Интернет-журнал «Науковедение» ISSN 2223-5167 http://naukovedenie.ru/
Том 7, №5 (2015) http://naukovedenie.ru/index.php?p=vol7-5
URL статьи: http://naukovedenie.ru/PDF/63TVN515.pdf
DOI: 10.15862/63TVN515 (http://dx.doi.org/10.15862/63TVN515)

УДК 336.025

Бирюков Александр Николаевич
ФГБОУ ВПО «Башкирский государственный университет»
Россия, Уфа
Филиал в г. Стерлитамак[1]
Профессор кафедры «Экономической теории и анализа»
Доктор экономических наук
E-mail: biryukov_str@mail.ru

Метод квазирешений для регуляризации нейросетевых моделей налогового контроля

[1] 453100, Республика Башкортостан, Стерлитамак, пр. Ленина 49 а

Аннотация. Вопросы, рассматриваемые в статье, возникли в связи с объективной необходимостью проведения исследований, направленных на повышение эффективности работы налоговой системы регионального уровня.

Государство не может тратить большие средства на сбор налогов, поэтому сама структура налогов и государственная система налогового администрирования (СНА), обеспечивающие их сбор, должны при минимальных затратах обеспечивать высокую эффективность работы, которая невозможна без хорошей информационно-аналитической системы поддержки принимаемых решений. Построение такой системы предполагает интеграцию в единое информационное пространство всех структурных подразделений налоговых и других государственных органов. Необходимое единое информационное пространство создавалось в последние годы в виде системы электронной обработки данных (ЭОД), разработанной ФНС РФ. Эта система, которая является основой автоматизации в работе налоговых органов, описана практически во всех учебниках российских экономических вузов.

Однако ЭОД имеет одно узкое место – в ней слабо форматизирован аналитический блок, ядром которого должна служить математическая модель анализа финансово-экономического состояния налогоплательщиков, выявления нарушений налогового законодательства в декларациях, синтеза оптимального плана выездных налоговых проверок. Такие модели должны служить достаточно достоверной и объективной основой для поддержки принятия управленческих решений.

В настоящее время существует мощный математический инструментарий (универсальный аппроксиматор и кластеризатор) – *нейронные сети.* При использовании нейросетей требуется их обучение на примерах, что с математической точки зрения является некорректно поставленной по Адамару обратной задачей типа задачи интерпретации. Здесь возникает дилемма нахождения компромисса между ошибками обобщения модели и ее робастностью (устойчивостью к вариации данных в заданном диапазоне).

Применительно к нейросетевым моделям СНА, которые отличаются сильным зашумлением данных, отягченным в ряде случаев дефицитом наблюдений, исследования по регуляризации нейросетей в обратных задачах не проводились.

В работах В.К. Иванова [8] дано строгое математическое обоснование двух методов решения некорректно поставленных задач при условии, что имеется дополнительная априорная информация об искомом решении. Если известно, что решение является элементом заданного компакта, им был разработан метод квазирешений. В этом случае возможна и оценка погрешности приближенного решения.

В статье развивается идея новой компьютерной технологии предварительной (камеральной) налоговой проверки предприятий-налогоплательщиков, предложенная на основе нейросетевого моделирования. Использование этих моделей создает основу для повышения достоверности и объективности налогового контроля в налоговых органах, и повысить результативность выездных налоговых проверок.

Ключевые слова: нейросеть (НС); нейросетевая модель (НСМ); байесовский подход; метод вложенных математических моделей (ВММ); нейросетевая субмодель (НССМ)**;** метод квазирешений; алгоритм.

Ссылка для цитирования этой статьи:

Бирюков А.Н. Метод квазирешений для регуляризации нейросетевых моделей налогового контроля // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 7, №5 (2015) http://naukovedenie.ru/PDF/63TVN515.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ. DOI: 10.15862/63TVN515

Введение

Предметом исследования в статье являются вопросы регуляризации нейросетевых моделей (НСМ) в задачах ранжирования объектов налогового контроля юридических лиц по степени нарушения налогового законодательства, и соответственно, ожидаемых доначислений. Эти задачи имеют характерную особенность – сильное зашумление данных, которое усугубляется дефицитом наблюдений [1]. Как показано в [2] в столь сложных условиях моделирования НСМ подлежит регуляризации с целью устранения в ней чрезмерной чувствительности выхода к небольшим изменениям входных данных.

В работе автора [3] была сформулирована концепция обеспечения состоятельности алгоритмов регуляризации НСМ с сильным зашумлением данных. В работе автора [4] была описана идея применения байесовского подхода [2] к управлению оценкой погрешности аппроксимации в НСМ.

В данной статье концепция регуляризации НСМ из [3, 4] и реализующий её метод вложенных математических моделей (ВММ) детализированы в замкнутой форме и апробированы на реальных данных налоговых деклараций строительных предприятий.

1. Задача обучения НСМ как обратная задача восстановления многомерной нелинейной параметризованной функции и проблема её регуляризации

Согласно предлагаемому подходу к построению устойчивых НСМ налогового контроля требуется восстановить некоторую многомерную, в общем случае нелинейную, функцию $Y(\vec{X})$, где $\vec{X} = (X_1, X_2, ..., X_j, ..., X_n)$ - вектор объясняющих (входных) переменных, Y – эндогенная (объясняющая) переменная, для простоты считающаяся скалярной.

Функция Y должна нести в себе информацию, прямую или косвенную, о величине налогооблагаемой базы. Причем, если НСМ строится на кластере примерно однородных по выбранной числовой мере объектов налогообложения, то относительные отклонения (эталонных) расчётных $\hat{Y}(\vec{X})$ и декларированных $Y(\vec{X})$ значений функции Y (см. ниже формулу (24)) несут важную информацию для камеральных налоговых проверок. В качестве Y можно использовать такие показатели, как выручка, налог на добавленную стоимость, либо линейные свёртки (агрегаты) из экономических показателей, включаемых в декларации.

