Вестник Евразийской науки / The Eurasian Scientific Journal https://esj.today

2018, №1, Том 10 / 2018, No 1, Vol 10 https://esj.today/issue-1-2018.html

URL статьи: https://esj.today/PDF/56ITVN118.pdf

Статья поступила в редакцию 19.02.2018; опубликована 12.04.2018

**Ссылка для цитирования этой статьи:**
Горбатков С.А., Фархиева С.А., Горбаткова Е.Ю. Метод агрегирования переменных нейросетевой модели в обратных задачах восстановления зависимости в условиях высокой размерности пространства признаков и зашумленности данных // Вестник Евразийской науки, 2018 №1, https://esj.today/PDF/56ITVN118.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

**For citation:**
Gorbatkov S.A., Farkhieva S.A., Gorbatkova E.Yu. (2018) The method of aggregating the variables of a neural network model in inverse problems of recovering the dependence in conditions of high dimensionality of the space of attributes and data noisiness. *The Eurasian Scientific Journal*, Vol. 10, No. 1. https://esj.today/PDF/56ITVN118.pdf (In Russ.)

**УДК 378.675**

**ГРНТИ 28.23.24**

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

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

**Горбаткова Елена Юрьевна**
ФГОБУ ВО «Башкирский государственный педагогический университет им. М. Акмуллы», Уфа, Россия
Доцент кафедры «Охрана здоровья и безопасности жизнедеятельности»
Кандидат педагогических наук
E-mail: gorbatkovaeu@mail.ru
РИНЦ: http://elibrary.ru/author_profile.asp?id=300637

# Метод агрегирования переменных нейросетевой модели в обратных задачах восстановления зависимости в условиях высокой размерности пространства признаков и зашумленности данных

**Аннотация.** Статья посвящена совершенствованию интеллектуальных информационных технологий (нейросетей) и носит методологический характер. В предшествующих статьях авторов были сделаны первые шаги в исследовании эмерджентных эффектов, порождаемых комплексированием нейросетевых технологий и агрегированием экзогенных переменных на базе обобщенных функций желательности Харрингтона. Первые

шаги исследований оказались успешными и были подтверждены вычислительными экспериментами и сравнением с классической (базовой) эконометрической моделью. Однако не был получен ответ на основной вопрос комплексирования указанных выше методов, предложенный авторами: сработает ли эта идея при ужесточении условий моделирования, характерных для некоторых классов прикладных задач? Речь идет о сильном зашумлении в данных (вплоть до их сознательного искажения), отягощённых отсутствием априорных сведений о виде закона распределения шумов, неполнотой данных, неопределенностью, большой размерностью пространства признаков (до нескольких сот экзогенных переменных и нескольких десятков эндогенных переменных). В данной статье получены четкие ответы на указанные выше вопросы и дополнительно исследованы вопросы об адекватности получаемой нейросетевой модели. Специфической особенностью рассмотренного класса нейросетевых моделей является то, что и эндогенные, и экзогенные переменные являются случайными величинами. Подробно исследованы вопросы регуляризации предложенного метода и его апробации на реальных данных обработки опросных анкет для экзогенных переменных и изменений с помощью сертифицированных измерительных средств для эндогенных переменных. Предложена концепция и оригинальный реализующий ее метод иерархического трехуровневого каскадного агрегирования эндогенных и экзогенных переменных нейросетевой модели, которые повышают прогностическую силу модели в очень сложных условиях моделирования. Концепция и реализующий ее метод апробированы на реальных данных из области профилактической медицины.

**Ключевые слова:** нейросетевая модель; агрегирование; экзогенные переменные; эндогенные переменные; обобщенная функция желательности Харрингтона; регуляризация; обратные задачи; восстановление зависимостей

## Введение. Постановка задачи нейросетевого моделирования. Актуальность темы исследования

В статье рассматривается обратная задача восстановления зависимостей, скрытых в данных

$$\hat{\vec{y}} = \varphi(\vec{x});\ \vec{y} = (\ y_1, \dots, y_u, \dots, y_m);\ \vec{x} = \left(\ x_1, \dots, x_j, \dots, x_n\right). \quad (1)$$

Здесь $\hat{\vec{y}}$ – вектор восстановленных (расчетных) значений экзогенных переменных; $\vec{x}$ – вектор экзогенных (объясняющих) переменных; $\varphi(\cdot)$ – оператор восстанавливаемой нелинейной зависимости (нейросетевая модель системы); $x_j \epsilon R^n$, $y_u \epsilon R^m$ – пространства вещественных чисел.

Таким образом, с позиций системного подхода целью моделирования в статье является получение новых знаний о моделируемой системе здоровьесохраняющей технологии для социального кластера обучающейся молодежи (студентов). Для достижения этой цели разрабатываются оригинальные нейросетевые модели.

В статье авторов [1] было показано, что объединение двух расчетных систем – нейросети и агрегирующей обобщенной функции желательности Харрингтона [2] для экзогенных переменных – порождает эмерджентный эффект. Он заключался в сокращении размерности факторного пространства: $R^n \to R^k, k \ll n$ и, соответственно, уменьшении энтропии в расчетной системе. Эндогенная переменная $Y$ в [1] рассматривалась как скалярная величина – вероятность $P$ банкротства корпорации.

