

Алгоритм ранжирования налогоплательщиков на основе общесистемных закономерностей асимметрии и неполного подавления побочных дисфункций структурирования информационной системы

© 2010 А.Н. Бирюков

кандидат экономических наук

Башкирский государственный университет

Стерлитамакский филиал

E-mail: guzsa@ufamts.ru

Вероятностный критерий ранжирования объектов налогового контроля по числовой мере искажения ими отчетной документации и внесение в этот критерий эвристической априорной информации с использованием доверительных интервалов для отклонений между расчетными, полученными в рабочей нейросетевой модели (НСМ), и декларированными значениями моделируемого показателя, позволяют повысить достоверность процедуры ранжирования. Эта идея базируется на общесистемном законе асимметрии, а также на общесистемной закономерности неполного подавления дисфункций структурируемой информационной системы.

Ключевые слова: нейросетевая модель (НСМ), нейронная сеть (НС), синтез плана выездных проверок (СПВП), налоговый контроль.

В общей теории систем¹ известно, что при образовании системы (в нашем исследовании это система налогового контроля) элементы системы обладают большим количеством свойств, одни из которых при формировании связей могут подавляться, а другие, наоборот, усиливаться. Однако подавление системой незначимых (паразитных) свойств, как правило, не бывает полным, поэтому в системе возникают не только “полезные” функции, но и “дисфункции”. Так, например, в НСМ аппроксимации производственной функции объекта налогового контроля полезное взаимодействие (многоуровневое иерархическое структурирование, “эксплуатация” свойств нелинейной взаимосвязи компонента вектора входных факторов и обучения НСМ в условиях сильной зашумленности) порождает неоднозначную (вероятностную) оценку моделируемого показателя. Это практически приводит к неопределенности в интерпретации больших относительных отклонений между расчетными (“эталонными”) и декларированными значениями моделируемого показателя $Y(\bar{X})$.

Действительно, в силу вероятностного характера оценки случайной величины Y в НС даже хорошо обученная, адекватная и имеющая высокую точность НСМ “имеет право” на существование небольшого числа точек тестового множества, где имеются большие отклонения. Число этих точек задается пользователем при обучении НСМ, и его нельзя сделать равным 0, ибо время обучения может стать недопустимым. Тогда не-

ясно, чем вызваны большие отклонения в данных точках: ошибками расчета НС или это случай искажения деклараций на фоне условного математического ожидания? Должен быть предусмотрен некий защитный (иммунный) механизм подавления указанных выше дисфункций математико-информационной модели. В качестве такого механизма может быть предложен вероятностный критерий ранжирования.

Действительно, использование отклонения δ как меры недостоверности отчетной документации имеет следующие недостатки²:

1) специфика предметной области такова, что наиболее информативны в аспекте увеличения налоговых сборов субъекты, у которых отклонения являются результатом сознательного систематического искажения информации, т.е. необходимо учитывать предысторию каждого g -го объекта налогового контроля;

2) возникают случайные искажения показателей декларации;

3) имеется некоторая вероятность появления больших случайных ошибок обобщения, обусловленная самим алгоритмом НС (квадратичный критерий аппроксимации) и ее применением.

Отметим, что возможны следующие случаи появления больших отклонений δ : 1) большое отклонение δ вызвано кратковременными случайными дестабилизирующими факторами в момент прогнозных оценок $Y_{ТС}$; 2) большое от-

клонение δ обусловлено систематическим сознательным искажением отчетной документации.

Для снижения чувствительности δ к помехам, т.е. для подавления дисфункций, предлагается ввести меру риска, получаемую методами математической статистики, учитывающую предысторию конкретного объекта налогового контроля по следующему алгоритму. В формулу $\rho_0 = \varphi(\delta)$ критерия ψ_g ранжирования налогоплательщиков по степени ожидаемых доначислений введено смещение:

$$\tilde{\delta}_{gt} = M\{\delta_{gt}\} + U_{gt_0}, \quad (1)$$

где U_{gt} - полуширина доверительного интервала для отклонения δ_{gt} ;
 $M\{\delta_{gt}\}$ - математическое ожидание по времени;