Нейросетевую модель можно записать в виде:

$$\hat{Y}(\vec{X}) = F(\vec{X}, W), \quad \vec{X} \in X^{(n)} \subset R^n, \tag{1}$$

где W – матрица параметров (синаптических весов связи между нейронами) НСМ, элементы которой представляют собой вещественные числа; R^n - n-мерное пространство вещественных чисел.

Оператор НСМ, отображающий пространство $\vec{X}$ на пространство $\vec{Y}$ при эталонно заданных параметрах W является композицией двух операторов – проецирования входных сигналов нейронов и затем нелинейной аппроксимации результатов проецирования:

$$F = F_1 \circ F_2, \tag{2}$$

$$F_1(\vec{X}) \equiv S_p = \sum_{j=1}^{n} w_{pj} x_{jp} - \Theta_p; \tag{3}$$

$$F_2(S_p) \equiv Y_p = f(S_p), \tag{4}$$

где w_{pj} - элемент матрицы W, т.е. синаптический вес p-го нейрона по j-му входу; S_p - функция состояния p-го нейрона; Θ_p - порог его возбуждения; $f(S_p)$ - активационная (передаточная) функция.

Нелинейная функция $f(S_p)$ в промежуточных слоях НСМ выбиралась из класса непрерывно дифференцируемых согласно байесовскому подходу из [4] (см. ниже в вычислительных экспериментах).

В режиме обучения сети матрица синаптических весов W модифицируется (адаптируется) к подаваемым на вход обучающим примерам – кортежам $\langle y_i, \vec{x}_i \rangle, i = \overline{1,N}$, где N – объём обучающей выборки (здесь и далее конкретные числовые реализации случайных величин X_j, Y обозначаются малыми латинскими буквами. Использовался известный алгоритм обратного распространения ошибки (англ. *backpropagation* (BP)) для обучения сети, в котором веса w_{pj} исправляются итерационно с помощью градиентного метода, в котором минимизируется квадратичный функционал E:

$$W^* : \left\{ w_{pj}^k - \eta^{(k)} \frac{\partial E}{\partial w_{pj}}; E(W) = \frac{1}{2} \sum_{pj} (Y_{pj}^N - d_{pj})^2 \to \min_w E(W), k = 0,1,2,... \right\}, \quad (5)$$

где $\eta^{(k)}$ - длина шага обучения на k-той итерации; $Y_{pj}^{(N)}$ - расчётное значение p-го нейрона в выходном N-ном слое при подаче на его входы i-го обучающего примера; d_{pj} - идеальное (экспериментальное) значение выхода в i-том примере. При уменьшении длины шага обучения, например, пропорционально $1/k$, итерационная процедура (5) приводит к нахождению локального минимума ошибки аппроксимации E_u соответственно нахождению оптимальных весов $W*$ [5].

Таким образом, в режиме обучения используется композиция трёх операторов: F_1, F_2 *и* F_3 по (5).

Для конкретности и наглядности изложения предлагаемого подхода к регуляризации НСМ перейдём к её описанию в терминах функционального анализа [6,7]. НСМ можно представить в виде операторного уравнения:

$$Az = u; \quad (6)$$

где u – наблюдаемые на выходе сети характеристики изучаемого объекта (процесса); $Z = Z(\vec{X}, W)$ восстанавливаемые НСМ параметризованные многомерные нелинейные функции; $A(\cdot)$ - оператор связи «вход-выход» сети, который можно представить в виде:

$$A = \begin{cases} F = F_1 \circ F_2 \text{ в режиме расчёта с заданной матрицей } W \\ F = F_1 \circ F_2 \circ F_3 \text{ в режиме обучения сети.} \end{cases} \quad (7)$$

$$Z = Z(\vec{X}, W) \in Z; u \in U. \quad (8)$$

Как видно из (1) - (6) оператор $A(\cdot)$ является нелинейным в следствие нелинейности активационных функций $f(\cdot)$ в (3).

Пусть пространство Z является подмножеством в пространстве непрерывных дифференцируемых по своим аргументам функций:

$$Z \subset C^1(\Omega);\ \Omega = \left\{a_j \le x_j \le b_j; \left|w_{pj}\right| \le B\right\},\ j = \overline{1,n};\ p = \overline{1,m}, \tag{9}$$

где a_j, b_j, B - заданные положительные числа.

Алгоритм НСМ (1) - (8) и смысл её функционирования [5] допускают введение в модель априорной информации вида (9).

Согласно представлению (6) - (9) НСМ реализует две подзадачи:

а) ***прямую подзадачу аппроксимации*** элементов по известным после обучения сети характеристикам процесса $z \in Z$ с помощью оператора $A(\vec{x}, W)$ в (1) - (4). Эта задача корректно поставлена.

б) ***обратную задачу аппроксимации*** для режима обучения сети, в которой заданы кортежи $\left\langle \vec{x}_i, y_i \right\rangle$, зафиксированы архитектура сети, активационные функции $f(s)$ в промежуточных слоях и в выходном слое, а также правило обучения сети, например алгоритм обратного распространения (*backpropagation* (BP)), а искомыми элементами является параметризованная функция $Z(\vec{x}_i, w)$ по алгоритму (1) - (6). Эта задача является в общем случае некорректно поставленной, поскольку по известному следствию – совокупности элементов $\{U_i\}, i = \overline{1,N}$ - требуется найти причину, т.е. восстановить элементы $Z(\vec{X}, W)$. Как известно [6,7] такие задачи относятся к классу некорректно поставленных по Адамару, и в условиях зашумления данных, отягченного их дефицитом, требуют специальных процедур регуляризации [2].

Рассмотрим вопрос о регуляризации НСМ подробнее. Условия корректно поставленной задачи по Адамару:

1. $\forall u \in U$ решение существует, т.е. обратная задача $Z = A^{-1}u$ разрешима;

2. $\forall u \in U$ решение Z единственно;

3. Решение Z непрерывно зависит от исходных данных.