В данной статье идея комплексирования расчетных методов из [1] развивается и углубляется, что обусловлено спецификой специально выбранной моделируемой системы с особо сложными условиями моделирования – объекта профилактической медицины для

социального кластера студентов вузов. Исследуемый социально-биологический кластер требует учета системного нелинейного воздействия различных факторов на индивидуум (студента), причем с оценкой результатов воздействия в разных аспектах. Соответственно, исходная эндогенная переменная *Y* должна быть не скаляром, а вектором в (1). При этом для содержательного описания моделируемого кластера набирается более 170 экзогенных переменных (факторов) и несколько десятков эндогенных переменных. Результаты наблюдений сильно зашумлены в силу того, что экзогенные переменные получаются путем обработки опросных анкет. Вид закона распределения шумов априорно неизвестен, что исключает возможность разработки аналитических методов регуляризации моделей.

Такая постановка задачи порождает, помимо практического интереса для профилактической медицины, два положительных момента для теории нейросетевого моделирования сложных систем:

1) Создается благодатная почва для проверки идей усовершенствования нейросетевых технологий в сложных условиях моделирования.

2) Векторный характер эндогенной переменной $Y$ в (1) создает предпосылки разработки концепции ступенчатого (каскадного) агрегирования переменных модели ($\{y_u\} \cup \{x_j\}$); $u = \overline{1,m}$; $j = \overline{1,n}$ (см. ниже в разделе 2 статьи). Здесь «∪» – символ дизъюнкции, т. е. объединения множеств.

Концепция ступенчатого (каскадного) агрегирования эндогенных и экзогенных переменных нейросетевых моделей на основе обобщенной функции желательности Харрингтона [1] предложена и исследована в данной статье впервые. Это обосновывает актуальность и научную новизну материала статьи.

## 1. Идея предлагаемой концепции и метода каскадной компрессии эндогенных и экзогенных переменных

Предлагаемая концепция, представляющая собой методологическую основу метода и разработанной нейросетевой модели (НСМ), формулируется так: «Концепция уменьшения размерности исходного пространства переменных НСМ [$R^n \times R^m$] » представляет собой алгоритм, организованный по иерархической трехуровневой схеме так, что предыдущий уровень (каскад) агрегирования переменных создает предпосылки для усиления полезного эффекта в последующем каскаде. В итоге на выходе НСМ степень компрессии переменных примерно равна произведению степеней компрессии каждого каскада (рис. 1). Порождаемый эмерджентный эффект – повышение качества НСМ в сложных условиях моделирования за счет сильной компрессии пространства переменных и сглаживания ошибок измерения переменных операторами частных функций желательности $\{d_j(x_j)\}$ и $\{d_u(y_u)\}$ и затем операторами обобщенных функций желательности $D_y$ и $D_x$ (см. ниже (2) и (3))».

Как отмечалось в [1], суммарная длина описания НСМ по информационному критерию Куллбака-Лейблера [3] определяет прогностическую силу модели, ограничивая сверху ожидаемый риск неверной оценки. Другими словами, следует уменьшать размерность пространства переменных с финишным контролем ее качества на тестовом множестве [9].

Заметим, что сформулированная выше концепция отражает общесистемный закон каскадного принципа усиления положительного эффекта [4]. Наглядный пример из техники: коэффициент усиления полезного сигнала в каскадном электронном усилителе примерно равен произведению коэффициентов усиления последовательных каскадов.



56ITVN118
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Сформулированная выше концепция послужила методологической основой оригинального гибридного метода, названного авторами «Метод каскадного агрегирования эндогенных и экзогенных переменных построения нейросетевой модели (МКА)». Алгоритм МКА, организованный по иерархической схеме рис. 1, детально описан ниже.


***Рисунок 1.*** *Метод каскадного агрегирования построения нейросетевой модели (составлено авторами)*

**На иерархическом уровне I**, назовем его «базовым», производится отбор показателей (признаков), включаемых в НСМ, т. е. $\Pi = [\{y_u\} \cup \{x_j\}]$ . На эвристическом уровне производится разделение пространства признаков $\Pi$ на две группы: $y_u \in \Pi_1$ эндогенных переменных и $x_j \in \Pi_2$ экзогенных переменных. Дискриминантным правилом разбиения $\Pi$ на две группы $\Pi_1$ и $\Pi_2$ служит способ получения информации, который аккумулирует профессиональный опыт аналитика, разрабатывающего модель. Здесь необходим глубокий анализ и учет специфики моделируемого объекта: пространство эндогенных переменных $\{y_u\} \in \Pi_1$ формируется с помощью приборных измерений отдельных показателей $y_u$ и образованием из них сверток и удельных безразмерных критериев (см. ниже в иллюстрационном примере раздела 3 статьи (жизненный индекс, адаптационный показатель, гармоничность телосложения и др.)) всего более 30 показателей. При образовании таких «мини-агрегатов» используются общепринятые формулы профилактической и спортивной медицины [5, 6].

