Дюличева Ю.Ю. **УДК 519.7**

МОДЕЛИРОВАНИЕ КРЕДИТНОГО СКОРИНГА НА ОСНОВЕ РЕШАЮЩИХ ДЕРЕВЬЕВ

Разработка и внедрение автоматизированных систем оценки кредитоспособности заемщиков является актуальным и перспективным направлением для продвижения банков Украины на рынке потребительского кредитования. По данным Нацбанка Украины [1] «просроченная задолженность по кредитам, предоставленным физическим лицам за 9 месяцев 2006 года увеличилась на 94%, а темпы роста кредитования физических лиц выросли на 86,4%». С наступлением мирового финансового кризиса наблюдалось снижение числа кредитов и рост уровня невозврата кредитов. В современных условиях эффективная работа банка невозможна без внедрения автоматизированных скоринговых систем, оценивающих риск невозврата кредитов и вероятность мошенничества со стороны физических лиц.

Перечислим основные типы современного скоринга [1]: application-скоринг ориентирован на оценку кредитоспособности заемщиков; collection-скоринг ориентирован на разработку комплекса мероприятий по работе с заемщиками, неспособными своевременно вернуть кредит до момента передачи таких заемщиков в коллекторское агенство; behavioral-скоринг ориентирован на исследование динамики кредитного счета заемщика; fraud-скоринг ориентирован на выявление мошенников среди потенциальных заемщиков. Application-скоринг и fraud-скоринг подразумевает разработку модели на основе технологий интеллектуального анализа данных Data Mining типа деревьев решений (леса решений), нейронных сетей, генетических алгоритмов, байесовских сетей и т.п., а также комбинирования этих моделей. В работе [2] обосновывалась эффективность применения логических алгоритмов и, в частности, лесов решений для оценки кредитоспособности заемщиков на основе кредитных историй.

Целью данной работы является разработка алгоритма синтеза леса решений с высокой обобщающей способностью и исследование эффективности применения такой модели для разработки автоматизированной системы application-скоринга.

Сложность переборной задачи построения решающего дерева (дерева решений) с минимальным числом листьев обосновывает существование целого класса алгоритмов синтеза деревьев и лесов решений. Любой алгоритм из этого класса направлен на поиск компромисса между сложностью структуры решающего дерева и точностью распознавания объектов. В работе [3] был предложен алгоритм синтеза леса решений, реализующий такой компромисс. Несмотря на высокую обобщающую способность модели, предложенной в работе автора [3], оставался открытым вопрос о подборе ранга ветвей каждого дерева решений.

Рассмотрим модификацию алгоритма синтеза леса решений с определением ранга ветвей (числа внутренних вершин) на основе «голосования» критериев редукции деревьев решений.

1. BuildTree(d_i)
2. For each vertex v from d_i
   {
   estimation1(v) = ComputePEP(v);
   estimation2(v) = ComputeCCP(v);
   estimation3(v) = ComputeMEP(v);
   }
3. For each vertex v from d_i
   {
   v_pruning = majority_vote(estimation1(v),estimation2(v),estimation3(v));
   }
4. PruningTree(d_i)
5. Rejection_set($\cap d_i$), priority_vector ($\cap d_i$)
6. if (Rejection_set($\cap d_i$) $\neq \varnothing$) {
   i++;
   goto 1;
   }

Процедура BuildTree(d_i) реализует построение решающего дерева с помощью энтропийного критерия, процедуры ComputePEP(v), ComputeCCP(v), ComputeMEP(v) реализуют стратегию редукции:

- критерий PEP [4] основан на сравнении числа ошибок в вершине v и числа ошибок в поддереве $T(v)$ с учетом стандартной ошибки для поддерева, вычисленной в предположении биномиального распределения ошибок

$$error(v) \le error(T(v)) + SE(error(T(v))), \quad (1)$$

где $SE(error(T(v))) = \left(error(T(v)) \cdot (n(v) - error(T(v)))/n(v)\right)^{1/2}$, $n(v)$ – число обучающих объектов в вершине v.

- критерий CCP [4] основан на редукции поддеревьев $T(v)$ с наименьшей величиной α, вычисляемой по формуле

$$\alpha = \frac{r(v) - r(T(v))}{\mu(T(v)) - 1}, \quad (2)$$

где $r(v)$ – коэффициент ошибок в вершне v, $r(T(v))$ – коэффициент ошибок в поддереве с корнем v, $\mu(T(v))$ - число листьев в поддереве с вершиной v.

- критерий MEP [4] основан на вычислении коэффициента ожидаемой ошибки для случая обучающего множества с k классами по формуле

$$E_k = \frac{n(v) - n_c(v) + k - 1}{n(v) + k}, \quad (3)$$

где $n(v)$ – число обучающих объектов в вершине v, $n_c(v)$ – число обучающих объектов в вершине v, принадлежащих мажоритарному классу c.