Нарушение любого из трёх условий ведёт к некорректной постановке задачи. Заметим, что условия 1) и 2) характеризуют математическую определённость задачи, а условие 3) – её экономическую детерминированность (неслучайность).

Поясним смысл условия 3) и понятие квазирешения на широко известном примере решения интегрального уравнения [7]. Пусть $A(\cdot)$ - интегральный оператор и требуется найти функцию $Z(s)$ из интегрального уравнения вида [6, 7]:

$$\int_a^b K(x,s)z(s)ds = u(x), \quad x \in [c;d] \tag{10}$$

по известной правой части $u(x)$. Пусть ядро $K(x,s)$ данного интегрального уравнения и искомая функция $Z(s)$ удовлетворяют условиям: $K(x,s), K_x'(x,s), K_z'(x,s)$ непрерывны в прямоугольнике $c \le x \le d, a \le s \le b,$ а $u(x) \in C[c,d]$, т.е. функция $u(x)$ непрерывна, но условие её непрерывной дифференцируемости на $[c,d]$ не наложено. Обратная задача (10) является некорректной. Действительно, для неё не выполняется условие 1), поскольку решение (10) существует не для любой непрерывной функции $u(x) \in C[c,d]$, а только для непрерывно дифференцируемой функции $u(x) \in C^1[c,d]$. Если последнее условие не наложено, то интегральное уравнение (10) не может иметь непрерывное решение $Z_0(x)$. Это

следует из того, что для любой непрерывной функции $Z(s)$ и оговорённых выше условий на ядро $K(x,s)$ интеграл в левой части (10) представляет собой функцию непрерывную и дифференцируемую. Значит и правая части $u(x)$ должна быть непрерывной и дифференцируемой, что противоречит наложенному условию $u(x) \in C[c,d]$. Другими словами, классическое решение обратной задачи (10) в классе функций правой части $u(x) \in C[a,b]$ не существует для непрерывно дифференцируемых ядер $K(x,s)$.

Теперь уточним понятие приближённого решения при неточном наблюдении характеристик процесса - правой части (6), которую обозначим $\tilde{u}(x)$. В прикладных задачах имеет место именно такая ситуация. В рассматриваемом примере (10) мы априори полагали, что существует точное решение $Z(s)$ уравнения (10), отвечающее точной правой части $u_T(x)$ и требуется найти приближение к нему $\tilde{Z}(s)$, если вместо $u_T(x)$ известна приближённая правая часть $\tilde{u}(x)$ с оценкой:

$$\rho_U(u_T, \tilde{u}) \le \delta, \quad u \in U \tag{11}$$

где ρ - расстояние между элементами u_T и $\tilde{u}$ в пространстве *U*.

На практике может не быть информации о существовании точного искомого решения уравнения (10), но имеется информация о классе возможных правых частей *U* и можно ставить вопрос о нахождении приближённого «решения» $\tilde{Z}(s)$ уравнения (10). Под «приближённым» решением надо понимать некоторое обобщённое решение, которое уточняется ниже.

В [7] определено понятие обобщённого решения квазирешения уравнения (10) на множестве *Z* как такого элемента $\tilde{z} \in Z$, на котором расстояние $\rho_U(Az, \tilde{u})$ достигает точной нижней границы т.е.:

$$\rho_U(A\tilde{z}, \tilde{u}) = \inf_{z \in Z} \rho_U(Az, \tilde{u}) \tag{12}$$

$$Az \equiv \int_a^b K(x,s) z(s) ds. \tag{13}$$

Очевидно, что при $u = \tilde{u}$ квазирешение совпадает с обычным точным решением $z_T \in U$. Таким образом, условие 3) для обратной задачи сводится к нахождению таких алгоритмов построения обобщённых решений (квазирешений), которые устойчивы к малым изменениям правой части $u(x)$.

Изложим общие соображения по построению квазирешения применительно к процессу обучения нейросети. Если в обратной задаче (6) *A* – вполне непрерывный оператор [6,7], тогда обратный к нему оператор A^{-1}, вообще говоря, не будет непрерывным на *U* и решение уравнения (6) не будет устойчивым к малым изменениям правой части и в метрике пространства *U*. Действительно, если оператор *A* – вполне непрерывный, по малым возмущениям *Az* в (6), и соответственно, ρ в (11) могут отвечать возмущения z, а (значит, и $\tilde{z}$), далёкие от точного решения задачи.

В рассматриваемой обратной задаче восстановления нелинейной многомерной функции $Y(\vec{X})$ с помощью нейросети исходными данными являются правая часть уравнения (6) и оператор *A*. В обратной задаче обучения сети этот оператор $F_1 \circ F_2 \circ F_3$ по (7) можно считать заданным точно. Предположим, что правая часть уравнения (6) $\tilde{u}$ известна с

точностью δ такой, что $\rho_U(u_T, \tilde{u}) \le \delta$. По имеющимся данным $(\tilde{u}, \delta)$ требуется найти такой элемент $z_\delta \in Z_\delta$, который стремился бы (в метрике *Z*) к точному решению Z_T при $\delta \to 0$. Такой элемент по терминологии [6, 7] называется приближённым к Z_T решением уравнения $Az = \tilde{u}$.

Элементы $\tilde{z} \in Z$, удовлетворяют условию:

$$\rho_U(A\tilde{z}, \tilde{u}) \le \delta \tag{14}$$

Называются *сопоставимыми по точности* с исходными данными $(\tilde{u}, \delta)$. Пусть Z_δ - совокупность всех таких элементов $\tilde{z} \in Z$. Естественно приближённые решения уравнения $Az = \tilde{u}$ искать в классе Z_δ элементов *Z*, сопоставимых поточности с исходными данными $(\tilde{u}, \delta)$. В [6] такой класс Z_δ называется *множеством практической эквивалентности*.

