1	0,4172	0	-6
15	0,37	0,38	0,18	1	0,73	0,7	1	1	0,3749	0,18	0,8455	1	0,4887	1	6
16	0,38	0,379	0,26	1	1	1	0,8	0,9	0,3795	0,26	0,9457	0,9	0,5383	0	-6
17	0,37	0,371	0,1	0,63	0,25	0,5	0,1	0,71	0,3704	0,1	0,2979	0,71	0,6393	1	6
18	0,48	0,379	0,25	1	0,9	0,9	0,9	1	0,4265	0,25	0,924	1	0,5602	0	-6
19	0,495	0,01	0,18	1	1	1	1	1	0,0703	0,18	1	1	0,335	1	6
20	0,39	0,378	0,17	1	0,83	0,98	0,3	0,75	0,3839	0,17	0,7028	0,75	0,4307	0	-6
21	0,31	0,371	0,1	1	0,72	0,9	0,3	0,76	0,3391	0,1	0,664	0,76	0,3617	1	6
22	0,43	0,38	0,5	1	1	1	1	1	0,4042	0,5	1	1	0,6704	0	-6
23	0,35	0,37	0,12	1	0,3	0,92	0,23	0,71	0,3599	0,12	0,5019	0,71	0,3522	1	6
24	0,38	0,37	0,1	0,95	1	1	1	1	0,375	0,1	0,9872	1	0,4386	0	-6
25	0,41	0,379	0,1	1	0,9	0,9	0,3	0,72	0,3942	0,1	0,7021	0,72	0,3757	1	6
26	0,39	0,45	0,4	0,72	0,45	0,94	0,72	1	0,4189	0,4	0,6843	1	0,5819	0	-6
27	0,45	0,39	0,23	1	0,9	1	0,73	0,9	0,4189	0,23	0,9003	0,9	0,5286	1	6
28	0,5	0,37	0,1	1	0,8	0,75	0,71	0,97	0,4381	0,1	0,8079	0,97	0,4304	0	-6
29	0,37	0,371	0,38	1	0,68	0,95	0,98	1	0,3705	0,38	0,892	1	0,5953	1	6
30	0,37	0,378	1	1	1	1	0,8	0,95	0,374	1	0,9457	0,95	0,761	0	-6
31	0,72	0,37	0,17	0,95	0,5	0,91	0,51	0,83	0,5161	0,17	0,6852	0,83	0,4726	1	6
32	0,46	0,39	0,17	1	0,97	0,85	1	1	0,4235	0,17	0,953	1	0,5118	0	-6
33	0,39	0,37	0,16	1	0,9	0,91	1	1	0,38	0,16	0,9513	1	0,4904	1	6

5. Количественные оценки

Использовалась база данных строительной отрасли из работы [10]. Вероятность риска банкротства оценивалась по логистической модели:

$$P = 1/[1 + \exp(-Y(D_u))], P \in [0; 1]; U = \overline{1,4}. \quad (13)$$

В таблице 2 приведены результаты агрегирования экзогенных переменных, в таблицах 3, 4 – результаты построения нейросетевых и регрессионной моделей байесовского ансамбля.

Критерий качества моделей при фильтрации байесовского ансамбля из табл. 3 – это число верно идентифицируемых точек на тестовом множестве («банкрот-небанкрот»), которых модель «не видела» при обучении, т.е. построении моделей 1, 2 и 3 из табл. 3. Другим словами, точки тестового множества $\Omega_{test} = 0{,}2 \cdot 99$ не участвовали при построении моделей 1, 2 и 3, но потом предъявлялись для идентификации.

Фильтр $\frac{N^*}{N} \to max \leq \omega$, где N_{test} – общее число наблюдений в тестовом множестве нейросети; N^* – число верно идентифицированных точек тестового множества нейросети.

Таблица 3

Характеристики логистических моделей банкротства в байесовском ансамбле (таблица составлена авторами статьи)

№ модели	Наименование модели	Учет нелинейной зависимости эндогенной переменной Y от вектора факторов $\vec{x}$ в (1)	Учет нелинейного влияния факторов друг на друга в составе агрегата $X_{ag}^{(k)}$	Примечание
1	Классическая (базовая) эконометрическая модель множественного линейного уравнения регрессии для показателя экспоненты $Y(\vec{x})$ в (13) (рассчитывается в MS Excel по исходным (сырым) данным ($N = 3 \times 33 = 99$) наблюдений с учетом «бустреп-процедуры»)	Отсутствует	Отсутствует	Есть возможность обобщения модели введением вторых и более высоких степеней факторов и их произведений, однако это затруднено ростом «длины описания модели» при числе факторов 10 и более
2	Нейросетевая модель с включением всех исходных 8 факторов для показателя экспоненты $Y(\vec{x})$ в логистической модели (N=99)	Имеется	Отсутствует	Устраняется проблема мультиколлинеарности факторов в уравнении $Y(\vec{x})$ показателя экспоненты в (13)
3	Нейросетевая модель с агрегированием факторов в виде обобщенных функций желательности Харрингтона по 4-м группам (кластерам) факторов для показателя экспоненты $Y(D_u), u = \overline{1,4}$ ($N = 99$)	Имеется	Имеется	Существенно уменьшается «длина описания» данных и самой модели по числовой мере Куллбака-Лейблера, повышаются прогностические свойства нейросетевой модели

