УДК 519.226.3:336.201.2

Удаление противоречивых наблюдений как процедура предрегуляризации нейросетевой модели налогового контроля

С. А. Горбатков[1], И. И. Белолипцев[2], Е. А. Мурзина[3]

[1] sgorbatkov@mail.ru, [2] red7315@gmail.com, [3] murzinaea@ufa.uralsib.ru

[1,2] ФГБОУ ВПО «Финансовый университет при Правительстве Российской Федерации»
[3] ОАО «Уралсиб»

Поступила в редакцию 22.03.2013

Аннотация. Предлагается алгоритм поиска и удаления противоречивых наблюдений из данных. Предлагаемая процедура является одной из процедур регуляризации нейросетевой модели налогового контроля и направлена на повышение устойчивости модели к возмущению входных данных.

Ключевые слова: противоречивые данные; регуляризация обучения; устойчивость; нейросетевая модель.

ВВЕДЕНИЕ

До настоящего времени нейросети не применялись в задачах налогового контроля, кроме как в публикациях авторов статьи и представителей Уфимской школы нейромоделирования. Известна также работа [1], в которой предложен подход к прогнозированию налоговых доходов бюджета региона на основе адаптивной базы знаний иерархического типа с механизмом обучения на основе нейронечеткой сети. В работах [2, 3] авторами был предложен и подробно описан метод вложенных математических моделей (МВММ) для построения гибридной нейросетевой модели (ГНСМ) налогового контроля. Исходной информацией для построения модели являются данные бухгалтерской отчетности, на основе которых рассчитывается ряд показателей, значимо влияющих на моделируемую величину Y, по которой диагностируются нарушения налогового законодательства.

Целью данной работы является исследование вопроса об обеспечении однородности данных для улучшения качества ГНСМ, ее точности и устойчивости. В работе предлагается модифицированный алгоритм удаления из данных локальных неоднородностей (противоречивых вектор-строк). Проблема неоднородности данных характерна для экономических задач и задач налогового контроля в частности. Она вызвана сознательным искажением данных бухгалтерской отчетности налогоплательщиками в целях уменьшения налогооблагаемой базы. Как показывает практика, чаще всего искажаются расходные статьи бухгалтерского баланса, то есть налогоплательщик «дописывает» несуществующие расходы с целью снижения налогооблагаемой базы. Если локальные неоднородности оставить в данных, то нейросеть будет искажать восстанавливаемую многофакторную зависимость, скрытую в данных.

НЕОБХОДИМОСТЬ УДАЛЕНИЯ ПРОТИВОРЕЧИВЫХ ДАННЫХ

Предлагаемый алгоритм удаления противоречивых данных относится к процедурам предрегуляризации модели и проводится до непосредственного обучения байесовского ансамбля нейросетей [3]. Формируется исходная база данных $\langle \vec{x}_i, y_i \rangle, i = \overline{1,N}$, где $\vec{x}_i$ – векторы значений входов нейросети, y_i – значения моделируемой величины.

Суть «противоречивости данных» в следующем: двум близким по некоторой числовой мере вектор-строкам $\vec{x}_\alpha$ и $\vec{x}_\beta$ могут соответствовать существенно отличающиеся значения выходной величины y_α и y_β. То есть пара вектор-строк $\vec{x}_\alpha$ и $\vec{x}_\beta$ «растягиваются» при нейросетевом отображении, что ухудшает качество обучения сети и негативно влияет на устойчи-

вость модели к изменению входных факторов (рис. 1).

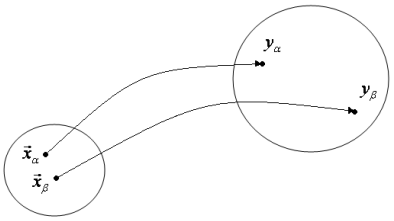

Рис. 1. Иллюстрация зависимости $y(\vec{x})$, скрытой в данных в области локальной неоднородности

Для выявления локальных неоднородностей предлагается использовать аналог константы Липшица

$$L_{\alpha,\beta} = \left\| y_\alpha - y_\beta \right\|_{E_n} / \left\| x_\alpha - x_\beta \right\|_{E_n}, \qquad (1)$$

где $\vec{x}_\alpha, \vec{x}_\beta$ – близкие по евклидовой норме вектор-строки, y_α, y_β – соответствующие им значения выходной величины, E_n – n-мерное евклидово пространство; $\|\bullet\|$ – норма в E_n. Все записи исходной базы данных предварительно сортируются в порядке возрастания $\|\vec{x}_i\|$. Тогда если $L_{\alpha,\beta} > 1$, то это значит, что двум близким по норме векторам $\vec{x}_\alpha$ и $\vec{x}_\beta$ соответствуют существенно различающиеся значения y_α и y_β, что может говорить о сознательном искажении данных налогоплательщиком.