Однако в ряде случаев класс Z_δ может быть слишком широким. Например, в задаче налогового контроля, рассматриваемой в данной статье, в силу сознательного искажения данных налоговых деклараций, погрешность исходных данных $(\tilde{u}, \delta)$ может оказаться слишком большой и условию (14) сопоставимости по точности решений обратной задачи будут удовлетворять даже очень грубые нейросетевые модели, восстанавливающие скорее шум, чем латентные многомерные функции $Y(\vec{X})$, «зашитые» в данных.

2. Понятие о квазирешении обратной задачи

Определение. Элемент $\tilde{z} \in M$, минимизирующий приданном и функционал $\rho_U(Az, u)$ на компакте *M*, называется *квазирешением* уравнения $Az = u$ на *M*:

$$\rho_U(A\tilde{z}, u) = \inf \rho_U(Az, u). \tag{15}$$

Если *M* – компакт, то квазирешение, очевидно, существует для любого $u \in U$. Если, кроме того, $u \in AM$, то квазирешение $\tilde{z}$ совпадает с обычным (точным) решением уравнения (6). Квазирешение может быть и не одно. В этом случае под квазирешением будем понимать любой элемент из множества квазирешений *D*.

Данное определение даёт широкий простор для построения прикладных обратных задач, в том числе и для разработки НСМ. Если уравнение $Az = u$ может иметь на компакте *M* не более одного решения u и проекция каждого элемента u на множество $N = AM$ единственна, то квазирешение $\tilde{z} : \rho_U(A\tilde{z}, u) = \inf\limits_{z \in M} \rho_u(Az, u)$ единственно и непрерывно зависит от правой части u.

Здесь проекция u понимается в смысле следующего определения. Элемент q из множества *N* называется проекцией элемента u на множество *N* $(q = Pu)$, если:

$$\rho_U(u, q) = \rho_U(u, Q) = \inf_{h \in Q} \rho_U(u, h). \tag{16}$$

Таким образом, при переходе от обычного решения к квазирешению восстанавливаются все три условия корректности обратной задачи (6) по Адамару, т.е. задача нахождения квазирешения уравнения $Az = u$ на компакте *M* является корректно поставленной.

Если уравнение единственности решения уравнения (15) не выполнено, то квазирешения $\{\tilde{z}\}$ образуют, некоторое множество *D* элементов в компакте *M*. В этом случае имеет место непрерывная зависимость множества квазирешений $\tilde{z} \in D$ от правой части в смысле непрерывности многозначных отображений [7].

Практически поиск квазирешения означает использование методов минимизации функционалов (при параметризации *z*-функции многих переменных) на множестве с ограничениями. В случае квадратичной метрики (5) для НСМ [3] в конечномерном пространстве нескольких переменных, удобно положить:

$$\rho^2_{\tilde{U}} = E. \tag{17}$$

Для обратных задач класса интерпретации условно-корректная или обобщённо-корректная постановка исчерпывает проблему построения регуляризирующего алгоритма, поскольку для таких задач можно применять общие алгоритмы. Так любой алгоритм $z_\delta = R(\tilde{u}_\delta, \delta)$ выбора элемента z_δ из множества практической эквивалентности Z_δ при решении приведённого ниже функционального неравенства является регуляризирующим по Тихонову [6]:

$$z_\delta : \{\rho_{\tilde{U}}(Az, \tilde{u}) \le \delta, \quad z \in M\}. \tag{18}$$

Таким образом, основным условием нахождения квазирешения в методе (18) является принадлежность точного решения z_T и квазирешения z_δ к одному и тому же компакту $M \subset Z$. Метрическое пространство *M* называется компактным, если из всякой последовательности в *M* можно извлечь сходящуюся подпоследовательность. Компактное подпространство метрического пространства будем называть также компактным множеством. Компактные пространства (и подпространства) обладают двумя важными свойствами:

- компактное пространство является ограниченным;
- компактное пространство *Y* метрического пространства *X* является замкнутым.

Введение условия компактности в постановку обратной задачи означает практически использование количественной априорной информации об искомом решении. В НСМ множество *U* приходиться вводить искусственным путём. Поскольку при этом требуется обеспечить существование решения на априори заданном компакте *M*, решается вопрос: какими свойствами должно обладать множество *U* для заданного компакта *M*? Если ответ на него получен, хотя бы на качественном уровне, то из множества реальных наблюдений $\tilde{U}$ выделяется подмножество *U*, обладающее нужными свойствами. В [6] сформулировано предложение: «в конкретных обратных задачах при этом решается задача «сглаживания» заданного элемента $\tilde{u}$ ».

В данной статье и в прежних работах автора [3, 4] идея «сглаживания» данных развита и воплощена в форме метода вложенных математических моделей (ВММ), в котором строятся итерационные процедуры «сглаживания» и повышения информативности данных с помощью вспомогательных нейросетевых субмоделей (НССМ).

Замечание. В решении обратной задачи (6) участвует разнородная информация об изучаемом явлении: его НСМ $A(\cdot)$, некоторые общие свойства искомого решения $z \in M$, оценка погрешности данных δ. Если эта информация не согласована, т.е. $\rho_{\tilde{U}}(Az, \tilde{u})$ и $A(\cdot)$ задаются независимо, то может возникнуть ситуация несостоятельности задачи регуляризации (18):

$$\rho_0 = \inf\ \rho_{\tilde{U}}(Az, \tilde{u}) \rangle\ \delta. \quad (19)$$

Связанная с этой ситуацией потеря устойчивости НСМ подробно анализировалась в [3, 4]. Следовательно, для практической регуляризации А.Н. Тихонова по (18) необходимо разработать три алгоритма:

1) *алгоритм* I обеспечения состоятельности задачи регуляризации (18), исключающий ситуацию (19);
2) *алгоритм* II построения подходящей числовой меры оценки погрешности δ для данных и инструмент управления этой погрешностью;
3) *алгоритм* III построения числовой меры оценки качества НСМ и инструмента управления этим качеством.

