Уфа: УГАТУ Вестник УГАТУ 2001. № 2 (4). С. 32–37

УДК 519.86

ОБ АДЕКВАТНОСТИ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ НАЛОГОВОГО КОНТРОЛЯ И УПРАВЛЕНИЯ В УСЛОВИЯХ ДЕФИЦИТА НАБЛЮДЕНИЙ

С. А. ГОРБАТКОВ, Н. Т. ГАБДРАХМАНОВА

Уфимский филиал Всероссийского заочного финансово-экономического института
Тел: (3472) 51 08 23 E-mail: bublik@ufacom.ru

Исследуются вопросы использования кластеризации в качестве усиливающего блока к нейросетевым математическим моделям экономических объектов. Из учета специфики объекта моделирования (налогооблагаемых предприятий) предлагается эффективный способ кластеризации по аналогии построения вложенных отрезков. Рассмотрен реальный пример построения адекватных нейромоделей по данному способу

Нейросеть; адекватная модель; кластер; сложный объект

ВВЕДЕНИЕ

Предметом исследования являются вопросы повышения адекватности нейросетевых моделей налогооблагаемых предприятий. Построение математической модели налогооблагаемых объектов (ОН) с хорошими ассоциативными свойствами является основной подзадачей задачи автоматизации системы налогового контроля и управления (СНКУ) и оптимизации на ее основе доначислений в бюджет. Нейросетевые модели ОН позволяют заменить существующую в СНКУ операцию камеральных проверок.

В СНКУ камеральным проверкам подвергаются все предприятия с момента предоставления отчетности в налоговую инспекцию в сроки, установленные законодательством. В ходе камеральной проверки инспектор анализирует предоставленную финансовую отчетность для выявления возможных нарушений в налоговых отчислениях. По результатам камеральных проверок преимущественно по эвристическим критериям и алгоритмам формируется план документальных проверок, т. е. определяется перечень предприятий для документальных проверок. Экспертный характер камеральной проверки обуславливает возникновение ряда недостатков в составлении плана.

Предлагается автоматизировать процедуру камеральных проверок на основе математического моделирования. В работе решалась задача получения «эталонных» адаптивных моделей предприятий налогоплательщиков. На основе имеющейся информации по бухгалтерским отчетам предприятий налогоплательщиков за предыдущие периоды (кварталы) построена математическая модель предприятия для прогноза экономического показателя, в частности выручки. Разность (отклонение) между расчетным значением выходной величины по модели и выходным экономическим показателем предприятия из отчетной документации используется как показатель достоверности финансовой отчетности предприятия. На основе этих вычислений составляется оптимальный план документальных проверок предприятий.

При построении математической модели налогооблагаемых предприятий необходимо учесть ее специфические свойства: неопределенность внешней и внутренней среды; сложность самого объекта; недостаточность и зашумленность информации. Под сложностью объекта понимается его сложная структура, стохастический и динамический характер процессов в объекте, большое число входных факторов (десятки и сотни), мультиколлинеарность вектора входных факторов. Наиболее эффективным инструментарием из арсенала новейших информационных технологий для рассматриваемых задач являются нейросети.

1. ПОСТАНОВКА ЗАДАЧИ

Пусть имеется некоторое множество ОН $\{S_j\}$, $j = 1, 2, ..., m$, где m число объектов в достаточно однородном кластере. Для каждого объекта имеются ряд наблюдений с номерами $i = 1, 2, ..., n$, где зафиксированы значения моделируемого показателя $Y_j(t_k)$, а также вектор входных факторов $\{\vec{x}(t_k)\}_j$ в различные дискретные моменты времени $\{t_k\}$, $k = 1, 2, ..., l$. Требуется построить нейросетевое отображение F

$$Y(t) = F(\vec{x}_t), \qquad (1)$$

где $\vec{x}_t$ — вектор входных воздействий или экзогенных (внешних переменных), $Y(t)$ — выходной параметр характеристики системы (эндогенных или внутренних переменных), так, чтобы обеспечить допустимую погрешность прогноза, оцениваемую доверительным интервалом $\delta_\alpha Y(t_{\text{пр}})$:

$$Y(t_{\text{пр}}) \approx \hat{Y}(t_{\text{пр}}) \pm \delta_\alpha Y(t_{\text{пр}}). \qquad (2)$$

