Информационно-математическая система прогнозирования кредитоспособности заемщиков банка

В.А. Алексеева[a], Ю.Е. Кувайскова[a]

[a] *Ульяновский государственный технический университет, 432027, ул. Северный Венец, 32, Ульяновск, Россия*

Аннотация

В статье проводится исследование алгоритмов, методов классификации и прогнозирования классов объектов и описание информационно-математической системы, разработанной на основе этих алгоритмов. Для решения задачи классификации, в частности, прогнозирования кредитоспособности заемщиков банков, используются всевозможные методы машинного обучения, а также их комбинации – так называемые агрегированные классификаторы. Реализованный программный комплекс позволяет: осуществлять предварительную подготовку исходных данных, включающую в себя дискретизацию, восстановление пропущенных данных и выявление статистически значимых факторов; применять методы классификации и строить комбинированные модели; проводить анализ качества построенных моделей с использованием ряда статистических критериев; прогнозировать классы исследуемых объектов.

Ключевые слова: машинное обучение; агрегированный классификатор; статистический анализ данных; классификация; кредитоспособность; прогнозирование

1. Введение

Рассмотрим задачу бинарной классификации объектов [1], в которой каждый объект $K_i (i = 1,..., N)$ характеризуется m-мерным вектором признаков $(X_1 ... X_m)$, принимающих как числовые, так и нечисловые значения и образующих выборку для дальнейших исследований. По значениям данных признаков нужно предсказать значение бинарной характеристики объектов y. В качестве примеров таких задач можно привести задачи технической диагностики и классификации состояния объекта [7,8], обнаружения факта передачи или отсутствия единственного сигнала, нормального или патологического состояния органа и т.д.

В данной статье рассматривается решение задачи бинарной классификации на примере кредитного скоринга, который заключается в оценке кредитоспособности заемщиков банка [10].

Увеличение размеров задолженностей по кредитам, рост рисков невозврата, а также конкуренция на рынке кредитных услуг требуют совершенствования известных методик оценки и прогнозирования кредитоспособности заемщиков с целью более точной оценки кредитного риска и принятия верного решения при выдаче кредита. Ни один из известных подходов не позволяет выявить наиболее точные модели для решения данной задачи.

В целях уменьшения задолженностей заемщиков и обеспечения возврата кредитов разработана информационно - математическая система, позволяющая оценить кредитоспособность заемщика на этапе принятия решения о выдаче кредита. Для оценки кредитоспособности использовались методы машинного обучения [2] с агрегированием различных классификаторов на основе деревьев решений, нейронной сети, дискриминантного анализа, байесовского классификатора, метода опорных векторов, логистической регрессии и др.

2. Агрегированные классификаторы

Для решения задачи прогнозирования класса исследуемого объекта в настоящее время существует множество методов и моделей. Для исследования оценки кредитных рисков были взяты следующие методы: деревья принятия решений [14], нейронные сети [13], дискриминантный анализ [2], байесовский классификатор [5], метод опорных векторов, логистическая регрессия [6], бэггинг деревьев решений, модели нечеткого вывода [10], метод эмпирической функции. Каждый из этих методов имеет свои преимущества и недостатки. Например, нельзя применять метод эмпирической функции для прогнозирования данных, набор значений признаков которых не совпадает хотя бы с одним набором из обучающей выборки, а для применения байесовского подхода необходимо прежде привести исходные данные к интервальной шкале, чтобы переменные были дискретными, иначе это может привести к потере значимой информации. Нет универсальной модели, с помощью которой можно было бы с высокой точностью оценить принадлежность объекта к тому или иному классу.

Так как в зависимости от конкретной ситуации наилучшим с точки зрения точности прогнозирования может оказаться любой из методов машинного обучения, то предлагается совместное использование различных классификаторов, построенных на разных частях обучающей выборки [3]. Используя девять выше перечисленных методов, можно методом полного перебора получить $2^9 - 9 - 1 = 502$ всевозможных комбинаций различных моделей.