Большие значения величины $L_{\alpha,\beta}$ могут быть вызваны двумя причинами: 1) для двух близких по норме векторов $\vec{x}_\alpha$ и $\vec{x}_\beta$ норма $\left\| y_\alpha - y_\beta \right\|$ велика, что говорит о сознательном искажении данных; 2) два вектора $\vec{x}_\alpha$ и $\vec{x}_\beta$ очень близки по норме, т. е. $\left\| \vec{x}_\alpha - \vec{x}_\beta \right\|$ очень мало при существенно большем значении $\left\| y_\alpha - y_\beta \right\|$. В этом случае $L_{\alpha,\beta} >> 1$. Введем дополнительное условие. Для пары близких по норме векторов $\vec{x}_\alpha$ и $\vec{x}_\beta$ должно выполняться условие:

$$\left\| \vec{x}_\alpha - \vec{x}_\beta \right\| / \left\| \vec{x}_\alpha \right\| \cdot 100\% \geq \zeta, \zeta \approx 1\%. \quad (2)$$

Вектор-строки, удовлетворяющие критериям (1) и (2), удаляются из исходной базы данных и не участвуют в последующем обучении нейросети.

При обнаружении пары вектор-строк, для которых $L_{\alpha,\beta} > 1$, без дополнительного исследования неясно, какая из них «растягивает» нейросетевое отображение. В [2] предлагалось удалять из базы данных обе строки – при этом вместе с противоречивыми, мешающими обучению данными удаляются и «хорошие», информативные вектор-строки. Такой подход допустим, если данных достаточно для качественного обучения. Однако для задач налогового администрирования характерен дефицит наблюдений, и, кроме того, алгоритм построения ГНСМ налогового контроля на этапе обучения нейросетей предусматривает дополнительную чистку базы данных для обеспечения лучшей точности и устойчивости модели. В такой ситуации на этапе предобработки данных желательно сохранить как можно больше наблюдений. Для решения этой задачи предлагается следующее: при обнаружении пары вектор-строк $\vec{x}_s$ и $\vec{x}_{s+1}$, для которых $L_{s,s+1} > 1$, вычисляется критерий Липшица в соседних точках $L_{s-1,s+1}$. Если $L_{s-1,s+1} < 1$, то это означает, что для пары векторов с номерами s и $s+1$ именно строка с номером s приводит к тому, что $L_{s,s+1} > 1$. Вектор-строка $\vec{x}_s$ признается противоречивой, удаляется из базы данных и не участвует в дальнейшем обучении. И наоборот, если $L_{s-1,s+1} > 1$, то противоречивой является строка с номером $s+1$.

В результате выполнения предложенной процедуры из базы данных удаляются сознательно искаженные налогоплательщиком наблюдения, которые могут негативно повлиять на качество обучения нейросети. Сформированная база данных $\langle \vec{x}_i, y_i \rangle, i = \overline{1, N_0}$ будет более однородной и обученная на ней нейросеть будет менее чувствительна к изменениям входных факторов. Впоследствии противоречивые вектор-строки, не участвовавшие в обучении, предъявляются обученной сети для выявления налогоплательщиков-нарушителей и окончательного синтеза плана выездных налоговых проверок.