3. Алгоритм I обеспечения состоятельности задачи регуляризации

Постулируется, что в некоторых пределах интенсивности шума и объема сильнозашумленных вектор-столбцов данных $x_{ij,}\, y_i, i = \overline{1,N}\quad j = \overline{1,n},$ где x_{ij}, y_i - соответствующие значения компоненты вектора входов нейросети (НС) $\vec{X}$ и выхода НС Y в i-том наблюдении, независимо от закона распределения шума существует непрерывная зависимость меры ρ_0 по (19), характеризующей качество аппроксимации в НСМ, от меры оценки погрешности данных δ:

$$\rho_0 = \varphi(\delta). \quad (20)$$

Предлагается следующая концепция разработки методов и алгоритмов обеспечения состоятельности задач регуляризации: *уменьшение числовых мер ошибок эксперимента δ и ошибок аппроксимации ρ_0, должно производиться взаимосвязано с использованием (20), причем числовая мера δ должна быть связана с процедурой управления «сглаживанием» и структурированием данных, в аспекте улучшения качества будущего обучения НС, а числовая мера ρ_0 должна быть связана с управлением качеством аппроксимации восстанавливаемой функции $Y(\vec{X})$ и, соответственно, с прогностическими свойствами сети.*

Управление качеством данных по мере δ предлагается осуществлять на основе вспомогательных нейросетевых субмоделей (НССМ), в которых реализуются следующие итерационные процедуры структурирования данных:

- оптимальная кластеризация;
- оптимальная очистка данных в образованных кластерах;
- «ремонт» сильнозашумленных вектор - столбцов данных с помощью НССМ.

Указанные процедуры, разработанные с использованием общесистемных законов энтропийного равновесия, подавления дисфункций структурируемой системы и фоновой закономерности [8] описаны достаточно подробно в [9], поэтом излагать их здесь не будем.

4. Алгоритм II построения числовой меры оценки погрешности данных на основе байесовского подхода

Краткое изложение формализма байесовского подхода к сравнению моделей содержится в [4].

В общем виде алгоритм II, предлагаемый в настоящей работе можно сформулировать следующим образом: *в качестве числовой меры δ погрешности данных выбирается обобщенный (векторный) мультипликативный критерий Ф, оцениваемый согласно алгоритму I в НССМ путем осреднения в ансамбле НС и определенный на тестовом множестве данных*:

$$\Phi(Y(\vec{X},W)) = E \cdot S \cdot R,\ (\vec{X},Y) \in \Omega^{test}, \tag{21}$$

где $E = \|\hat{y} - y\| / \|y\|$ - ошибка обобщения НСМ, которая имеет смысл относительной нормы ошибок аппроксимации на тестовом множестве наблюдений, не используемых при обучении НСМ в евклидовой метрике E^n; $S = |\hat{y}_\alpha - \hat{y}_\beta| / \|\vec{x}_\alpha - \vec{x}_\beta\|_{E^n}$ - мера сжимающих свойств НСМ (аналог константы Липшица связи «вход - выход» НС); $E = 1 - (r_y)^2$ - мера отклонения коэффициента детерминации от его идеального значения, равного 1 [2]. Ошибка обобщения *E* характеризует прогностические свойства НС: чем меньше *E,* тем ближе расчетные значения $\hat{y}_i$ к экспериментальным *y* на новых наблюдениях. Частный критерий *S* характеризует устойчивость НСМ к вариациям независимых переменных $\vec{x}$: чем меньше *S*, тем меньше «разбегание» траектории $\hat{Y}(\vec{X})$ на новых наблюдениях после обучения НС. *Однако заметим, что при малых S, т.е. при сильных сжимающих свойствах НС-отображения (6), в режиме обучения НС, т.е. в обратной задаче поиска параметров W НСМ оператор A(·) ведет себя как вполне непрерывный (компактный) оператор* [6]*, что является индикатором некорректности обратной задачи.* Критерий *R* характеризует качество аппроксимации «зашитых» в данных истинной зависимости *Y(X)*, т.е. гиперповерхности с помощью нейросетевого отображения $\hat{Y}(\vec{X})$.

Таким образом, обобщенный критерий качества НСМ *Ф* оценивает как точностные и прогностические свойства НСМ, так и ее устойчивость к вариации данных.

Осреднение в ансамбле гипотез $\{h_q\}$ о порождении данных [4] проводится как вычисление среднего арифметического:

$$\overline{\Phi} = \sum_{q=1}^{Q} \Phi_q / Q, \tag{22}$$

где *Q* – число вспомогательных НССМ в байесовском ансамбле.

Алгоритм вычисления критерия *S*, следующий. Номера вектор - строк α и β приравниваются в панельных данных, образуемых налоговыми декларациями к соседним номерам вектор - строк, т.е. близко расположенным точкам по времени $(\alpha \equiv i,\ \beta = i+1,\ i = \overline{1,N})$. Для соседних точек *i* и *i*+1 вычисляется $|\hat{Y}_\alpha - \hat{Y}_\beta|$. Критерий *S* равен:

$$\max_{i \in 1,N} |\hat{Y}_\alpha - \hat{Y}_\beta| / \|\vec{x}_\alpha - \vec{x}_\beta\| \tag{23}$$

5. Алгоритм III построения числовой меры оценки качества НСМ ρ_0 и инструмента управления этим качеством

Конструирование меры ρ_0 осуществлено на основе байесовского подхода [2, 4], т.е. использованы усреднённые оценки на ансамбле априорных гипотез о порождении данных $\{h_q\}$, как и в процедурах предобработки данных. В качестве ρ_0 взято среднее значение вероятности $\overline{P}$ получения в НСМ «плохих» точек, в которых относительная ошибка расчёта Δ_i превышает заданный экспертно уровень ε:

$$i^*: \Delta_i = \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \cdot 100\% \rangle \varepsilon, i = \overline{1, N}; \quad N^* = \sum_{i^*=1}^{m} i^*; \tag{24}$$

$$P_q = (N^* / N)_q; \quad \overline{P} = (\sum_{q=1}^{Q} P_q) / Q \tag{25}$$

В качестве инструментов управления числовой мерой $\overline{P}$ по (24)-(25) выбраны:

- байесовская регуляризация на ансамбле сетей, различающихся архитектурой, видом активационных функций, числом нейронов в скрытых слоях (см. ниже);
- оптимизация параметров обучения сети (шага градиентного спуска, коэффициента «тяжелого веса», начальных весов $W^{(0)}$).