Для принятия решения о принадлежности заемщика к одному из классов (кредитоспособен или некредитоспособен) на основе результатов параллельного применения к исходной выборке отдельных методов классификации возможно агрегирование результатов по трем признакам:

- по среднему значению (вероятность принадлежности исследуемого объекта классу $y = 1$ («кредитоспособный клиент») считается как среднее арифметическое значений вероятностей принадлежности объекта классу $y = 1$, найденных по всем девяти методам классификации);

- по медиане (сначала ранжируется ряд, содержащий результаты базовых методов классификации в комбинации, вероятность находится путем вычисления результата срединного классификатора в случае их нечетного количества или полсуммы результатов срединных базовых классификаторов в четном случае);

- с помощью процедуры голосования (результат агрегированного классификатора по голосованию представляет собой среднее значение результатов базовых методов классификации, которые определили факт принадлежности заданного объекта классу $y = 1$ с вероятностью $\geq 0{,}1$).

Для оценки кредитоспособности клиентов на основе агрегированных классификаторов предлагается алгоритм решения, включающий следующие этапы:

1) формирование и обработка исходной выборки. Данный этап включает в себя разбиение выборки на обучающую (применяется для построения моделей классификации) и тестовую (применяется для проверки точности построенных моделей), восстановление пропущенных данных [9], дискретизацию некоторых признаков и поиск факторов, наиболее существенно влияющих на выходную характеристику *y*;
2) построение на обучающей выборке параллельно девяти моделей классификации;
3) построение агрегированных классификаторов;
4) прогнозирование на тестовой выборке кредитоспособности новых клиентов на основе всех построенных моделей;
5) получение результата прогнозирования кредитоспособности каждого клиента. На данном этапе определяется среднее значение вероятностей всех построенных моделей;
6) выбор наилучшей модели, т.е. модели с наиболее высокой точностью прогнозирования. Точность модели определяется по ряду критериев [12].

3. Информационно-математическая система кредитного скоринга

На основе рассмотренного выше алгоритма была разработана информационно-математическая система кредитного скоринга [4]. Она позволяет прогнозировать класс рассматриваемого объекта (в частности, кредитоспособоность заемщиков) на основе обучающей выборки. Программный комплекс был разработан в среде программирования Matlab R2014a, содержащей все методы обработки исходных данных и большинство алгоритмов машинного обучения, необходимых для решения задачи классификации. Исходные данные представляют собой информацию о клиентах, которая включает анкетные данные и соответствующий класс кредитоспособности «старых» клиентов; анкетные данные «новых» клиентов; анкетные данные, кредитную историю и условия по кредитной сделке заемщиков, погашающих кредит.

Программа позволяет осуществлять предварительную подготовку исходных данных: восстановление пропущенной информации; дискретизацию характеристик; кодирование нечисловых данных; выбор статистически значимых признаков. В программе реализованы все рассмотренные выше методы классификации, а также агрегированный классификатор с возможностью выбора критерия агрегирования (по среднему значению, по медиане или с помощью процедуры голосования).

При построении классификаторов для получения несмещенных оценок показателей их качества используется метод L-кратной перекрестной проверки. Суть данного метода заключается в разделении исходной выборки на L непересекающихся частей, приблизительно равных по объему. В программе можно выбрать значение L, оно варьируется от 3 до 10. Далее в порядке очереди каждая часть выступает в роли контрольной выборки, а остальные части объединяются в обучающую выборку. Итоговая оценка качества классификатора определяется усреднением ошибок по всем L контрольным выборкам. Эта процедура позволяет исключить возможность «подгонки» модели к наилучшим прогнозным характеристикам.

В результате работы программы формируются значения показателей качества построенных моделей для трех порогов отсечения (порог отсечения – значение, выше которого объект признается принадлежащим классу $y = 1$): порог отсечения 0,5; оптимальный порог отсечения; порог отсечения, заданный пользователем. Оптимальный порог классификации – это наименьшее отклонение между ошибками I рода и II рода.

Оценка качества построенных моделей классификации и агрегированных классификаторов производится с помощью следующих критериев [12]: ошибки первого и второго рода, ROC-кривые, показатель AUC, среднеквадратическая ошибка прогнозирования (MSE), а также процент верных прогнозов кредитоспособных клиентов и процент верно предсказанных некредитоспособных клиентов.