ИТЕРАЦИОННЫЙ АЛГОРИТМ УДАЛЕНИЯ ПРОТИВОРЕЧИВЫХ ВЕКТОР-СТРОК

1. Для всех строк исходной базы данных $\langle \vec{x}_i, y_i \rangle, i = \overline{1, N}$ вычисляются евклидовы нормы векторов $\vec{x}_i$

$$\left\|\vec{x}_i\right\| = \sqrt{\sum_{j=1}^{n} x_{ij}^2}, i \equiv \overline{1, N},$$

где j – номер компоненты вектора $\vec{x}_i$

2. Все строки базы данных ранжируются в порядке возрастания норм $\left\|\vec{x}_i\right\|$

$$\left\|\vec{x}\right\|_1, \left\|\vec{x}\right\|_2, ..., \left\|\vec{x}\right\|_s, ..., \left\|\vec{x}\right\|_{N-1}, \left\|\vec{x}\right\|_N, s = \overline{1, N},$$

где s – номер члена ранжированного ряда

3. Вдоль ряда вычисляются нормы разностей

$$\left\|\vec{x}_1 - \vec{x}_2\right\|, \left\|\vec{x}_2 - \vec{x}_3\right\|, ..., \left\|\vec{x}_s - \vec{x}_{s+1}\right\|, ... \left\|\vec{x}_{N-1} - \vec{x}_N\right\|.$$

4. Для каждой пары векторов $\left\|\vec{x}_S\right\|$ и $\left\|\vec{x}_{S+1}\right\|$ вычисляется величина $L_{s,s+1}$ по критерию (1).

5. Проводится итерационный процесс поиска и удаления противоречивых наблюдений. Определяются пары вектор-строк, для которых $L_{s,s+1} > 1, s = \overline{1, N_k - 1}$, где N_k – количество наблюдений в базе на *k*-й итерации.

6. Организуется вложенный цикл: для каждой пары вектор-строк с $L_{s,s+1} > 1$ проверяется условие (2). Если оно не выполняется, переходим к следующей паре вектор-строк с $L_{p,p+1} > 1$. Если же условие (2) выполняется, вычисляется $L_{s-1,s+1}$. Если $L_{s-1,s+1} < 1$, то строка с номером s удаляется из базы данных, в противном случае удаляется строка с номером $s+1$.

7. В конце каждой *k*-й итерации проверяется условие репрезентативности данных по правилу

$$N_k > \xi n, \qquad (3)$$

где N_k – количество оставшихся наблюдений на k -й итерации; ξ – коэффициент запаса; n – количество входных факторов. В некоторых работах [4] предлагается эмпирическое правило в соответствии с которым количество наблюдений должно быть в 10 раз больше количества входов нейросети, если относительная погрешность в сети порядка 10 %. Но с учетом того, что на этапе обучения будет проводиться дополнительная чистка базы данных, коэффициент запаса ξ предлагается принять в интервале $\xi \approx 15...20$. Если условие (3) не выполняется, фиксируется состояние, достигнутое на предыдущей $(k-1)$-й итерации, процесс удаления противоречивых точек останавливается.

Шаги 3–7 повторяются до тех пор, пока в базе данных не останется пар строк, для которых $L_{\alpha,\beta} > 1$, либо пока процедура не будет прервана по правилу (3).

РЕЗУЛЬТАТЫ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

Для построения исходной базы данных $\langle \vec{x}_i, y_i \rangle, i = \overline{1, N}$ использовались данные квартальной бухгалтерской отчетности 24 сельскохозяйственных предприятий в период 2006–2009 гг. Компоненты векторов $\vec{x}_i$ представляют собой относительные величины, характеризующие финансовое состояние предприятия (аналоги коэффициентов финансового анализа) [5]. В качестве моделируемой величины Y было взято отношение суммарных затрат к величине выручки предприятия. Всего в базе имелось 276 наблюдений.

После удаления противоречивых вектор-строк в базе данных осталось 231 наблюдение. Используя алгоритм, изложенный в [2, 3], в базе остается только 211 наблюдений. То есть предлагаемая модификация алгоритма позволила «сохранить» 20 информативных наблюдений, которые должны способствовать лучшему обучению сетей байесовского ансамбля.

Для оценки эффективности предложенной выше процедуры на разных наборах данных была обучена нейросеть (персептрон, 3 скрытых слоя, активационная функция в скрытых слоях – тангенс; в выходном слое – линейная). Результаты обучения представлены в табл. 1.

