Литература

1. *Габасов Р., Кириллова Ф.М., Павленок Н.С.* Синтез оптимальных обратных связей в классе инерционных управлений//АиТ, № 2, 2003. с. 22 – 49
2. *Балашевич Н.В., Габасов Р., Кириллова Ф.М.* Численные методы программной и позиционной оптимизации линейных систем управления // ЖВМ МФ. 2000. Т.40. № 6. С. 838 – 859.

ПОСТРОЕНИЕ МОДЕЛИ ПОВЕДЕНЧЕСКОГО СКОРИНГА С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ DATA MINING

А. Н. Никитюк

ВВЕДЕНИЕ

Одним из основных видом деятельности банка является кредитная деятельность, обеспечивающая высокую доходность всех активов, однако сопровождающаяся повышенным риском. С увеличением объемов кредитования актуализируются задачи управления кредитным риском банка. Разработка методов оценки и механизма регулирования кредитных рисков связана с задачей классификации многомерных данных.

Целью данной работы является сравнительный анализ точности классификации заемщиков банка с использованием нейронных сетей *(neural networks)* и деревьев решений *(decision trees)*. На этапе предварительного статистического анализа обучающей выборки также исследовалась проблема выбора информативных признаков. Для тестирования использовались реальные данные по заемщикам немецкого коммерческого банка.

1. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ НАБЛЮДЕНИЙ И ПОСТАНОВКА ЗАДАЧИ

Пусть имеется некоторое множество из n клиентов банка $X = (x_1, x_2, \ldots, x_n)^T \in \Re^{nN}$, и задано множество номеров $Y = (y_1, y_2, \ldots, y_L)^T \in \Re^L$, каждый из которых соответствует определённому классу заемщиков $\{\Omega_i\}$, $i = \overline{1, L}$. Классы Ω_i представляют собой непересекающиеся подмножества множества X, и клиенты из разных классов существенно отличаются по степени надежности. Каждый k-й имеющийся заемщик банка характеризуется набором индивидуальных признаков, из которых образован N-мерный вектор факторов $x_k = (x_{k1}, x_{k2}, \ldots, x_{kN})^T \in \Re^N$, где x_{ki} – признаки (анкетные и поведенческие), характеризующие k-го клиента[3].

Задача поведенческого скоринга: для каждого k-го заемщика банка из множества X спрогнозировать номер y_j одного из классов на основании совокупности имеющихся факторов x_k.

В данной работе рассмотрен случай, когда $L = 2$. Будем полагать, что: Ω_G – класс надежных заемщиков, обладающих высокой степенью платежеспособности; Ω_B – проблемные заемщики, неадекватно оценивающие свои возможности и обладающие низкой степенью платежеспособности.

2. ИНСТРУМЕНТЫ DATA MINING

2.1 Нейронные сети

Для решения описанной задачи классификации из семейства нейронных сетей был выбран *многослойный персептрон Розенблатта* – многослойная нейронная сеть прямого распространения сигнала, в которой входной сигнал преобразуется в выходной, проходя последовательно через несколько слоёв. Сеть состоит из трех слоёв. Нейроны каждого слоя соединяются с нейронами предыдущего и последующего слоя по принципу «каждый с каждым». Работа многослойного персептрона описывается формулами:

$$NET_{jl} = \sum_{i=1}^{N_{l-1}} w_{ijl} \cdot x_{ijl}, \; l = \overline{1,3}, \; j = \overline{1, N_l},$$

$$OUT_{jl} = F\left(NET_{jl}\right), \; l = \overline{1,3}, \; j = \overline{1, N_l},$$

$$x_{ij(l+1)} = OUT_{il}, \; l = \overline{1,3}, \; j = \overline{1, N_{l+1}},$$

i – номер входа, j – номер нейрона в слое, l – номер слоя, N_l - количество нейронов в слое, x_{ijl} – i-й входной сигнал j-го нейрона в слое l, w_{ijl} – весовой коэффициент i-го входа нейрона номер j в слое l, F – логистическая активационная функция, NET_{jl} - сигнал NET j-го нейрона в слое l, OUT_{jl} - выходной сигнал нейрона[2].

В качестве значений входных нейронов используются значения вектора-факторов x_k. Обрабатывающих элементов скрытого слоя также может быть сколько угодно. Выходной слой представляет цель, которая должна быть спрогнозирована.