По заданным критериям пользователь может определить, какой метод или комбинация методов дают оптимальный результат для исследуемых объектов, и построить прогноз для исходного набора значений признаков. Работа агрегированного классификатора производится программой, т.е. по ряду критериев автоматически формируется оптимальная комбинация методов, после чего пользователь может сравнить результаты агрегированного классификатора с базовыми методами классификации.

4. Примеры применения разработанной системы кредитного скоринга

В качестве первого примера рассмотрены результаты работы программы по реализации агрегированного классификатора для выборки по клиентам немецкого банка, включающей 900 заемщиков, описанных 20 признаками (статус текущего чекового счета, кредитная история, цель кредита, срок кредита, сумма кредита, средний баланс на накопительном счете, стаж работы на последнем месте, доход в %, семейное положение, поручители, постоянное проживание на последнем месте, данные об имуществе, возраст, имеющиеся кредиты, вид жилья, количество предыдущих кредитов в этом банке, вид деятельности, количество иждивенцев, наличие телефона, гражданство), и одной зависимой бинарной переменной (заемщик кредитоспособен или некредитоспособен). С помощью программы проведена предварительная обработка данных, включающая в себя дискретизацию ряда признаков и кодирование нечисловых данных, таких как гражданство клиентов, образование, семейное положение и т.д. с помощью чисел. Проанализированы все девять отдельных методов классификации и агрегированный классификатор. Агрегирование проводилось по среднему значению. Также возможны варианты применения агрегирования по всем трем признакам. При классификации использовалась 10-кратная перекрестная проверка.

Для исследуемой выборки получен оптимальный из всех возможных агрегированный классификатор при пороге отсечения 0,5, состоящий из следующих методов: нейронные сети, логистическая регрессия, бэггинг деревьев решений, метод эмпирической функции и метод нечеткого логического вывода. Результаты работы программы представлены в таблице 1. Лучший результат классификации получен с помощью агрегированного классификатора, так как среднеквадратическая ошибка агрегированного классификатора меньше, чем у остальных методов; самый высокий процент верных прогнозов кредитоспособных клиентов наблюдается для двух методов: агрегированного классификатора и бэггинга деревьев решений, но при этом ошибка первого рода у агрегированного классификатора ниже; по некредитоспособным клиентам агрегированный классификатор дает средний результат по прогнозу, но с минимальной ошибкой второго рода.

Таблица 1. Результаты классификации по заемщикам немецкого банка

Классификатор	Среднекв. ошибка (MSE)	Кредитоспособные ($y=1$)		Некредитоспособные($y=0$)	
		Верный прогноз, %	Ошибка I рода, %	Верный прогноз, %	Ошибка II рода, %
Нейронная сеть (НС)	0,1743	84,1	56,8	44,2	15,8
Дискриминантный анализ (ДА)	0,1862	84,5	48,0	57,0	16,7
Байесовский классификатор (БК)	0,2012	76,2	32,8	62,4	28,5
Метод опорных векторов (МОВ)	0,1653	88,5	47,7	63,2	15,1
Деревья решений (ДР)	0,2395	79,1	46,1	57,6	26,8
Логистическая регрессия (ЛР)	0,1852	88,7	50,2	50,4	13,3
Бэггинг деревьев решений (БДР)	0,1532	88,1	49,3	51,1	11,9
Метод эмпирической функции (МЭФ)	0,4576	35,7	4,8	95,3	70,2
Нечеткая логика (НЛ)	0,1845	79,3	39,1	68,2	23,5
Агрегированный классификатор (АК)	0,1552	88,5	36,5	61,9	11,0

В таблице представлены только три показателя качества построенных классификаторов. Программа позволяет также формировать диаграммы, отображающие площади под ROC-кривыми (AUC). На рис. 1 представлена такая диаграмма для исследуемой выборки. ROC-кривая [12], также известная как кривая ошибок, показывает соотношение между долей верных положительных классификаций от общего числа положительных классификаций и долей ошибочных положительных классификаций от общего числа отрицательных классификаций при варьировании порога решающего правила. Показатель AUC позволяет количественно оценить график ROC-кривой. Чем выше показатель AUC, тем точнее классификатор. По графику видно, что наиболее точный результат классификации дают агрегированный классификатор и бэггинг деревьев решений, но все же выше значение AUC у агрегированного классификатора.