Таблица 4

Характеристика моделей байесовского ансамбля по вероятности верной идентификации на тестовом множестве (таблица получена авторами статьи)

Номер модели в таблице 3	N^*	N_{test}	N^*/N_{test}
1	11	20	0,55
2	14		0,70
3	**18**		**0,9**

Из таблицы 4 видно, что классическая модель 1 множественного уравнения регрессии практически неработоспособна при сильно зашумленных (реальных) данных.

Для лучшей модели № 3 (нейросетевая модель с агрегированием факторов в виде обобщенных функций желательности Харрингтона по 4-м группам) приведены результаты расчета (табл. 4 и рис. 2), полученные с помощью программы NeuroSolutions (демоверсия). Множественный коэффициент детерминации в модели составил 0,9, что вполне допустимо для сложных условий моделирования.

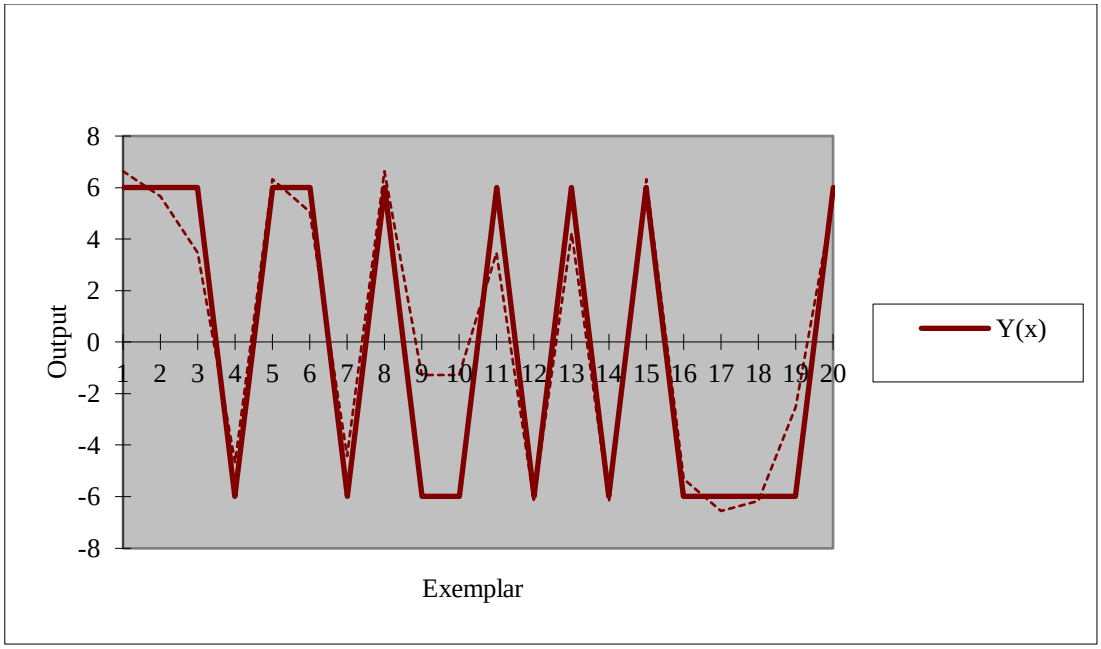

Рисунок 2. *Графическое представление фиксированных и рассчитанных с помощью нейросети показателей экспоненты $Y(D_u)$ логистической функции модели № 3 (Exemplar – номер i наблюдения в тестовом множестве Ω_{test}, Output – значение функции $Y_i(D_u)$, рассчитанное нейросетью; Y(x) – экспериментальное значение, Y(x) Output – расчётное значение; рисунок получен авторами)*

На рисунке 2 показано сравнение рассчитанных в нейросети Y и «экспериментальных» $Y(\vec{x})$ значений показателя экспоненты $Y(D_u)$ в (13). Видно, что кроме точек i=9 и i=10 все точки тестового множества идентифицированы верно.

Выводы

1. Эмерджентный эффект, получаемый за счет предложенного метода агрегирования переменных (ДМАП) в виде обобщенных функций желательности Харрингтона, проявился довольно четко: число верно идентифицированных точек на тестовом множестве в модели 3 (с агрегированием по 4-м кластерам факторов) из табл. 4 возросло на 25% по сравнению с моделью № 2 без агрегирования факторов.