Пусть задано множество пар векторов $\left\{X^s, T^s\right\}$, $s = \overline{1,S}$, где X^s – входной пример, T^s – известное решение для этого примера. Совокупность пар $\left\{X^s, T^s\right\}$ составляют обучающее множество. Величина S – количество элементов в обучающем множестве – должна быть достаточной

для обучения сети, чтобы сформировать набор весовых коэффициентов сети, дающий нужное отображение $X \to Y$.

На выходном слое находится один нейрон, значение которого определяет, к какому классу относится клиент. Результат имеет бинарный тип (хороший/плохой кредит)[1].

2.2 Деревья решений

Деревья решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение. Для построения дерева решений был выбран *алгоритм CART (Classification and Regression Tree)*. В алгоритме CART каждый узел дерева решений имеет двух потомков. На каждом шаге построения дерева правило, формируемое в узле, делит заданное множество примеров на две части: часть, в которой выполняется правило и часть, в которой правило не выполняется. Для выбора оптимального правила используется функция оценки качества разбиения.

Алгоритм CART использует индекс Gini который оценивает «расстояние» между распределениями классов:

$$GINI(c) = 1 - \sum_j p_j^2$$

где c – текущий узел, а p_j – вероятность класса j в узле c. Данный индекс основан на идее уменьшения неопределенности в узле.

Правило разбиения. Алгоритм CART работает с числовыми и категориальными атрибутами. В каждом узле разбиение может идти только по одному атрибуту. Если атрибут является числовым, то во внутреннем узле формируется правило вида $x_{ki} \le c$. Если атрибут относится к категориальному типу, то во внутреннем узле формируется правило $x_{ki} \in V(x_{ki})$, где $V(x_{ki})$ – некоторое непустое подмножество множества значений переменной x_{ki} в обучающем наборе данных.

Механизм отсечения. Метод заключается в получении последовательности уменьшающихся деревьев: находят любую пару узлов с общим предком, которые могут быть объединены, т.е. отсечены в родительский узел без увеличения ошибки классификации. Так получается дерево, имеющее такую же стоимость как *исходное,* но менее ветвистое, чем *исходное*.

3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Для построения модели использовались данные по прошлым клиентам. Каждый клиент характеризуется 19-тью признаками. К анкетным относятся 17 признаков, к поведенческим относятся 2 признака. Из них непрерывными являются 4 признака, все остальные категориальные. Для анализа платежеспособности заемщиков банка были отобраны 10 наиболее информативных предикторов.

Всего имеется данные по тысячи клиентов. Для каждого клиента определена бинарная переменная «кредитоспособность». Эта переменная включает информацию о том, привлекателен или нет рассматриваемый клиент. При этом 30% относится к проблемным заемщикам, а остальные 70% – к платежеспособным. Процент невыплат по всей совокупности данных составляет приблизительно 3%.

Статистические оценки вероятностей ошибочной классификации в процентах (ошибок первого и второго рода $\hat{P}_1$, $\hat{P}_2$, а также безусловной вероятности ошибки $\hat{P}$) для двух методов представлены в таблице. Под ошибкой классификации первого рода понимается вероятность отнесения заемщика из класса Ω_B в класс Ω_G, соответственно под ошибкой второго рода понимается вероятность отнесения заемщика из класса Ω_G в класс Ω_B.

Таблица

Оценки вероятностей ошибочной классификации

Метод классификации	$\hat{P}_1$	$\hat{P}_2$	$\hat{P}$
Нейронная сеть	10.00	7.80	8.32
Дерево решений CART	15.53	12.99	13.50

Таким образом, метод классификации на основе нейронной сети обладает более высокой точностью по сравнению с деревом решений CART.

Литература

1. Руководство по кредитному скорингу // под ред. Элизабет Мэйз; пер. с англ. И.М. Тикота; науч. ред. Д.И. Вороненко. – Минск: Гревцов Паблишер, 2008. – 464 с.
2. *Ежов А. А., Шумский С. А*. Нейрокомпьютинг и его применения в экономике и бизнесе. Учебное пособие. – Москва: 1998. – 236 с.
3. *Бэстенс Д.-Э., Ванденберг В.-М., Вуд Д.* Нейронные сети и финансовые рынки. Учебное пособие. – Москва: Научное издательство 1997. – 222 с.