Отметим два важных свойства изложенного способа формирования пространства эндогенных переменных $\Pi_1$ , которые предопределяют качество создаваемой НСМ: 1) получаемая информация, содержащаяся в $\Pi_1$, носит объективный характер; 2) образуемые



56ITVN118
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

«мини-агрегаты» $\{y_u\}$ уже подвергают компрессии исходные, измеренные приборами, показатели.

Пространство экзогенных переменных $x_j \in \Pi_2$, которое отражают нелинейное влияние внешней среды на объект (биосистему) – студента, формируется совсем другим способом – путем обработки данных опросных анкет. Здесь возможно сильное зашумление (вплоть до сознательного искажения). Это и служит посылом к совершенствованию нейросетевых информационных технологий в этой прикладной области.

**Агрегирование переменных на уровне II**. Здесь в блоке II.2 на базе частных функций желательности $d_{y,u}$ и обобщенной функции желательности Харрингтона $D_Y$ множество эндогенных переменных $\{y_u\} \in \Pi_1$ трансформируется в один обобщенный скалярный показатель $D_Y$ по формуле (2).

$$D_Y = \sqrt[m]{d_1 \cdot d_2 \cdot \ldots \cdot d_u \cdot \ldots \cdot d_m} \in [0;1]. \quad (2)$$

Функции частных желательностей $d_u(y'_u)$ осуществляют преобразование натуральных нормированных (кодированных) значений эндогенных переменных $y'_u$ в безразмерную шкалу желательности, имеющий интервал от 0 до 1. Значение $d_u = 0$ соответствует абсолютно неприемлемому уровню $u$-го признака, а значение $d_u = 1$ идеальному (самому лучшему) значению данного признака. Функции $d_u$ вычисляются либо по формуле (3), либо по графику (рис. 2).

$$d_u = \exp[-\exp(-y')]. \quad (3)$$

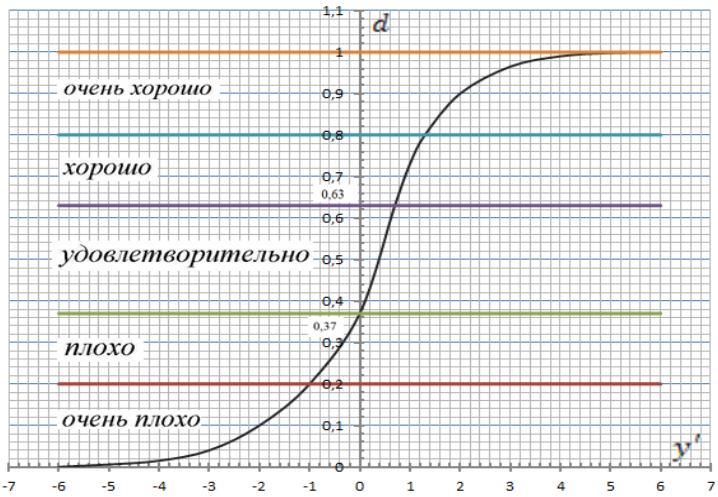


***Рисунок 2.*** *Функция желательности (рисунок заимствован из [2])*

Принято считать лингвистическую оценку «очень хорошо» соответствующей по шкале желательности диапазону $d_u \in [0{,}8; 1]$ , «хорошо» $d_u \in [0{,}63; 0{,}8]$ ; «удовлетворительно» $d_u \in [0{,}37; 0{,}63]$, «плохо» $d_u \in [0{,}2; 0{,}37]$, «очень плохо» (неприемлемо) $d_u \in [0; 0{,}2]$. Число интервалов на кодированной шкале $y'$ принято выбирать от 3 до 6. Выбор числа интервалов на кодированной шкале определяет крутизну кривой желательности в средней зоне.

Функция частных желательностей $d_u(y'_u)$ имеет три положительных свойства:

- Она хорошо интерпретируется в терминах качественных (лингвистических) оценок меры влияния $u$-ой эндогенной переменной на обобщеную функцию желательности Харингтона $D$ по (2).
- Функция $d_u \in [0; 1]$, поэтому она удобна в расчетах; функция по (3) непрерывно дифференцируема любое число раз.



56ITVN118
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

- Функция $d_u(y'_u)$ реализует сжимающий оператор по отношению к своему аргументу $y'_u$: например, на рис. 2 видно, что интервал $[-6; 6]$ отображается в интервал $[0; 1]$, т. е. коэффициент сжатия получается примерно равным 12. Данное свойство особенно важно для исследуемого класса задач с сильным зашумлением данных.

Отметим, что лингвистическая интерпретация значений обобщенной функции желательности Харрингтона *D* также производится в соответствии с рис. 2. Указанные выше положительные свойства частных функции желательностей характерны и для функции Харрингтона *D*.

В блоке II.1 алгоритма происходит преобразование экзогенных переменных $x_j \in \Pi_2$ в частные функции желательности $\{d_j\}$, аналогично операциям в блоке II.1. Разница состоит лишь в том, что в пространстве $\Pi_2$ на эвристическом уровне выделяются множества подгрупп экзогенных переменных $\left\{x_j^{(k)}\right\}$, $j = \overline{1,n}; (k) = 1,2,\dots,l$. Каждая *k*-я подгруппа отражает свой аспект влияния экзогенных показателей на обобщенный показатель *D* по (2).