2. РЕШЕНИЕ ЗАДАЧИ

С помощью нейросетей аппроксимировалось динамическое отображение вида (1). Сам оператор $F()$ считался стационарным. Влияние вектора внешних возмущающих воздействий было отнесено к шумам. Важнейшим показателем качества нейросетевой модели являются ее ассоциативные свойства, т. е. относительная погрешность вычисления выходной величины за пределами тестового множества:

$$\delta_* = \max_{j=\overline{1,m}} \left| \frac{Y_j(t_k) - \hat{Y}_j(t_k)}{Y_j(t_k)} \right|, \qquad (3)$$

где $t_k = t_{\text{пр}}$; $Y_j(t_k)$ — выручка j-го предприятия в момент времени t_k, т. е. в строке матрицы планирования расчетов с номером $i > N_1$, где N_1 — номер последней точки тестового множества; $\hat{Y}_j(t_k)$ — расчетное (вычисленное с помощью нейросети) значение выручки j-го предприятия в момент времени t_k.

Опыт нейросетевого моделирования СНКУ позволяет сформулировать некоторые предложения по улучшению ассоциативных свойств нейросети:

1) Выбрать наиболее информативные показатели для включения в вектор входных данных нейросети.

Как указывалось, для исходной задачи характерна большая размерность вектора входных факторов. Это может отрицательно повлиять на качество настраиваемой нейросети. Для хорошей настройки нейросети рекомендуется соотношение

$$N \cong \frac{n}{E}, \qquad (4)$$

где N — число обучающих примеров, n — число входных факторов, E — относительная допустимая погрешность настройки нейросетевой модели. В моделях налогового контроля существует проблема дефицита обучающих примеров, так как число наблюдений ограничено числом однородных по финансовому состоянию налогоплательщиков в группе и числом временных отсчетов наблюдения (кварталов за 2 года наблюдения). Требуемый уровень погрешности $E = (0{,}1...0{,}15)$ [1]. Следовательно, для построения нейросети с хорошими ассоциативными качествами необходимо сократить размерность входного вектора. Задача решается с использованием корреляционного и факторного анализа [2].

2) Предварительно обработать исходные данные.

Необходимо провести предварительную обработку данных, т. е. исключить из файла данных грубые ошибки. Данные аномальные точки находятся путем эвристического анализа входных данных.

3) Использовать принцип обобщенного перекрестного подтверждения [1].

4) Повысить однородность выборки путем решения задачи кластерного анализа.

Качество получаемой нейросети зависит от двух взаимно противоречивых характеристик:

- числа учитываемых признаков, характеризующих разнообразие самих объектов и влияние внешней среды;
- однородность объектов в группе (кластере), охватываемой данной моделью, которую можно оценить различными числовыми мерами.

Суть взаимной противоречивости состоит в том, что если образовать один кластер с достаточно большим числом качественных и количественных входных переменных (факторов), то возникают две неприятности. Первая состоит в том, что часть факторов все равно ускользнет из поля зрения исследователя и, следовательно, будет отнесена к шумам. Вторая заключается в том, что модель с очень большим числом факторов (сотни) оказывается «неподъемной» на стадии построения, верификации.

Другой крайний случай — очень большое число кластеров.

Может быть достигнута высокая однородность входных данных. Однако резко снижается число обучающих примеров при построении нейромодели, что плохо сказывается на качестве модели.

Для исходной задачи кластеризация исходных данных проводилась поэтапно. Вначале с использованием метода морфологического анализа [6] по основным признакам объекта (классификационным, количественным, качественным и порядковым) выделены кластеры, которые на первой стадии моделирования считаются достаточно однородными. На более поздних стадиях моделирования решалась задача дробления кластера на подкластеры. Дробление может быть реализовано по различным признакам и с использованием различных методов.

В данной статье предлагается один из способов разбиения достаточно однородного кластера на подкластеры. При кластеризации использовался метод k-средних, реализованный в пакете Statistica 5.0 [3, 4]. В качестве классификационного признака использовался функционал $\operatorname{th} \bar{Y}$. Кластеризация позволила построить адекватные модели.

3. АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ

1. Построить нейромодель всего кластера. Если погрешность выше допустимых норм, то положить $i = 0$ и перейти к 2, иначе — к 4.

2. Методом k-средних разбить весь кластер на два подкластера. Вычислить $i := i+1$. Обозначить подкластер, лежащий справа на числовой $\operatorname{th} Y$, через $A(i)$, а слева — $B(i)$ и построить для каждого кластера нейромодель. Если нейромодель для подкластера $B(i)$ адекватна, присвоить номера полученным подкластерам и перейти к шагу 4, иначе — к шагу 3.

3. Из двух подкластеров для подкластера $A(i)$ присвоить очередной номер, для подкластера $B(i)$ выполнить шаг 2.