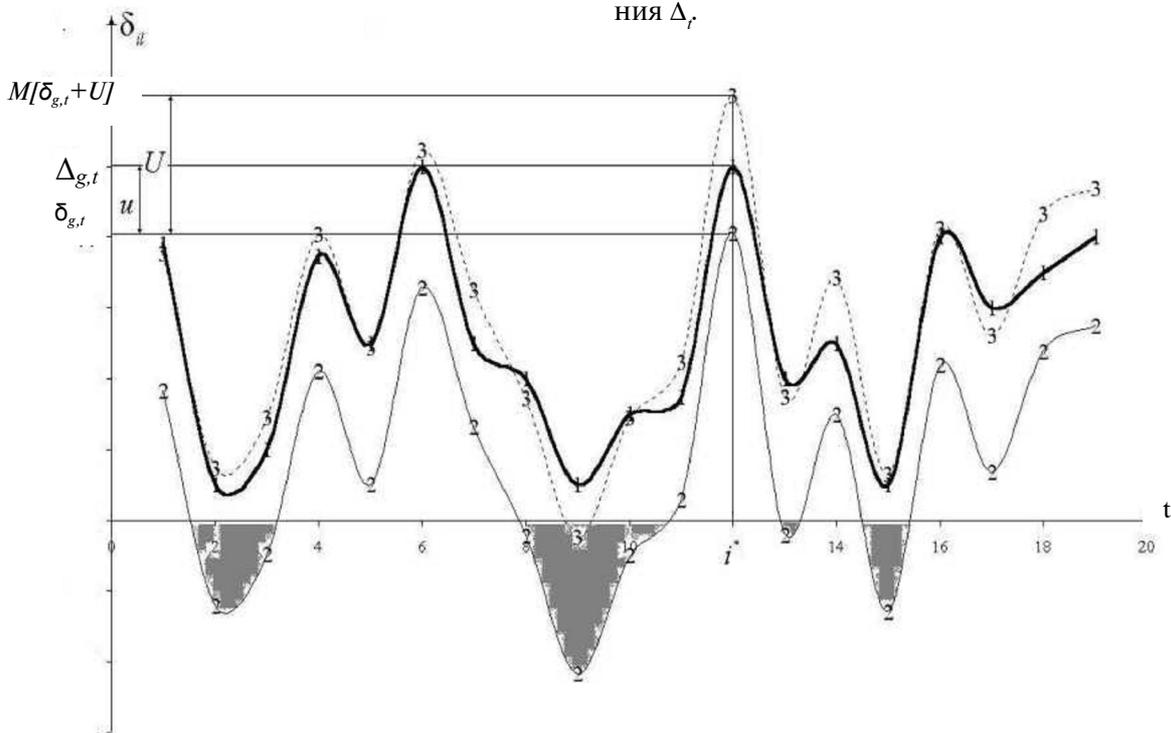


Рис. Использование концепции асимметрии в алгоритме ранжирования

t_0 - момент времени получения оценки. Поясним смысл добавления величины U_{gt} к отклонению δ_{gt} (см. рисунок).

Рассмотрим величину неизвестного нам истинного значения выходной случайной моделируемой величины \bar{Y}^{true} . Тогда $\Delta_t = \frac{y_t^{true} - y_t}{y_t}$ -

неизвестное нам значение отклонения между истинным и декларированными значениями выходной случайной величины Y .

По смыслу задачи наиболее вероятно, что налогоплательщик стремится исказить налогооблагаемую базу в сторону ее уменьшения. Не-

рационально платить налогов больше, чем это требуется согласно законодательству. Отсюда вытекает, что $\Delta_t \geq \delta_t$. Тогда введение "доверительного интервала" для отклонений приподнимает кривую δ_t на величину полуширины доверительного интервала U_t . На рисунке кривая (1) - это истинное значение отклонения Δ_t ; кривая (2) - расчетное значение отклонения δ_t ; кривая (3) - верхняя граница масштабированного смещенного отклонения $\tilde{\delta}_t$;

$$\tilde{\delta}_t = M[\delta_t] + U_t, \quad (2)$$

где $M[\delta_t]$ - математическое ожидание от $\{\delta_t\}$.

Смещение отклонения δ_t на величину полуширины доверительного интервала U_t приближает отклонение δ_t в сторону истинного значения Δ_t .

В алгоритме ранжирования, согласно $\rho_0 = \varphi(\delta)$, используется смещенное на U_t значение отклонения $\tilde{\delta}_{g,t_0}$ и вычисляется вероятность того, что усредненное на отфильтрованном байесовском ансамбле отклонение δ_g^* будет больше числа $\tilde{\delta}_g$:

$$\tilde{\delta}_g = M\left\{\delta_{t,g}^* \right\}_{t=t_1}^{t=t_0} + U_{g,t_0}. \quad (3)$$

Отклонения, лежащие ниже оси абсцисс, не участвуют в последующей процедуре ранжирования.