Так в иллюстрационном примере из раздела 3 статьи обобщенной оценки здоровья студентов подгруппа $\Pi_{2.1}$ характеризует качество питания студентов, подгруппа $\Pi_{2.2}$ – упорядоченность образа жизни (режим), подгруппа $\Pi_{2.3}$ – приверженность к вредным привычкам и т. д.

Соответственно числу подгрупп $l$ для каждой из них образуются свои обобщенные функции желательности Харрингтоны $D_X^{(k)}(d_j), (k = \overline{1,l}; j = \overline{1,n})$ по (1).

Отметим два важных свойства образуемых на уровне II агрегатов $D_Y$ и $D_{x_j}$ в виде обобщенных функций желательности Харрингтона: 1) агрегируемые частные функции желательности $d_u$ и $d_j$, играющие роль аргументов в (2), должны быть однонаправленными; 2) в общей нелинейной НСМ должно учитываться влияние эндогенных частных желательностей $d_u$, $d_p$ $(u \neq p;\ u,p = 1,2,\dots,m)$ друг на друга. Первое свойство учитывается при кодировании (нормировании) натуральных величин аргумента формулы (2): $y \in [y_{min}, y_{max}] \rightarrow y' \in [-6, +6]$. Второе свойство выполняется автоматически в силу структуры формулы (2).

**На уровне III** алгоритма МКА строится байесовский ансамбль НСМ, связывающая вход модели $\left\{D_X^{(k)}(d_j)\right\}$ с выходом $D_Y(d_u)$:

$$D_X^{(k)}(d_j) \in R^l \xrightarrow{F} D_Y(d_u) \in R^1, \quad (4)$$

где: $F(\cdot)$ – оператор нейросетевого отображения [7].

Сделаем некоторые замечания по поводу построения НСМ на уровне III алгоритма:

1) Согласно классификации [4] агрегатов нейросеть вида (3) следует отнести к «агрегатам-сетям», которая на заключительном каскаде III тоже вносит свой вклад в компрессию переменных: $R^l \rightarrow R^1$, т. е. степень компрессии здесь оценивается числом подгрупп $l$ экзогенных переменных. В итоге в МКА степень компрессии достигается порядка сотен. В примере из раздела 3 статьи она составляет 566. Компрессия переменных в НСМ сопровождается регуляризацией расчетной модели на байесовским ансамбле.

2) Итоговый выходной показатель описанного алгоритма $D_Y(d_u) \in [0; 1]$ и поэтому допускает трактовку в терминах субъективной вероятности желаемого события.



56ITVN118
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

## 2. Количественные оценки метода каскадного агрегирования переменных и построения НСМ

Всего обследовалось из ведущих вузов Республики Башкортостан 800 студентов. Фрагмент выборки объемом 90 студентов, по которой для иллюстрации предлагаемого МКА строилась нейросетевая модель, показан в таблице 1 (эндогенные, т. е. выходные переменные в виде агрегатов – частных функций желательности $\{d_{yu}\}$). Всего выделено 11 групп основных эндогенных переменных, в которых агрегированы от двух до нескольких первичных, измеряемых с помощью сертифицированных приборов, показателей:

$Y_1$ – индекс массы тела, вычисляемый по формуле $Y_1 = M/h$, кг/м$^2$, где $M$ – масса тела, кг; $h$ – поверхность тела, м$^2$;

$Y_2$ – гармоничность телосложения (индекс Леви) [5, 6], $Y_2 = [T/Z] \cdot 100$, где $T$ – обхват грудной клетки в спокойном состоянии, см; Z – рост, см;

$Y_3$ – индекс относительной силы (в %), вычисляемый по формуле $Y_3 = [F/M] \cdot 100$ %, где $F$ – сила кисти (кг);

$Y_4$ – жизненный индекс (в мл/кг), вычисляемый по формуле: ЖЕЛ/М; где ЖЕЛ – жизненная емкость легких;

$Y_5$ – частота пульса, ударов/мин.;

$Y_6$ – циркулярно-респираторный индекс Скибинской [5, 6]: [ЖЕЛ/100·А]/$Y_5$, где А – задержка дыхания после вдоха, с;

$Y_7$ – артериальное давление, мм ртутного столба;

$Y_8$ – адаптационный показатель, вычисляемый по формуле: $Y_8$ = 0,011 · $Y_5$ + 0,014 · · (САД) + 0,008 · (ДАД) + 0,009 · М - 0,009 · $Y_5$ + 0,014 · В - 0,27; (где САД, ДАД – систологическое и диастологическое артериальное давление; В – возраст. Эта формула представляет собой уравнение регрессии, где числовые значения коэффициентов оценены по методу наименьших квадратов и являются размерными величинами [5, 6]. В то же время каждый член в правой части уравнения является безразмерной величиной; факторы в правой части имеют размерность: [$Y_5$] = ударов пульса/мин.; [САД, ДАД] = мм ртутного столба; [М] = кг; [В] = возраст, полных лет).