4. Конец.

Результаты построения нейросети можно считать положительными при выполнении следующих условий:

1) средняя квадратическая ошибка обучения нейросети меньше 0,01;

2) средняя квадратическая ошибка перекрестной проверки меньше 0,01;

3) нормализованная средняя ошибка тестирования меньше 0,1;

4) доверительная вероятность нейросети P больше 0,85.

Для тестирующей выборки относительная погрешность вычисления настроенной нейросети

$$\delta_j = \left| \frac{Y_j(t_k) - \hat{Y}_j(t_k)}{Y_j(t_k)} \right|. \qquad (5)$$

Количество точек, в которых произошла большая ошибка вычисления $\delta_j > \varepsilon$, где ε — назначаемый экспертно уровень допустимых отклонений, не должно превышать 15 % всего объема тестирующей выборки).

4. ЧИСЛОВОЙ ПРИМЕР

В качестве апробации принципа ОПП построим нейромодель группы приближенно однородных торговых предприятий Уфы. В кластере из 78 предприятий были представлены три района Уфы: Советский, Кировский, Орджоникидзевский. Для каждого предприятия использованы наблюдения за 4 последних квартала, т. е. общее число наблюдений $78 \times 4 = 312$. Для обеспечения приемлемой достоверности исходных данных в основном использовались данные отчетной бухгалтерской документации после их коррекции в результате документальной проверки [1].

На основе экспертных оценок и предварительных цифровых экспериментов выбраны следующие наиболее информативные входные факторы: X_1 — основные средства производства; X_2 — себестоимость реализованных товаров и услуг; X_3 — среднесписочная численность сотрудников; X_4 — денежные средства; X_5 — среднегодовая стоимость имущества; X_6 — коммерческие расходы предприятия; X_7 — запасы готовой продукции; X_8 — дебиторская задолженность; X_9 — кредиторская задолженность; X_{10} — нераспределенная прибыль; X_{11} — торговый оборот; X_{12} — сумма износа основных средств производства. Выбран выходной показатель Y — выручка предприятия. Все значения независимых и зависимых переменных после нормировки становятся безразмерными. Таким образом, $Y(t) = f\left(\bar{X}(t)\right)$, где $\bar{X}(t) = (X_1(t), ..., X_{12}(t))$, t — номер временного интервала. Функция f аппроксимируется нейросетью.

Для задач аппроксимации хорошо подходят нейросети с архитектурой многослойный персептрон. Нейросети реализовывались с помощью пакета NeuroSolutions фирмы NeuroDimension Inc. (NS) [5].

4.1. Нейромодель исходного кластера

Первоначально по исходным данным была построена одна нейромодель следующей конфигурации: тип — многослойный персептрон, число скрытых слоев — 2; число нейронов 1-го слоя — 23; число нейронов 2-го слоя — 14; активационная функция скрытого слоя — гиперболический тангенс; активационная функция выходного слоя — линейная функция; алгоритм обучения — критерий квадрата ошибки. Результаты построения нейросети представлены на рис. 1 и табл. 1.

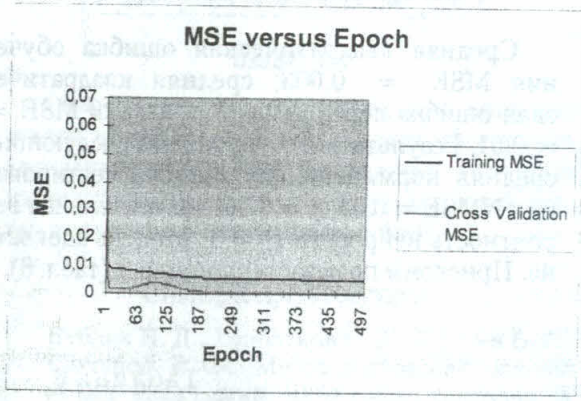

Рис. 1

Таблица 1

Best Networks	*Training*	*Cross Validation*
Epoch #	500,0000	354,0000
Minimum MSE	0,0052	0,0005
Final MSE	0,0052	0,0005

Результаты построения нейросети следующие: средняя квадратическая ошибка обучения MSE = 0,0052; средняя квадратическая ошибка перекрестной проверки MSE = = 0,0005. Эти результаты свидетельствуют о том, что обучение завершилось успешно. Затем нейросеть тестировалась. Результаты тестирования представлены на рис. 2 и в табл. 2 и 3. В колонках табл. 3 приняты следующие обозначения: Y — декларированное значение выручки (руб); $\hat{Y}$ — вычисленное значение выручки (руб); δ — относительная погрешность прогнозирования, вычисленная по формуле (5). Результаты тестирования следующие: средняя нормированная ошибка тестирования NMSE = 0,02256, коэффициент корреляции документированной и вычисленной выходной величины r = 0,9163, доверительная вероятность нейросети $P \approx 0{,}42$. Ясно, что такая модель не может быть признана пригодной в практическом использовании.