Тем самым, согласно общесистемной закономерности асимметрии, в алгоритм ранжирова-

ния вносится “асимметрия”, что позволяет внести в этот алгоритм ценную априорную информацию, приближающую его к реальным условиям отбора.

В (1) $P(\bar{\delta}_g^* > \tilde{\delta}_g)$ - вероятность того события, что усредненное отклонение $\bar{\delta}_g^*$ в момент t_0 будет больше выборочного среднего по времени $M\{\delta_{g,t}\}$ с учетом его смещения на полуширину доверительного интервала U_{g,t_0} . В момент времени t_0 осуществляется ранжирование налогоплательщика, т.е. это момент проверки, соответствующий последнему кварталу подачи декларации налогоплательщиком. Третий множитель в формуле (1) - M_g , это экспертно задаваемый коэффициент масштаба g -го налогоплательщика, который может быть введен по следующей формуле:

$$M_g = \frac{H_g}{\bar{H}}, \quad (4)$$

где H_g - налогооблагаемая база g -го налогоплательщика;

\bar{H} - та же величина, усредненная в кластере по всем его элементам и по времени наблюдения $t \in [t_1; t_N]$.

Ту же точку зрения о введении вероятностного критерия при ранжировании налогоплательщиков разделяет и автор работы³, который в качестве модели отклонений $\delta(\bar{x})$ предлагает регрессионную модель, а в качестве критерия (1) - модель бинарного отклика (“пробит” (probit) или “логит” (logit)). Автор иллюстрирует предложенную модель линейной множественной регрессии $\hat{y}(\bar{x}, \bar{b})$ на искусственно синтезированном примере, т.е. строит имитационную модель, не совпадающую с реальными данными⁴.

Вначале опишем сущность и цель процедуры ранжирования налогоплательщиков.

Критерии (1) для различных предприятий-налогоплательщиков ранжируются в порядке убывания, т.е. составляется вариационный ряд:

$$\psi_{g,1} \geq \dots \geq \psi_{g,n} \geq \dots \geq \psi_{g,G^*} \quad (5)$$

Для члена вариационного ряда ψ_g индекс $v = \overline{1, G^*}$ означает его порядковый номер в вариационном ряду, g - фиксируемый номер налогоплательщика в исходной базе данных.

Задача синтеза множества проранжированных налогоплательщиков ставится следующим образом: в соответствии с ψ -критерием присвоить каждому налогоплательщику ранг, показываю-

щий степень нарушения им налогового законодательства; найти совокупность номеров налогоплательщиков Θ , планируемых в дальнейшем для проведения выездной проверки, обеспечивающей максимум функции цели (ожидаемых доначислений Δc), при ограничениях на затраты, связанные с деятельностью налоговых органов при выездных проверках.

Формализованной записи этой задачи предположим несколько замечаний.

Функционал качества ранжирования Ψ . В принципе, ожидаемую сумму доначислений Δc можно выразить через критерий ψ_g в явном виде

$$\Delta c = \varphi(\psi_g) \cdot \delta = \overline{1, G^*}, \quad (6)$$

где G^* - количество налогоплательщиков в группе крупных нарушителей, число G^* может определяться по числу проверяющих бригад при проведении выездных проверок;

φ - некоторая нелинейная функция, связывающая отклонения декларируемых значений дискриминантной (моделируемой) функции Y от расчетных Y с величиной до начислений Δc .

Действительно, критерий ψ_g по (1) есть не что иное, как ожидаемое значение (с учетом вероятности его отклонения от среднего) моделируемой дискриминантной функции Y . Если в качестве Y выбрана величина платежей по какому-нибудь из налогов, например *налог на имущество* или *НДС*, то функция $\varphi(\psi_g)$ есть просто сумма (либо экспертно взвешенная сумма) критериев ψ_g на множестве G^* налогоплательщиков, отобранных в план:

$$\Delta c = \sum_{v=1}^{G^*} \psi_{g,v}, \quad g = \overline{1, G^*}. \quad (7)$$

В этом случае функционал качества отбора Ψ тождественно совпадает с Δc и для получения функции $\varphi(\cdot)$ не требуется построения специальной экономической модели.