6. Пример построения НСМ для налогового контроля с регуляризацией на основе алгоритмов I, II, III

Для расчётов использовались реальные данные из [1] налоговых деклараций строительных предприятий, закодированных числами. В качестве моделируемой «обобщённой производственной функции» кластера примерно однородных налогоплательщиков была выбрана функция выручки $Y(\vec{X})$. Для образованного по алгоритму I кластера Z^1 наблюдений компоненты вектора $\vec{X}$ независимых переменных имели следующий экономический смысл: X_1 - сумма основных средств; X_2 - себестоимость товаров, продукции, услуг предприятия; X_3 - среднесписочная численность работающих, чел.; X_4 - сумма оборотных активов; X_5 - среднегодовая стоимость облагаемого налогом имущества предприятия; X_6 - коммерческие расходы.

В таблице 1 показан фрагмент исходных данных. Все значения Y и X_j, $j = \overline{1,6}$, кроме X_3, приведены в тыс. руб.

Таблица 1

Исходные данные панельного типа для построения НСМ

Сквозной номер наблюдения i	Код предприятия p^r	Номер квартала (временного интервала) t_i	X_1	X_2	X_3	X_4	X_5	X_6	Y
1	1.1	5715,7	58459	49	47179,6	5676,8	3762,1	3762,1	62106,6
2	1.1	5645,7	13226	49	34079,2	12018	7432,9	7432,9	182534
…	…	…	…	…	…	…	…	…	…
349	3.13	216,93	7257,7	42	3123,2	699,23	1108,6	1108,6	8644,98
350	3.13	21,47	7276	42	1364,87	887,92	1654,4	1654,4	9322,02
351	3.13	211,23	6103	40	2940,89	279,01	1359,1	1359,1	7917,12

Здесь код предприятия обозначен 2^{-x}-значным числом: первая цифра соответствует номеру образованного кластера данных, а вторая – номеру предприятия.

В соответствии с байесовским подходом к регуляризации обучения и обеспечения состоятельности алгоритма регуляризации из [2, 4] была выбрана мета-гипотеза H и априорные гипотезы $\{h_q\}$ о порождении данных, т.е. о виде аппроксимации восстанавливаемой зависимости $Y(\vec{X})$:

H – многослойный персептрон (с алгоритмом обучения обратного распространения ошибки (*backpropagation* (BP)) ($H \equiv \{h_q\}, q = \overline{1,Q}$); (приложение 1).

h_1 - структура MLP с одним скрытым слоем и сигмоидной активационной функцией вида:

$$f(s) = 1/(1+\exp(-as)), \quad a \rangle 0; \tag{26}$$

h_2 - структура MLP с двумя скрытыми слоями и активационной функцией (26) в них;

h_3 - структура с двумя скрытыми слоями и активационной функцией (26) в первом слое и гауссовой во втором слое

$$f(s) = \exp(-s^2/b^2), \quad b \rangle 0; \tag{27}$$

h_4 - структура MLP с одним скрытым слоем и функцией (27) в нём;

h_5 - структура MLP с двумя скрытыми слоями и активационной функцией (11) в них;

h_6 - структура MLP с двумя скрытыми слоями и активационной функцией (26) в первом слое и (27) во втором слое.

В качестве управления оценкой погрешности данных δ в алгоритме предобработки данных были выбраны оптимальные итерационные процедуры кластеризации, очистки образованных кластеров от аномальных точек и «ремонта» наиболее сильно зашумлённого вектор - столбца $\{x_{i2}\}$ согласно алгоритму I [3, 4]. В качестве оценки погрешности данных использовался мультипликативный критерий (21) с его осреднением на ансамбле гипотез $\{h_q\}$ по (22). Данные вычислительных экспериментов показаны в таблице 2 и на рис.1. из [3]. Здесь обозначено: k – номер итерации в процедуре очистки кластера от аномальных точек [9]; n_1, n_2 - количество нейронов в скрытых слоях; Δ – относительная погрешность по (24); A – число «плохих» точек, которые на каждой k-той итерации находятся по условию (24)

(соответствующие $\varepsilon^{(k)}$ приведены в последней строке таблицы 5); $N^{(k)}$ - число примеров на *k*-той итерации, оставшихся после предыдущей итерации.

Эксперимент проводился на сети типа MLP с двумя скрытыми слоями. В качестве моделируемого экономического показателя (выходной величины) рассматривалась *Y* - выручка предприятия, тыс. руб. Всего имелось 351 наблюдение.