2. Важным обстоятельством апробации предложенных теоретических идей является то, что апробация проводилась в сложных условиях моделирования (имела место «триада НЕ-факторов») на реальны данных строительной отрасли [10], которая характеризуется высоким уровнем зашумленности данных.

3. В качестве направления дальнейших исследований можно указать введение эвристических операций предобработки данных в направлении выявления и исключения из данных обучающего множества противоречивых наблюдений, а также автоматизацию процедур агрегирования эндогенных и экзогенных переменных после соответствующих экспертных оценок по таблице 1.

ЛИТЕРАТУРА

1. Моделирование управленческих решений в сфере экономики в условиях неопределенности: Монография / И.И. Белолипцев, С.А. Горбатков, А.Н. Романов, С.А. Фархиева; ред. А.Н. Романов. – М.: ИНФРА-М, 2015. – 299 с.
2. Вентцель Е.С. Теория вероятностей: Учебник для вузов / Венцель Е.С.; 4-е изд. стереотипное. – М.: Наука: Физмалит, 1969. – 576 с.
3. Адлер Ю.П. Планирование эксперимента при поиске оптимальных условий: Монография / Ю.П. Адлер, Е.В. Маркова, Ю.В. Грановский; изд. 2-е, перераб. и доп. – М.: Наука, 1976. – 279 с.
4. Матвеев М.Г. Модели и методы искусственного интеллекта. Применение в экономике: Учебное пособие / М.Г. Матвеев, А.С. Свиридов, Н.А. Алейнокова. – М.: Финансы и статистика; ИНФРА-М, 2008. – 448 с.
5. Горбатков С.А., Белолипцев И.И., Макеева Е.Ю. О моделях диагностики банкротств организаций // Менеджмент и бизнес-администрирование. – 2014. – №1. – С. 151-172.
6. Rissanen J. Modeling bu shortest data description // Automatica. – 1978. Vol.14. P. 465-471.
7. Шумский С.А. Байесова регуляризация обучения // Лекции школы-семинара «Современные проблемы нейроинформатики» (23-25 января 2002 г., Москва). – М.: МИФИ, 2002. – С. 61-94.
8. Шевченко И.В. Создание виртуальной клиентской баз для анализа кредитоспособности российских предприятий / И.В. Шевченко, А.А. Халафян, Е.Ю. Васильева // Финансы и кредит. – 2010. – №1(385). – С. 13-18.
9. Тарков М.С. Понижение размерности пространства данных в задаче диагностирования заболевания щитовидной железы / М.С. Тарков, М.А. Чиглинцев // XIV Всероссийская науч.-техн. конференция «Нейроинформатика2012»; Сборник научных трудов. – М.: НИЯУ МИФИ. – 2012. – Часть 3. – С. 142-150.
10. Makeeva E.U., Neretina E.A. Binary model versus diskriminant analysis relating to corporate bankruptcies case of Russian Construction Industry // Journal of Accounting, Finance and Economics. – 2013. – Vol. 3. – №1. – P. 65-76.

Gorbatkov Stanislav Anatol'evich
Financial university under the government of the Russian Federation
Ufa branch, Russia, Ufa
E-mail: sgorbatkov@mail.ru

Farkhieva Svetlana Anatol'evna
Financial university under the government of the Russian Federation
Ufa branch, Russia, Ufa
E-mail: ok-xi@yandex.ru

The systems concept to aggregation of exogenous and endogenous variables in neural network models of bankruptcies on the basis of Harrington's functions

Abstract. In article on the basis of the systems concept the problem of a compression (aggregation) of exogenous and endogenous variables in the reverse tasks of restoration of the dependences hidden in data is considered. These tasks belong to the class incorrectly delivered on the Hadamard and therefore require regularization of the decision. In article the conceptual base of aggregation of factors in the form of the generalized function of desirability of Harrington is for the first time offered. The main emerdzhentny effect (idea) of this approach consists that the formed aggregate considers non-linear interaction of factors with each other, saving at the same time a possibility of an assessment of a contribution of each factor to models of interpretation of data. The Teoretiko-prakticheskу significance of the offered approach and the method of creation of model realizing it consists in support of working capacity and high quality model in the conditions of the high level of uncertainty (existence of a triad "NOT – factors": incompleteness, inaccuracy and uncertainty in data). The main idea is approved on real data of bankruptcies of corporations of construction branch of economy. The effect of improvement of quality of neural network model of bankruptcies due to joint influence of aggregation of variables and the accounting of non-linear influence at each other of the aggregated indices as a part of the formed aggregate is for the first time revealed.

Keywords: credit systems; neural network simulation; aggregation of variables; triad "NOT – factors"; the generalized function of desirability of Harrington; the system law of control of an entropy of an end-to-end information system; regularization of the reverse tasks of interpretation of data