$Y_9$ – проба Руфье [5, 6], соответствующая формуле: $Y_9$ = (4 · (Р1 + Р2 + Р3) - 200) / 10; где Р1 – число пульсаций сердца за 15 с до физической нагрузки, Р2 – число пульсаций после нагрузки за первые 15 с, Р3 – число пульсаций за последние 15 с первой минуты периода восстановления;

$Y_{10}$ – показатель гибкости позвоночника (расстояние от ладоней до пола в наклоне со скамейки без сгибания коленей, см);

$Y_{11}$ – оценка личностной тревожности в баллах (психологический тест «Шкала самооценки уровня тревожности» Ч.Д. Спилберга и Ю.Л. Ханина) [5, 6].



56ITVN118
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

**Таблица 1**

**Фрагмент агрегирования эндогенных переменных на основе обобщенной функции желательности Харрингтона $D_Y$**

| Номер наблюдения | Частные функции желательности эндогенных переменных $d_{Y,u,i}$ $(u = \overline{1,11}; i = \overline{1,10})$ | | | | | | | | | | | Произведения частных желательностей | Обобщенные функции желательности Харрингтона |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $i$ | $d_{1i}$ | $d_{2i}$ | $d_{3i}$ | $d_{4i}$ | $d_{5i}$ | $d_{6i}$ | $d_{7i}$ | $d_{8i}$ | $d_{9i}$ | $d_{10i}$ | $d_{11i}$ | $\prod_{u=1}^{n} d_{ui}$ | $D_{Y.i}$ |
| 1 | 0,9 | 0,7 | 0,75 | 0,95 | 0,45 | 0,75 | 0,5 | 0,65 | 0,35 | 0,985 | 0,72 | 0,01212 | 0,6695 |
| 2 | 0,95 | 0,65 | 0,86 | 0,81 | 0,75 | 0,74 | 0,5 | 0,7 | 0,55 | 0,63 | 0,1 | 0,00289 | 0,5878 |
| 3 | 0,85 | 0,5 | 0,85 | 0,3 | 0,5 | 0,5 | 0,6 | 0,14 | 0,68 | 0,8 | 0,39 | 0,00048 | 0,4995 |
| 4 | 0,35 | 0,9 | 0,68 | 0,36 | 0,21 | 0,67 | 0,5 | 0,25 | 0,75 | 0,7 | 0,6 | 0,00043 | 0,494 |
| 5 | 0,68 | 0,7 | 0,7 | 0,58 | 0,27 | 0,9 | 0,5 | 0,7 | 0,9 | 0,7 | 0,15 | 0,001553 | 0,555 |
| 6 | 0,45 | 0,37 | 0,68 | 0,19 | 0,15 | 0,45 | 0,19 | 0,3 | 0,5 | 0,53 | 0,51 | 0,000012 | 0,3547 |
| 7 | 0,65 | 0,5 | 0,37 | 0,36 | 0,29 | 0,5 | 0,37 | 0,17 | 0,4 | 0,5 | 0,18 | 0,000042 | 0,3625 |
| 8 | 0,83 | 0,6 | 0,68 | 0,7 | 0,46 | 0,68 | 0,5 | 0,81 | 0,81 | 0,285 | 0,25 | 0,00289 | 0,588 |
| 9 | 0,68 | 0,65 | 0,69 | 0,7 | 0,31 | 0,72 | 0,4 | 0,5 | 0,81 | 0,7 | 0,58 | 0,00313 | 0,592 |
| 10 | 0,83 | 0,68 | 0,72 | 0,3 | 0,71 | 0,68 | 0,75 | 0,7 | 0,9 | 0,98 | 0,81 | 0,0221 | 0,707 |

*Составлено авторами*

Как отмечалось выше, экзогенные первичные показатели $\{x_j\}$ получены из обработки анкетных карт студентов-респондентов. Каждая карта содержала 170 вопросов, сформированных с использованием общепринятых методик в профилактической и спортивной медицине [5, 6]. Ответы содержали как количественные, так и качественные показатели. Эти показатели пересчитывались в соответствующие тесты, пробы и величины (в размерных и безразмерных величинах) по формулам из [5, 6].

Образовано 9 подгрупп экзогенных переменных, для которых в каждом $i$-ом опыте получены соответствующие агрегированные показатели – обобщенные функции желательности Харингтона $D_{x_j}^{(k)}$, показанные в таблице 2. Эти показатели интерпретируются следующим образом: $D_{x_j}^{(1)}$ – качество питания (агрегировано 18 факторов); $D_{x_j}^{(2)}$ – режим использования телевизора и компьютеров (агрегировано 10 факторов); $D_{x_j}^{(3)}$–занятие спортом и закаливание (агрегировано 3 фактора); $D_{x_j}^{(4)}$ – режим труда и отдыха (агрегировано 5 факторов); $D_{x_j}^{(5)}$ – вредные привычки (агрегировано 11 факторов); $D_{x_j}^{(6)}$ – самооценка состояния здоровья (агрегировано 10 факторов); $D_{x_j}^{(7)}$ – оценка психического здоровья (агрегировано 7 факторов); $D_{x_j}^{(8)}$ – условие проживания и финансовое обеспечение (агрегировано 7 факторов); $D_{x_j}^{(9)}$ – нравственные установки респондента (агрегировано 19 факторов).