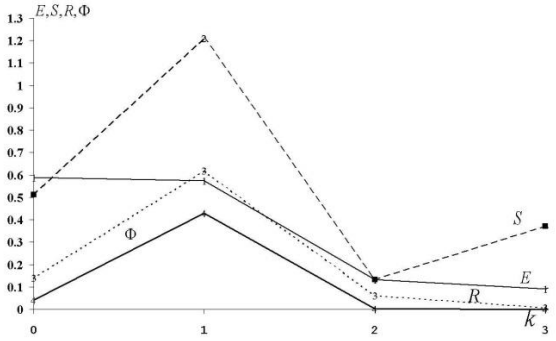

Рис. 1. *Зависимость частных критериев точности E, устойчивости S, детерминированности R и финишного критерия Ф от номера итерации k*

Таблица 2

Сводная характеристика каждой итерации

k	0	1	2	3	4
E	0,5895	0,5735	0,1317	0,0905	0,3181
S	0,5126	1,2113	0,1324	0,3704	0,0889
r	0,9279	0,6177	0,9695	0,9974	0,9352
R	0,139	0,6185	0,0601	0,0053	0,1254
Ф	0,0420	0,4297	0,0010	0,0002	0,0035
n_1	3	2	4	3	2
n_2	4	5	3	3	4
$\bar{\delta}$	149%	39%	29%	29%	50%
A	47	12	5	2	17
N	201	154	142	137	135
ε	100%	100%	100%	100%	100%

Оптимальное число нейронов в первом скрытом слое n_1 и во втором – n_2 подбирались в процессе обучения НС. Активационная функция в скрытых слоях – сигмоид, в выходном – линейная. Число примеров в обучающем множестве составило 80% от общего числа примеров в кластере, в множестве перекрестного подтверждения и тестовом – по 10%. Обобщенные результаты экспериментов представлены на рисунке 1 и в таблице 2, где отражены значения частных критериев *E*, *S* и *R* и обобщённого критерия *Ф* на каждой *k*-ой итерации.

В таблице также приведены значения коэффициента корреляции *r*, числа примеров в кластере N_k, числа аномальных наблюдений A_k, и среднего значения отклонения $\overline{\delta}^{(k)} = \delta_i / N$ на данной итерации. Анализ показывает, что обобщённый критерий *Ф* достиг своего минимального значения на третьей итерации, при этом его значение уменьшилось более чем в 5,89 раз. Последующее выбрасывание аномальных точек при обучении приводит только к ухудшению точности и устойчивости НСМ. Рост критерия Ф в 17,5 раз на 4-й итерации можно объяснить тем, что нарушается условие репрезентативности (23). Таким образом,

сформулированное утверждение обосновано численно, найден номер оптимальной итерации и получен кластер для создания рабочей модели, очищенный от аномальных наблюдений.

Анализ расчётов показывает, что выбранные инструменты регуляризации по методу ВММ весьма эффективны: мера оценки погрешности данных $\overline{\Phi}$ меняется от 0,000028 до 0,021, т.е. в 750 раз! Причём минимум $\overline{\Phi}$ достигается уже на третьей итерации. Чётко проявляется дефицит наблюдений: на четвёртой итерации, где $N^{(4)} = 135$, т.е. уменьшилось на 23,28% по сравнению с нулевой итерацией критерии *E, R, S*, и $\overline{\Phi}$ заметно ухудшаются. Характерно, что на второй итерации изменяется характер сжатия нейросетевого отображения (*S* начинает расти, и после $k \rangle 4$ возможно появление неустойчивости сети – ошибка входа будет «растягиваться» к выходу).

Ниже приведены результаты осреднения на байесовском ансамбле гипотез $\{h_q\}$ по критерию вероятности получения «плохих» точек расчёта (24) – (25). В расчётах использовалась программа NeuroSolutions – 4.0 (демоверсия). Построенные рабочие модели были проверены на адекватность по критерию $\overline{P}$. Экспертно задаваемый уровень ошибки ε возьмём равным 100%. Критическое значение доверительной вероятности в процедуре перекрёстной проверки (CV) P^{GCV} зададим равным 0,75. Значения доверительной вероятности, т.е. отношения числа «плохих» точек к общему числу наблюдений для НСМ типа *q* приведены в таблице 3.

Таблица 3

Результаты выполнения процедуры ОПП

Тип НСМ	НСМ1	НСМ2	НСМ3	НСМ4	НСМ5	НСМ6
P_q	0.8571	0.8889	0.8836	0.7989	0.8889	0.8889

Среднее значение $\overline{P}$ на ансамбле равно 0,8677, что вполне приемлемо для сильнозашумлённых данных.

Таким образом, из таблицы 3 следует, что все шесть типов НСМ прошли байесовскую процедуру перекрёстного подтверждения и на их основе можно синтезировать оптимальный план выездных налоговых проверок (приложение 2).

Выводы:

1. Полученные результаты являются обнадёживающими предпосылками в аспекте обеспечения состоятельности задачи регуляризации для нейросетевых задач с сильным зашумлением данных.
2. Изучение эффективности различных способов предобработки данных и их влияния на свойства регуляризованных нейросетей является востребованным и необходимым. Основополагающий принцип предобработки данных: снижение существующей избыточности всеми возможными способами. А это повышает информативность примеров и, тем самым, качество нейропредсказаний.