Таблица 1

Показатели качества НСМ для разных наборов данных

	Исходная база данных, 276 наблюдений	База данных, очищенная от противоречивых вектор-строк	
		231 наблюдение	211 наблюдений
MSE_{test}	0,268	0,162	0,105
$NMSE_{test}$	0,868	0,868	0,935
E	0,404	0,330	0,290
S	0,719	0,429	0,467
J	0,291	0,141	0,135

Качество обучения сети оценивалось по обобщенному критерию *J*, представляющему собой произведение частных критериев:

$$J = E \cdot S; \qquad (4)$$

$$E = \left\|\hat{Y}_i - Y_i\right\| / \left\|Y_i\right\|;$$

$$S = \frac{1}{N} \sum_{i=1}^{N} \frac{\left|\hat{y}_i - \hat{y}_{i+1}\right|}{\left\|\vec{x}_i - \vec{x}_{i+1}\right\|}; i = \overline{1, N}; N \in \Omega^{\text{test}}. \qquad (5)$$

Первый множитель в формуле (4) является ошибкой обобщения и характеризует точность и прогностические свойства сети. Критерий *S*

характеризует устойчивость сети к изменению входных факторов.

Как видно из табл. 1, после удаления противоречивых вектор-строк показатель J уменьшился более чем вдвое, то есть предложенную выше процедуру можно признать состоятельной.

При построении рабочей ГНСМ налогового контроля и выполнении всех процедур регуляризации, предусмотренных МВММ, ошибку обобщения E удалось снизить до величины 4–5 %, а величину J – до 2–3 % [3].

На этапе синтеза оптимального плана выездных налоговых проверок удаленные противоречивые данные вновь предъявлялись обученным сетям. Как показали вычислительные эксперименты [2], в этих точках наблюдаются значительные отклонения вида

$$\delta_i = \left| y_i - \hat{y}_i \right| / y_i,$$

где y_i – фактические наблюдаемые значения выходной величины; $\hat{y}_i$ – моделируемое значение, что является признаком нарушения налогового законодательства данным налогоплательщиком. Таким образом, ГНСМ уверенно идентифицирует нарушителей налогового законодательства. Результаты моделирования сравнивались с данными реально проводившихся налоговых проверок [3]. Из 12 предприятий, допускавших грубое нарушение налогового законодательства, в окончательный план проверки попало 10.

ВЫВОДЫ

Разработан алгоритм поиска локальных неоднородностей в данных, который применяется на этапе предрегуляризации нейросетевой модели налогового контроля. В вычислительных экспериментах доказана необходимость и эффективность этой процедуры в целях повышения точности и устойчивости модели к возмущению входных данных.

СПИСОК ЛИТЕРАТУРЫ

1. **Фаттахов Р. В., Черняховская Л. Р., Осипова Н. В.** Применение нейро-нечеткой системы прогнозирования налоговых доходов бюджета региона // Нейрокомпьютеры: разработка, применение. 2007. № 10. С. 75–80.

2. **Нейросетевое** математическое моделирование в задачах ранжирования и кластеризации в бюджетно-налоговой системе регионального и муниципального уровней: монография / С. А. Горбатков, Д. В. Полупанов, А. М. Солнцев, И. И. Белолипцев, М. В. Коротнева, С. А. Фархиева, О. Б. Рашитова. Уфа: РИЦ БашГУ, 2011. 224 с.

3. **Горбатков С. А., Белолипцев И. И., Фархиева С. А., Полупанов Д. В.** Приближенный метод байесовской регуляризации и двухступенчатая оценка адекватности гибридной нейросетевой модели // Нейроинформатика – 2011: сб. науч. тр. XIII Всерос. науч.-техн. конф. Науч. сессия НИЯУ МИФИ–2011: В 3 ч. Ч. 2. М.:НИЯУ МИФИ, 2011. С. 144–154.

4. **Ежов А. А., Шумский С. А.** Нейрокомпьютинг и его применение в экономике и бизнесе: учебник / под ред. проф. В. В. Харитонова. М.: Изд. МИФИ, 1998. 224 с.

5. **Шевченко И. В., Халафян А. А., Васильева Е. Ю.** Создание виртуальной клиентской базы для анализа кредитоспособности российских предприятий // Финансы и кредит. 2010, № 1 (385). С. 13–18.