**Таблица 2**

**Фрагмент агрегирования экзогенных переменных в виде функций Харрингтона $D_{x_j}^{(k)}$**

| $i$ | $D_{x_j}^{(1)}$ | $D_{x_j}^{(2)}$ | $D_{x_j}^{(3)}$ | $D_{x_j}^{(4)}$ | $D_{x_j}^{(5)}$ | $D_{x_j}^{(6)}$ | $D_{x_j}^{(7)}$ | $D_{x_j}^{(8)}$ | $D_{x_j}^{(9)}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,4816 | 0,4346 | 0,5521 | 0,5996 | 0,6260 | 0,6878 | 0,8133 | 0,6101 | 0,6388 |
| 2 | 0,5673 | 0,4577 | 0,5493 | 0,5915 | 0,3757 | 0,5856 | 0,8133 | 0,6424 | 0,4724 |
| 3 | 0,6085 | 0,3694 | 0,6383 | 0,7524 | 0,6594 | 0,5006 | 0,4177 | 0,6401 | 0,6760 |
| 4 | 0,5653 | 0,7336 | 0,5196 | 0,3985 | 0,7259 | 0,6296 | 0,8133 | 0,7165 | 0,7127 |
| 5 | 0,5678 | 0,4529 | 0,6663 | 0,6144 | 0,5262 | 0,6542 | 0,8133 | 0,7745 | 0,6307 |
| 6 | 0,6316 | 0,5490 | 0,5783 | 0,6811 | 0,6925 | 0,5986 | 0,8133 | 0,6052 | 0,7251 |
| 7 | 0,6398 | 0,5337 | 0,7773 | 0,4876 | 0,6957 | 0,5492 | 0,8133 | 0,5801 | 0,6437 |
| 8 | 0,5231 | 0,5897 | 0,4751 | 0,7225 | 0,6727 | 0,6418 | 0,5061 | 0,5648 | 0,6437 |

| $i$ | $D_{x_j}^{(1)}$ | $D_{x_j}^{(2)}$ | $D_{x_j}^{(3)}$ | $D_{x_j}^{(4)}$ | $D_{x_j}^{(5)}$ | $D_{x_j}^{(6)}$ | $D_{x_j}^{(7)}$ | $D_{x_j}^{(8)}$ | $D_{x_j}^{(9)}$ |
|---|---|---|---|---|---|---|---|---|---|
| 9 | 0,6587 | 0,4417 | 0,8500 | 0,6746 | 0,5055 | 0,5432 | 0,8133 | 0,7691 | 0,6004 |
| 10 | 0,5261 | 0,4671 | 0,3902 | 0,7822 | 0,5030 | 0,5924 | 0,7635 | 0,6720 | 0,5724 |

*Составлено авторами*

Теперь изложим алгоритм построения «нейросети-агрегата» в каскаде III рис. 1. Использовалась демоверсия программного продукта «Neurisolutions-4.0». Для регуляризации НСМ был построен байесовский ансамбль сетей, аналогичных [1, 10]. Все НСМ ансамбля принадлежали к одному классу – «многослойные персептроны (MLP) с обратным распространением ошибки (BP)». В скрытых слоях варьировалось их число и типы активационных функций – логистическая функция и гиперболический тангенс. Поскольку условия моделирования сложны, то заведомо нарушаются все предпосылки классического регрессионного анализа. Поэтому, для оценки качества НСМ мы использовали прямой критерий, не стесненный требованиями к статистическим свойствам остатков – процент правильной идентификации моделируемого показателя $D_Y$ на тестовом множестве примеров $\Omega_{test}$, которого сеть не знала при обучении. Использовался критерий оценки:

$$\eta = \frac{N^*}{N} \cdot 100\%;\ N_i^*: \left[\left(D_{Yi} - \widehat{D_{Yi}}\right)/D_{Yi}\right] \le \varepsilon;\ N^* = \left(\sum_{i=1}^{N} N_i^*\right)\Big|\Omega_{test}. \quad (5)$$

Здесь звездочкой «*» отмечены точки тестового множества, в которых НСМ верно идентифицирует обобщенный показатель здоровья, т. е. с достаточно малой погрешностью ε, назначаемой аналитиком (мы задавали ε = 0,05); $D_{Yi}, \widehat{D_{Yi}}$ – опытные и расчетные значения моделируемой величины $D_Y$; $N$ – общее количество точек тестового множества ($N$ = *20*). В вычислениях получено $\eta$ = 80 %.

Данная оценка качества дополнялась индексом множественной корреляции $r_{D,\hat{D}}$, выдаваемая НСМ при обучении. Получено достаточно высокое значение этого показателя: $r_{D,\hat{D}} = 0{,}94$.

На рис. 3-4 и таблице 3 показаны результаты обучения и тестирования НСМ, а в таблице 3 – данные по регуляризации модели согласно байесовскому подходу из [1, 8]. Модельное значение $D_Y$ усреднялось на отфильтрованном байесовском ансамбле нейросетей. При фильтрации отсеяны НСМ № 1, 4, 6, 8 из таблицы 3.