ЛИТЕРАТУРА

1. Букаев Г.И., Бублик Н.Д., Горбатков С.А., Сатаров Р.Ф. Модернизация системы налогового контроля на основе нейросетевых информационных технологий. – М.: Наука, 2001. – 344 с.
2. Шумский С.А. Байесова регуляризация обучения: Лекции для школы-семинара «Современные проблемы информатики» (23-25 января 2002 г., Москва). – М.: МИФИ, 2002. – 33 с. (file:// Нейро ОК Интелсофт.htm).
3. Бирюков А.Н. Теоретические основы разработки нейросетевых моделей в системе налогового администрирования. – Уфа: Академия наук РБ, Издательство «Гилем», 2011. – 380 с.
4. Бирюков А.Н. Байесовская регуляризация нейросетевых моделей ранжирования и кластеризации экономических объектов. – Уфа: Академия наук РБ, Издательство «Гилем», 2011. – 292 с.
5. Хайкин С. Нейронные сети: полный курс, 2-ое издание: Пер. с англ. – М.: Издательский дом «Вильямс», 2006. – 1104 с.
6. Горбатков С.А., Полупанов Д.В., Бирюков А.Н., Макеева Е.Ю. Методологические основы разработки нейросетевых моделей экономических объектов в условиях неопределенности - М.: Издательский дом «Экономическая газета», 2012. – 494 с.
7. Тихонов А.Н., Гончарский А.В., Степанов В.В., Ягола А.Г. Численные методы решения некорректных задач. М.: Наука, 1990.
8. Иванов В.К., Васин В.В., Танана В.П. Теория линейных некорректных задач и ее приложения. М.: Наука, 1978.
9. Прангишвили И.И. Системный анализ и общесистемные закономерности. – М.: СИНТЕГ, 2000. – 525 с.
10. Каллан Р. Основные концепции нейронных сетей = The Essence of Neural Networks First Edition. - М.: Вильямс, 2001. - 288 с.
11. Прангишвили И.В. Системный подход и общесистемные закономерности. Серия «Системы и проблемы управления». – М.: СИНТЕГ, 2000, 528 с.
12. Урманцев Ю.А. Общая теория систем: состояние, приложение и перспективы развития. Система, симметрия, гармония. – М.: Мысль, 1988.
13. Ясницкий Л.Н. Введение в искусственный интеллект. - М.: Издат. центр «Академия», 2005. - 176 с.

Рецензент: Статья рецензирована членами редколлегии журнала.

Biryukov Aleksandr Nikolaevich
FGBOU VPO «Bashkir state University»
Russia, Ufa
E-mail: biryukov_str@mail.ru

The method of quasi-solutions for the regularization of neural network models of tax control

Abstract. Issues covered in this article arose from an objective need for research aimed at improving the efficiency of the tax system at the regional level.

The state can not spend more on the collection of taxes, so the structure of taxes and state tax administration system (SNA) to ensure that they collect, must at minimum cost to provide high performance, which is impossible without good information and analytical system of support of decisions. The construction of such a system involves the integration into a single information space of all structural units of the tax and other state bodies. Required common information space created in recent years in the form of electronic data interchange (EDI), developed by the Federal Tax Service of the Russian Federation. This system, which is the basis for the automation of the tax authorities is described in almost all textbooks Russian economic institutions.

However, EDI is one bottleneck - it poorly formatizirovan analysis unit, the core of which should serve as a mathematical model for analyzing the financial condition of the taxpayer, of violations of the tax legislation in the declarations, the synthesis of the optimal plan of field tax audits. Such models should serve as a sufficiently reliable and objective basis for management decision support.

Currently, there is a powerful mathematical tool (universal approximator and clusterer) - neural network. By using neural networks require their learning by example, that from a mathematical point of view, is an ill-posed inverse Hadamard problems like interpretation. This raises the dilemma of finding a compromise between the generalization error model and its robustness (resistance to variations of the data in a given range).

With regard to the neural network model CHA, which differ very noisy data, aggravated in some cases, shortage of observations, research on regularization of neural networks in inverse problems were not carried out.

In the works V.K. Ivanova [8] a rigorous mathematical justification of the two methods for solving ill-posed problems, provided that there is an additional a priori information on the desired solution. If you know that the decision is an element of the set of the compact, he developed a method of quasi-solutions. In this case it is possible and error estimates for approximate solutions. The article develops the idea of a new pre-computer technology (desk) tax audit companies, taxpayers, proposed on the basis of neural network modeling. Use of these models provides the basis for improving the reliability and objectivity of the tax control in the tax authorities, and to improve the effectiveness of field tax audits.

Keywords: neural network; neural network model; Bayesian approach; nested mathematical models; Neural network submodel; the method of quasi-solutions; algorithm.

REFERENCES

1. Bukaev G.I., Bagel N.D. Gorbatkov S.A., Satarov R.F. Modernisation of tax control system based on neural network information technologies. - M.: Nauka, 2001 - 344 p.
2. Shumsky S.A. Bayesian regularization of study: Lectures for school-seminar "Modern problems of informatics" (23-25 January 2002, Moscow). - M.: MEPhI, 2002. - 33 p. (file: // Neuro Intelsoft.htm OK).
3. A.N. Biryukov The theoretical basis for the development of neural network models in the system of tax administration. - Ufa Academy of Sciences of Belarus, Publisher "Guillem", 2011. - 380 p.
4. A.N. Biryukov Bayesian regularization neural network models ranking and clustering of economic projects. - Ufa Academy of Sciences of Belarus, Publisher "Guillem", 2011. - 292 p.
5. Haykin C. Neural networks: a complete course, 2nd ed.: Trans. from English. - M.: Publishing House "Williams", 2006. – 1104 s.
6. Gorbatkov S.A., Polupanov D.V., Biryukov A.N., Makeyev E.Y. The methodological basis for the development of neural network models of economic objects in neopredelennosti - M.: Publishing house «Economic newspaper», 2012. – 494 s.
7. Tikhonov, Goncharsky A.V., Stepanov V.V., Yagola A.G. Numerical methods for solving ill-posed problems. M.: Nauka, 1990.
8. V.K. Ivanov, V.V. Vasin, V.P. Tanana, Theory of linear ill-posed problems and its applications. M.: Nauka, 1978.
9. Prangishvili II System analysis and system-wide patterns. - M.: SINTEG, 2000. - 525 p.
10. Kallan R. The basic concepts of neural networks = The Essence of Neural Networks First Edition. - M.: Williams, 2001. - 288 p.
11. Prangishvili I.V. The systems approach and system-wide laws. A series of "systems and control problems" - M.: SINTEG, 2000, 528 p.
12. Urmantsev Y.A. The general theory of systems: state and prospects of development of the application. The system, symmetry, harmony. - M.: Thought, 1988.
13. Yasnitsky L.N. Introduction to Artificial Intelligence. - M.: Izdat. center "Academy", 2005. - 176 p.