Каждой из этих процедур принимается решение о редукции (замене листом) поддерева (ветви) с корнем в вершине v. Критерии редукции PEP, CCP редуцирует дерево решений по направлению от листьев к корню, а MEP редуцирует дерево решений по направлению от корня к листьям, являясь по сути стратегией предредукции.

Процедурой majority_vote (estimation1(v), estimation2(v), estimation3(v)) принимается решение о целесообразности отсечения поддерева с корнем v на основе процедуры «голосования» критериев редукции.

Процедура PruningTree(d_i) реализует саму редукцию с последующим формированием области отказа Rejection_set(d_i) и вектора приоритетов priority_vector(d_i) признаков дерева d_i. В область отказа дерева попадают все объекты из обучающей выборки, которые попали в редуцированное поддерево с корнем v_pruning, в векторе приоритетов наивысшим приоритетом обладают признаки нулевого уровня всех деревьев леса решений, затем признаки первого уровня и т.д. Если совместная область отказа всех построенных деревьев леса решений ($\cap d_i$) не пуста, имеет смысл строить следующее дерево решений.

На основе предложенной модели синтеза леса решений с процедурой «голосования» критериев редукции было разработано «ядро» автоматизированной системы application-скоринга. Результаты моделирования апробированы на одной из выборок кредитных историй и представлены на рисунке 1.

Измерение процента ошибок на контрольной выборке показало, что в 67 экспериментах из 80 модель леса решений на основе «голосования» критериев редукции не хуже модели леса решений с фиксированным рангом 7.

Алгоритм синтеза леса решений с использованием процедуры голосования критериев редукции не гарантирует построение корректного (безошибочно распознающего все обучающие объекты) леса решений. В случае построения некорректного леса решений для классификации объектов могут быть использованы алгоритмы принятия решений на основе «голосования» деревьев решений леса.

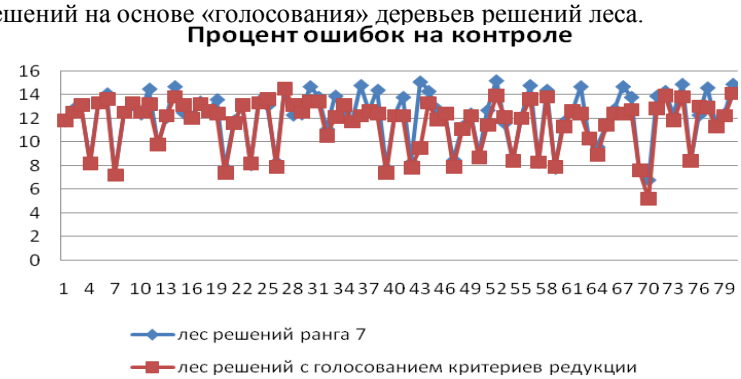

Рис.1. Эмпирическая оценка процента ошибок на контроле для леса решений ранга 7 и леса решений на основе процедуры голосования критериев редукции

Таким образом, синтез лесов решений – это одно из перспективных направлений, связанных с увеличением обобщающей способности алгоритмов этого класса. Высокая точность лесов решений достигается за счет упрощения структуры отдельных деревьев, но при этом усложняется процесс синтеза самого леса решений.

В дальнейшем представляет интерес изучение свойств лесов решений и исследование эффективности применения данной модели для оценки вероятности мошенничества заемщиков – fraud-scoring.

Источники и литература:

1. Пищулин А. Внедрение кредитного скоринга как один из факторов эффективного управления процессом кредитования / А. Пищулин // Современные кредитные технологии : доклады III Междунар. Бизнес-Форума. – 2005.
2. Дюличева Ю. Ю. Об эффективности применения ансамблей решающих деревьев в задаче оценки кредитоспособности физических лиц / Ю. Ю. Дюличева // Ученые записки Таврического национального университета им. В.И. Вернадского. – Симферополь : Информ.-изд. отдел ТНУ. – С. 23-28. – (Экономика).
3. Дюличева Ю. Ю. Модели коррекции редуцированных бинарных решающих деревьев : дис. … канд. ф.-м. наук по спец. 01.05.01 «Теоретические основы информатики и кибернетики» / Ю. Ю. Дюличева; Ин-т кибернетики им. В. М. Глушкова НАН Украины. – К., 2004.
4. Esposito F. A comparative analysis of methods for pruning decision trees / F. Esposito, D. Malerba, G. Seneraro // IEEE Transactions on Pattern Analysis and Machine Intelligence – 1997. – № 19 (5). – P. 476-491.