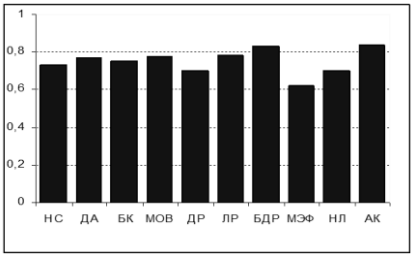

Рис.1. Площади под ROC-кривыми для исследуемой выборки.

В статье [4] исследуется выборка по заемщикам немецких банков, но большего объема (1000 наблюдений). Уменьшение числа наблюдений изменило результаты классификации незначительно.

Аналогично проведено исследование данных по кредитоспособности австралийских заемщиков. Все имена переменных и их значения закодированы в целях обеспечения конфиденциальности информации. Данные включают в себя одну бинарную зависимую переменную, характеризующую кредитоспособность (принимает значение 0 в случае некредитоспособного клиента или 1 в случае кредитоспособного клиента) и 14 независимых признаков. Всего имеется 690 наблюдений.

Для исследуемой выборки получен оптимальный из всех возможных агрегированный классификатор при пороге отсечения 0,5, состоящий из следующих методов: нейронные сети, логистическая регрессия, байесовский классификатор и метод нечеткого логического вывода. Результаты работы программы представлены в таблице 2. Лучший результат классификации получен с помощью агрегированного классификатора.

Таблица 2. Результаты классификации по заемщикам австралийского банка

Классификатор	Среднекв. ошибка (MSE)	Кредитоспособные ($y=1$)		Некредитоспособные($y=0$)	
		Верный прогноз, %	Ошибка I рода, %	Верный прогноз, %	Ошибка II рода, %
Нейронная сеть (НС)	0,1258	88,6	56,2	67,3	13,6
Дискриминантный анализ (ДА)	0,2511	75,8	42,1	54,2	25,2
Байесовский классификатор (БК)	0,1253	84,6	58,3	62,8	21,8
Метод опорных векторов (МОВ)	0,1648	87,2	42,2	55,1	20,1
Деревья решений (ДР)	0,3519	69,8	51,0	56,8	18,4
Логистическая регрессия (ЛР)	0,2157	78,9	42,9	61,5	12,6
Бэггинг деревьев решений (БДР)	0,1642	76,8	45,2	54,5	20,3
Метод эмпирической функции (МЭФ)	0,5862	58,1	31,2	66,8	25,8
Нечеткая логика (НЛ)	0,1683	87,6	48,6	57,0	16,1
Агрегированный классификатор (АК)	0,1146	89,1	31,8	63,1	12,1

Представленные примеры вкратце отображают возможности разработанной информационно-математической системы кредитного скоринга. Из всех возможных методов классификации выбирается тот, который позволяет с наиболее высокой точностью прогнозировать кредитоспособность и некредитоспособность клиентов одновременно, при этом минимизируя среднеквадратическую ошибку и ошибки первого и второго рода и максимизируя показатель AUC. Используя выбранный в программе метод, можно по заданному набору значений факторов определить, к какому классу относится исследуемый объект. Также разработанный программный комплекс позволяет обновлять модели по мере поступления новых данных.

5. Заключение

Для решения задачи бинарной классификации объектов предложено использование девяти известных методов машинного обучения, а также их всевозможные комбинации. Говорить об эффективности какого-либо из рассмотренных методов нецелесообразно, так как для разных выборок, даже для разных частей одной выборки, можно получить различные результаты. Эти методы и алгоритм построения агрегированных классификаторов реализованы в виде информационно-математической системы кредитного скоринга.