**Таблица 3**

**Характеристика моделей байесовского ансамбля на тестовом множестве из 20 точек**

| № п/п | Количество скрытых слоев / типы активационных функций в слоях | $N^*$ | $N$ | $N^*/N$ |
|---|---|---|---|---|
| 1 | 2 / Гиперболический тангенс – 1 слой, Линейная – 2 слой | 12 | 20 | 0,60 |
| 2 | 2 / Гиперболический тангенс – 1 и 2 слои | 19 | | 0,95 |
| 3 | 2 / Сигмоид – 1 и 2 слои | 17 | | 0,85 |
| 4 | 2 / Гиперболический тангенс – 1 слой, Сигмоид – 2 слой | 12 | | 0,60 |
| 5 | 2 / Сигмоид– 1 слой, Гиперболический тангенс – 2 слой | 15 | | 0,75 |
| 6 | 2 /Линейная – 1 слой, Гиперболический тангенс – 2 слой | 12 | | 0,60 |
| 7 | 2 / Линейная – 1 и 2 слои | 18 | | 0,90 |
| 8 | 2 / Сигмоид– 1 слой, Линейная – 2 слой | 13 | | 0,65 |
| 9 | 2 / Линейная – 1 слой, Сигмоид – 2 слой | 18 | | 0,90 |
| 10 | 1 / Гиперболический тангенс – 1 слой | 17 | | 0,85 |

*Составлено авторами*

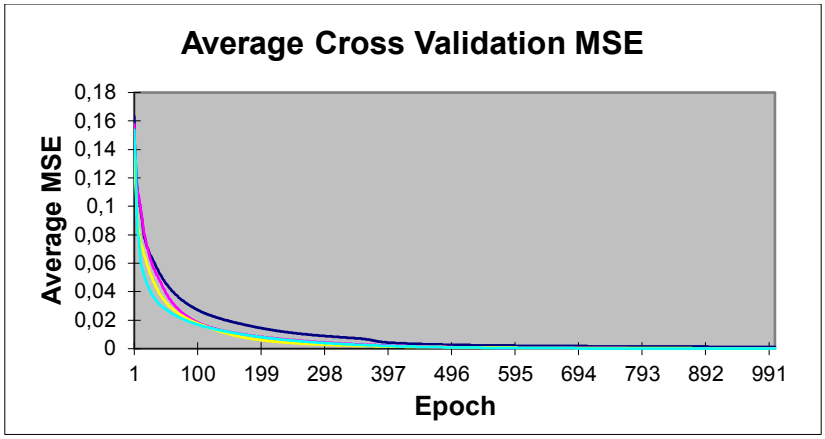


***Рисунок 3.*** *Зависимость средней квадратической ошибки (MSE) обучения от числа эпох (составлено авторами)*

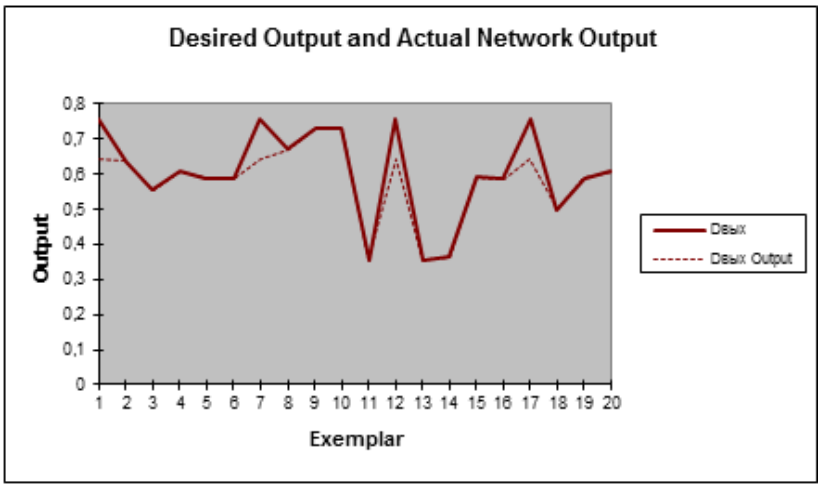


***Рисунок 4.*** *Сравнение опытных данных $D_{Yi}$ и расчетных $\widehat{D_{Yi}}$ значений моделируемого показателя на тестовом множестве НСМ (составлено авторами)*

Из рис. 3 видно, что сеть обучается успешно: MSE снижается до 0,002 за 595 эпох без эффекта «переобучения».

Из рис. 4 видно, что из 20 тестовых точек 16 точек идентифицируется верно при $\varepsilon < 0{,}05$ в формуле критерия (4).

## Выводы

1. Основная идея метода каскадного агрегирования (МКА) эндогенных и экзогенных переменных на основе обобщенных функций желательности Харингтона и НСМ подтверждена на реальных данных комплексного исследования студентов вузов. Получила степень компрессии переменных на трех каскадах агрегирования по рис. 1 примерно 566. Соответственно в сложных условиях моделирования, характерных для этой прикладной задачи, НСМ хорошо обучается и хорошо тестируется (процент правильно идентифицируемых точек достигает 80 % по рис. 4).