6. **Горбатков С. А., Белолипцев И. И.** Очистка данных наблюдений как процедура предрегуляризации нейросетевой модели налогового контроля // Социальная ответственность бизнеса: теория, методология, практика: мат-лы II Всерос. науч.-практ. конф. Уфа: ВЗФЭИ, 2012. С. 114–118.

ОБ АВТОРАХ

ГОРБАТКОВ Станислав Анатольевич, проф. каф. математики и информатики Уфимского филиала. Дипл. инж. по электрифик. пром. предпр. (ТПИ, 1960). Д-р техн. наук по упр. в техн. системах (МИЭМ, 1991). Иссл. в обл. нейросетевого моделир. в техн. и экон. системах с зашумлением данных

БЕЛОЛИПЦЕВ Илья Игоревич, преп. той же каф. Дипл. мат.-экономист (УГАТУ, 2004). Иссл. в обл. нейросет. моделир. в экон. системах с зашумлением данных

МУРЗИНА Елена Анатольевна, магистрант той же каф. Дипл. менедж.-экон. по работе с недвиж. имуществом (Башкирск. гос. ун-т, 2006).

METADATA

Title: Removal of conflicting data as a procedure of regularization of neural network model of tax control.

Authors: S. A. Gorbatkov[1], I. I. Beloliptsev[2], E. A. Murzina[3].

Affiliation:

[1,2] Financial University under the Government of the Russian Federation (Financial University), Russia.

[3] Public Corporation «Uralsib», Russia.

Email: [1] sgorbatkov@mail.ru.

Language: Russian.

Source: Vestnik UGATU (scientific journal of Ufa State Aviation Technical University), vol. 17, no. 5 (58), pp. 110-114, 2013. ISSN 2225-2789 (Online), ISSN 1992-6502 (Print).

Abstract: An algorithm for finding and removing conflicting observations from the data is proposed. The proposed procedure is a procedure of regularization neural network model of tax control and is aimed at improving stability of the model to a perturbation of the input data.

Key words: Conflicting data; regularization of training; sustainability neural network model.

References (English transliteration):

1. R. V. Fattahov, L. R. Chernahovskaya, and N. V. Osipova, "The use of neuro-fuzzy system for forecasting tax revenues in the region," *Neyrokomputery: Razrabotka, Primenenie* (Neurocomputers: development, application), no. 10 (37), pp. 75-80, 2007.
2. S. A. Gorbatkov, D. V. Polupanov, A. M. Solntsev, I. I. Beloliptsev, M. V. Korotneva, S. A. Farhieva, and O. B. Rashitova, *Neural network mathematical modeling in ranking and clustering of fiscal system, regional and municipal levels*, (in Russian). Ufa: Bashkir State University, 2011.
3. S. A. Gorbatkov, I. I. Beloliptsev, S. A. Farhieva, and D. V. Polupanov, "Approximate Bayesian regularization method and two-step evaluation of the adequacy of hybrid neural network model," in *Proc. 13th Workshop on Neuroinformatics*, vol. 2, pp. 144-154, 2011.
4. A. A. Yeghov and S. A. Shumsky, *Neurocomputing and its application in economics and business,* (in Russian). Moscow: Moscow Engineering Physics Institute, 1998.
5. I. V. Shevchenko, A. A. Khalafyan, and E. J. Vasilyeva, "Creating a virtual customer base for the analysis of the creditworthiness of Russian enterprises," *Finance and Credit*, no. 1 (385), pp. 13-18, 2010.
6. S. A. Gorbatkov and I. I. Beloliptsev, "Cleaning the data as the procedure of regularization of neural network model of tax control," in *Proc. 2nd Workshop on Social Responsibility of Business: Theory, Methodology, Practice*, Ufa, Russia, 2012, pp. 114-118.

About authors:

GORBATKOV, Stanislav Anatolievich, Prof., Dept. of Mathematics and Informatics. Dipl. Engineer electrification of industrial enterprises (Tomsk Polytechnic Ins., 1960). Dr. of Tech. Sci. (MIEM, 1991).

BELOLIPTSEV, Iliya Igorevich, assistant, Dept. of Mathematics and Informatics. Dipl. Mathematician & Economist (UGATU 2004).

MURZINA, Elena Anatolievna, Postgrad. (PhD) Student, Dept. of Finance & Credit. Dipl. Manager of the real estate (Bashkir State Univ., 2006).