Разработанная программа позволяет подобрать наилучшую модель или оптимальный агрегированный классификатор. Для исследуемых выборок наилучшим оказался классификатор. В случае данных по немецким заемщикам наиболее точный прогноз получен при использовании комбинации следующих методов: нейронных сетей, логистической регрессии, бэггинга деревьев решений, метода эмпирической функции и метода нечеткого логического вывода; для австралийских данных классификатор включает нейронные сети, логистическую регрессию, байесовский классификатор и метод нечеткого логического вывода. Применение агрегированного классификатора позволило достигнуть поставленной цели – повышения точности предсказания кредитоспособности клиентов банка.

Разработанную систему кредитного скоринга можно использовать для любой задачи бинарной классификации, в частности, для прогнозирования технического состояния объектов [7,8] или наличия и отсутствия сигналов.

Литература

[1] Айвазян, С.А. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. - М.: Финансы и статистика, 1989. - 607 с.

[2] Алексеева, В.А. Использование методов интеллектуального анализа в задачах бинарной классификации / В.А. Алексеева // Известия Самарского научного центра Российской академии наук. - 2014. – Т. 16, №6(2). – С. 354-356.

[3] Алексеева, В.А. Построение агрегированного бинарного классификатора / В.А. Алексеева // Современные проблемы проектирования, производства и эксплуатации радиотехнических систем. - 2015. - № 1-2 (9) . - С. 211-214.

[4] Алексеева, В.А. Использование методов машинного обучения в задачах бинарной классификации / В.А. Алексеева // Автоматизация процессов управления. - 2015. - № 3 (41) . - С. 58-63.
[5] Бидюк, П.И. Построение и методы обучения байесовских сетей / П.И. Бидюк, А.Н. Терентьев // Информатика и кибернетика. – 2004. – № 2. – С. 140-154.
[6] Васильев, Н.П. Опыт расчета параметров логистической регрессии методом Ньютона-Рафсона для оценки зимостойкости растений / Н.П. Васильев // Математическая биология и биоинформатика. – 2011. – Т. 6, №2. – С.190-199.
[7] Клячкин, В.Н. Применение методов машинного обучения при решении задач технической диагностики / В.Н. Клячкин, И.Н. Карпунина, Ю.Е. Кувайскова, А.С. Хорева // Научный вестник УВАУ ГА(И). – 2016. Т.8. – С. 158-161.
[8] Кувайскова, Ю.Е. Применение методов нечеткой логики и машинного обучения при решении задачи технической диагностики / Ю.Е. Кувайскова, А.Д. Барт, К.А. Федорова // Информатика и вычислительная техника: сборник научных трудов VIII Всероссийской научно-технической конференции аспирантов, студентов и молодых ученых ИВТ-2016. – 2016. – С. 160-166.
[9] Литтл, Р.Дж. А. Статистический анализ данных с пропусками / Р.Дж. А. Литтл, Д.Б. Рубин. – М.: Финансы и статистика, 1990. – 336 с.
[10] Штовба, С.Д. Идентификация нелинейных зависимостей с помощью нечеткого логического вывода в системе Matlab / С.Д. Штовба // Научно-практический журнал Exponenta Pro: математика в приложениях. – 2003. – №2(2). – С. 9-15.
[11] Шунина, Ю.С. Прогнозирование кредитоспособности клиентов на основе методов машинного обучения / Ю.С. Шунина, В.А. Алексеева, В.Н. Клячкин // Финансы и кредит. – 2015. – №27(651). – С. 2-12.
[12] Шунина, Ю.С. Критерии качества работы классификаторов / Ю.С. Шунина, В.А. Алексеева, В.Н. Клячкин // Вестник Ульяновского государственного технического университета. – 2015. – № 2(70). – С. 67-70.
[13] Ясницкий, Л.Н. Введение в искусственный интеллект / Л.Н. Ясницкий. – М.: Издательский центр «Академия», 2005. – 176 с.
[14] Якупов, А.И. Применение деревьев решений для моделирования кредитоспособности клиентов коммерческого банка / А.И. Якупов // Искусственный интеллект. – 2008. – № 4. – С. 208–213.