С целью повышения адекватности нейросети данные кластера разбиты на подкластеры по алгоритму, предложенному выше.

Внизу приведены краткие результаты построения нейросетей подкластеров.

На первом шаге кластер разбит на два кластера A1 и B1.

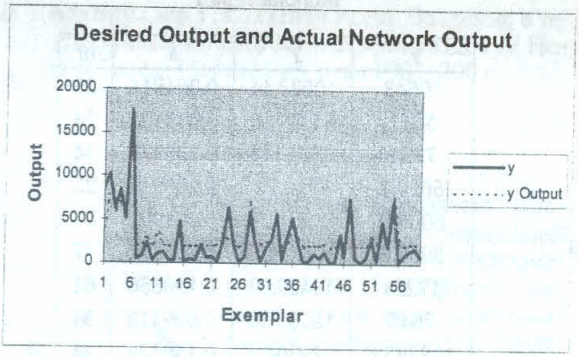

Рис. 2

Таблица 2

Performance	*y*
MSE	2306593,0000
NMSE	0,2256
MAE	1284,7893
Min Abs Error	29,6492
Max Abs Error	4620,1128
r	0,9163

Таблица 3

Результаты вычисления нейромодели всего кластера

Y	$\hat{Y}$	δ	pr
406,6	2014,293	3,953992	73
725,1	2015,766	1,779983	39
2333	2660,907	0,140552	53
369	2142,164	4,805323	11
1166,8	3283,646	1,814232	43
1197,3	2077,414	0,735082	69
314,8	1993,185	5,331592	45
281,6	1968,693	5,991096	22
4793,8	3875,706	0.191517	50
147	1951,476	12,27534	20
459,9	1984,5	3,315068	78
284,9	1978,652	5,945074	78

4.2. Нейромодель подкластера A1

Средняя квадратическая ошибка обучения MSE = 0,0034; средняя квадратическая ошибка перекрестной проверки MSE = 0,023. Эти результаты свидетельствуют о том, что обучение завершилось успешно. Затем нейросеть

тестируется. Результаты тестирования следующие: средняя нормированная ошибка тестирования NMSE = 0,05, r = 0,99, доверительная вероятность нейросети $P \approx 0{,}93$. Модель адекватна. Подкластеру присвоим номер 1 (табл. 4).

Таблица 4

Результаты вычисления нейромодели подкластера 1

Y	$\hat{Y}$	δ	pr
10643	10023,44	0,061811	25
5641	8285,768	0,319194	34
7354	6525,418	0,126978	44
6630,3	8960,28	0,260034	25
6005,3	6395,427	0,061001	44
5950,6	5904,092	0,007877	77
17224	14469,62	0,190356	61
7645	15583,36	0,509413	34
102243,6	120304,9	0,150129	24
12551	18365,38	0,316594	74

4.3. Нейромодель подкластера B1

Средняя квадратическая ошибка обучения: MSE = 0,001; перекрестной проверки: MSE = 0,007. Результаты тестирования: NMSE = 0,06, r = 0,8, доверительная вероятность нейросети $P \approx 0{,}8$, большие погрешности принадлежат объектам с маленькой выручкой. Модель не адекватна.

Подкластер B1 вновь разбит на подкластеры, которые обозначены A2 и B2.

4.4. Нейромодель подкластера A2

Средняя квадратическая ошибка обучения MSE = 0,0003; перекрестной проверки MSE = = 0,04. Результаты тестирования: NMSE = = 0,06, r = 0,99, доверительная вероятность нейросети $P \approx 1$. Модель адекватна. Подкластеру присвоим номер 2 (табл. 5).

Таблица 5

Результаты вычисления нейромодели подкластера 2

Y	$\hat{Y}$	δ	pr
1813	2425,299	0,337727	29
2581,1	2477,857	0,04	60
4611,1	4574,08	0,008029	52
4414,3	4282,048	0,02996	52
6106	5604,898	0,082067	55
7519	6724,62	0,10565	29
3977	3652,995	0,08147	50
4715	4673,805	0,008737	5

4.5. Нейромодель подкластера B2

Средняя квадратическая ошибка обучения MSE = 0,008; перекрестной проверки MSE = = 0,25. Результаты тестирования NMSE = = 0,49, r = 0,7, доверительная вероятность нейросети $P \approx 0{,}7$, большие погрешности принадлежат объектам с маленькой выручкой. Модель не адекватна.