Однако для других видов налогов, например *на прибыль*, *единый социальный налог*, в ряде случаев (например, для розничной торговли) более информативна модель дискриминантной функции ранжирования Y , косвенно связанная с налогооблагаемой базой. Такой функцией Y может быть, например, выручка⁵.

В этом случае алгоритм вычисления функции связи $\varphi(\cdot)$ по (6) сложен и громоздок. Он будет содержать ряд логических условных операторов начислений (или льгот) в процентах для различных секторов экономики, что неоправданно усложняет модель СПВП.

Исходя из того, что главной целью модели СПВП является не точное вычисление доначис-

лений (это будет сделано в процессе самих выездных проверок по анализу первичной документации), а ранжирование налогоплательщиков для получения информации, куда следует ехать с проверкой, оценка суммы доначислений Δc может носить условный (приближенный) характер и фигурировать в алгоритме отбора неявно. Таким образом, функция $\varphi(\cdot)$ в (4) может носить неявный характер. Это существенно упрощает построение функционала качества ранжирования. Выберем в качестве его величину:

$$\Psi = \sum_{v=1}^{G^*} \psi_{g,v} \rightarrow \max_g, \quad (8)$$

где $\psi_{g,v}$ - критерий ранжирования для g -го налогоплательщика с номером v в вариационном ряду (3).

В силу эвристических соображений ясно, что, если дискриминантная функция Ψ выступает информативной с точки зрения выявления нарушителей налогового законодательства, то существует неявная связь Δc с ψ_g согласно (4), причем увеличение критерия ранжирования Ψ в (6) соответствует увеличению ожидаемых доначислений Δc .

Функция доначислений. Из всех реальных ограничений учтем главные - ограничения по затратам на выездные налоговые проверки. Другие затраты (на поддержание информационного обеспечения технологии компьютерных проверок, цензурирование синтезированного плана и др.) носят постоянный характер, т.е. не зависят от вида плана проверок и могут не учитываться в задаче оптимизации.

Тогда затраты на выездные проверки будем приближенно считать пропорциональными числу m выездных налоговых бригад.

Ограничение запишется просто:

$$m \leq G^*, \quad (9)$$

где G^* - число субъектов, отобранных для первоочередной выездной проверки.

Формулировка задачи синтеза плана выездных проверок. Требуется найти множество Θ таких номеров g налогоплательщиков, чтобы кри-

терий оптимизации Ψ был максимальным при наложенных ограничениях:

$$\Theta = \left\{ g^* : \Psi = \sum_{v=1}^m \psi_{g,v} \rightarrow \max_g \right\}. \quad (10)$$

Таким образом, задача условной оптимизации (6), (7) сведена к задаче безусловной оптимизации (10), где ограничение $m \leq G^*$ непосредственно введено в формулу для вычисления функционала Ψ .

Поисковый алгоритм. Формулировка задачи оптимизации (10) допускает применение простого поискового алгоритма:

1. Все члены $\{\psi_{g,v}\}$ с помощью, например, инструмента *MS Excel*.

Данные | Сортировка располагаются в порядке убывания, т.е. образуется вариационный ряд:

$$\psi_{g,1}, \dots, \psi_{g,v}, \dots, \psi_{g,G} \quad g = \overline{1, G}. \quad (11)$$

2. Редуцируются первые m членов вариационного ряда (11) по условию $m \leq G^*$.

3. Вычисляется сумма $\Psi = \sum_{v=1}^m \psi_{g,v}$ редуци-

рованных членов ряда. Это и есть решение задачи оптимизации плана выездных проверок, т.е. определения множества номеров Θ по (10).

¹ Тихонов А.Н., Леонов А.С., Ягола А.Г. Нелинейные некорректные задачи. М., 1995.

² См.: Бирюков А.Н. Обобщение метода вложенных математических моделей на основе байесовского подхода к регуляризации задач нейросетевого моделирования налогового и финансового контроля // Экономика и управление. 2010. □ 1. С. 85-89; Бирюков А.Н. Особенности постановки задач моделирования бюджетных процессов на региональном и муниципальном уровне, ориентированных на конечный результат // Экономика и управление. 2010. □ 2. С. 80-85.

³ Теоретические основы разработки технологии налогового контроля и управления / Н.Д. Бублик [и др.]. Уфа, 2004.

⁴ Там же.

⁵ Там же.

Поступила в редакцию 15.10.2010 г.