2. На основе построенной обученной, протестированной и регуляризированной НСМ могут решаться прикладные задачи:

- Выявление групп риска потери (или ухудшения) здоровья студентов при групповом и индивидуальном обследовании по опросным анкетам, т. е. без



56ITVN118
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

применения дорогостоящих лабораторных измерений с использованием только информации из опросных анкет.

- Разработки методик комплексного обследования.
- Оценка и сравнение степени влияния агрегированных экзогенных показателей $\left\{D_X^{(k)}\right\}, k = \overline{1,l}$ на результативный обобщенный показатель $D_Y$ здоровья кластера респондентов (студентов). Для этого в НСМ численно определяются средние значения коэффициентов эластичности

$$\overline{Э}_k = \frac{1}{M}\sum_{q=1}^{M} \frac{\Delta\widehat{D}_{Y\cdot q}}{\Delta D_{X\cdot q}^{(k)}} \cdot \frac{D_{X\cdot q}^{(k)}}{\widehat{D}_{Y\cdot q}};\ k = \overline{1,l}. \quad (6)$$

где: $\Delta\widehat{D}_Y$, $\Delta D_X^{(k)}$ – конечные приращения функции результативной переменной $D_{Y\cdot q}$ и ее аргумента $D_{X\cdot q}$ в текущей точке $q$; $M$ – число дискретных точек для расчета коэффициентов эластичности для каждой $k$-ой группы экзогенных переменных.

## ЛИТЕРАТУРА

1. Горбатков С.А., Фархиева С.А. Системный подход к агрегированию экзогенных и эндогенных переменных в нейросетевых моделях банкротств на основе функций Харрингтона // журнал «Вестник Евразийской Науки», 2017 № 3 (9) [Электронный ресурс] – М.: Науковедение, 2017. Режим доступа: http://naukovedenie.ru/PDF/100TVN317.pdf, свободный. – Загл. с экрана. – Яз. рус., англ.
2. 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.
3. Деза Е., Деза М.М. Энциклопедический словарь расстояний / Елена Деза, М.М. Деза; пер. с англ. В.И. Сычева. – М.: Наука, 2008. – 448 с.
4. Перегудов Ф.И., Тарасенко Ф.П. Основы системного анализа: Учебник. – Томск: Изд-во НТЛ, 1997. – 367 с.
5. Баевский Р.М., Берсенева А.П. Оценка Адаптационных возможностей организма и риска развития заболеваний: Монография, 1997. – 234 с.
6. Дубровский В.Н. Спортивная медицина. М.: Владос, 2007. – 512 с.
7. Горбатков С.А. Методологические основы разработки нейросетевых моделей экономических объектов в условиях неопределенности / С.А. Горбатков, Д.В. Полупанов, Е.Ю. Макеева, А.Н. Бирюков; под ред. С.А. Горбаткова. – М.: Издательский дом «Экономическая газета», 2012. – 494 с.
8. Горбатков С.А., Горбаткова Е.Ю. Использование байесовской регуляризации модели анализа условий и образа жизни обучающейся молодежи // Современные проблемы науки и образования. – 2015. – №3(59). Режим доступа: https://science-education.ru/ru/article/view?id=18842, свободный.
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. Шумский С.А. Байесова регуляризация обучения // Лекции школы-семинара «Современные проблемы нейроинформатики» (23-25 января 2002 г., Москва). – М.: МИФИ, 2002. – С. 61-94.

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

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

**Gorbatkova Elena Yurevna**
Bashkir state pedagogical university named after M. Akmullah, Ufa, Russia
E-mail: gorbatkovaeu@mail.ru

# The method of aggregating the variables of a neural network model in inverse problems of recovering the dependence in conditions of high dimensionality of the space of attributes and data noisiness

**Abstract.** The article is devoted to the improvement of intellectual information technologies (neural networks) and is methodological in nature. In the previous articles of the authors, the first steps were taken to investigate emergent effects caused by the integration of neural network technologies and the aggregation of exogenous variables on the basis of generalized Harrington desirability functions. The first steps of the research were successful and were confirmed by computational experiments and comparison with the classical (basic) econometrics model. However, no answer was received to the main question of the combination of the above methods proposed by the authors: will this idea work if the modeling conditions that are characteristic of certain classes of applied problems are tightened? We are talking about a strong noise in the data (down to their conscious distortion), burdened by the lack of a priori information about the form of the noise distribution law, the incomplete data, the uncertainty, the large dimensionality of the feature space (up to several hundred exogenous variables and several dozen endogenous variables). In this article, clear answers have been received to the above questions and further questions have been investigated about the adequacy of the resulting neural network model. A specific feature of the considered class of neural network models is that both endogenous and exogenous variables are random variables. The questions of regularization of the proposed method and its approbation on real data of processing questionnaires for exogenous variables and changes with the help of certified measuring means for endogenous variables are studied in detail. The concept and the original method of hierarchical three-level cascade aggregation of endogenous and exogenous variables of neural network model realizing it which increase the predictive force of model in very difficult conditions of modeling is offered. The concept and the method that implements it are tested on real data from the field of preventive medicine.

**Keywords:** neural network model; aggregation; exogenous variables; endogenous variables; the generalized function of desirability of Harrington; regularization; the return tasks; restoration of dependences