Подкластер B2 разбит на подкластеры A3 и B3.

4.6. Нейромодель кластера A3

Средняя квадратическая ошибка обучения MSE = 0,003; средняя квадратическая ошибка перекрестной проверки MSE = = 0,01. Результаты тестирования следующие: средняя нормированная ошибка тестирования NMSE = 0,09, r = 0,96, доверительная вероятность нейросети $P \approx 1$. Модель адекватна. Присвоим подкластеру номер 3 (табл. 6).

Таблица 6

Результаты вычисления нейромодели подкластера 3

Y	$\hat{Y}$	δ	pr
2023,2	1783,003	0,118721	15
2386	2398,717	0,00533	53
1242,7	1368,379	0,101134	33
2340	2362,201	0,009488	66
1769	1681,81	0,049288	41
2105	2187,464	0,039175	4
1748	1911,884	0,093755	14
2086	1700,169	0,184962	42

4.7. Нейромодель кластера B3

Средняя квадратическая ошибка обучения MSE = 0,001; средняя квадратическая ошибка перекрестной проверки MSE = = 0,002. Результаты тестирования следующие: средняя нормированная ошибка тестирования NMSE = 0,05, r = 0,97, доверительная вероятность нейросети $P \approx 1$. Модель адекватна. Присвоим подкластеру номер 4 (табл. 7).

Таким образом, вместо одной модели были построены 4 нейромодели 1, 2, 3, 4, при этом результаты тестирования показали, что все 4 нейромодели адекватны.

Таблица 7

Результаты вычисления нейромодели подкластера 4

Y	$\hat{Y}$	δ	pr
379,5	373,8646	0,01485	17
507	598,4532	0,180381	19
573,2	509,1809	0,111687	78
225,6	228,5098	0,012898	65
631	482,141	0,23591	38
366,1	392,1212	0,071077	31
978,9	834,8677	0,147137	54
183,5	217,0746	0,182968	18
664	701,6781	0,056744	70
359	316,7926	0,117569	40

ВЫВОДЫ

Синтез кластерного анализа и нейросетевого моделирования позволяет обеспечить адекватность моделей в сложных условиях изменчивости самого объекта и внешней среды, а также дефицита наблюдений.

СПИСОК ЛИТЕРАТУРЫ

1. **Бублик Н. Д., Горбатков С. А., Колбин Б. Г., Саттаров Р. Ф.** Методологические основы новых технологий налогового контроля и управления для юридических лиц на основе нейросетевых технологий // Стохастические модели объектов налогообложения. Статистический анализ сложных экономических объектов: Сб. науч. тр. ВЗФЭИ. Вып. 5. Уфа: Баш. территор. ин-т проф. бухгалтеров, 2000. С. 40–67.
2. **Айвазян С. А. Мхитарян В. С.,** Прикладная статистика и основы эконометрики // М.: ЮНИТИ, 1998. С. 1022.
3. **Боровиков В. П., Боровиков И. П.** Statistica — Статистический анализ и обработка данных в среде Windows // М.: Информ.-издат. дом «Филинъ», 1998. 608 с.
4. **Сошникова Л. А., Тамашевич В. Н., Уебе Г., Шефер М.** Многомерный статистический анализ в экономике // М.: ЮНИТИ-ДАНА, 1999. 598 с.
5. **Логовской А. С.** Зарубежные пакеты: современное состояние и сравнительные характеристики // Нейрокомпьютер. 1998. № 1. С. 13–25.
6. **Альтшуллер Г. С.** Найти идею. Введение в теорию решения изобретательских задач // Новосибирск: Наука. Сиб. отд., 1991. 209 с.

ОБ АВТОРАХ

Горбатков Станислав Анатольевич, зав. региональной каф. математики и информатики ВЗФЭИ (Уфимск. филиал). Дипл. инж. по автоматике и телемеханике (Томск. политехн. ин-т, 1960). Д-р техн. наук по управлению в технических системах (защ. в МИЭМ, 1990). Академик Академии инженерных наук РФ. Исследования в области информационных технологий и моделирования экономических систем

Габдрахманова Неля Талгатовна, ст. преп. кафедры высшей математики ВЗФЭИ. Дипл. математик (БГУ, 1982). Исследования в области нейросетевого моделирования