Система кредитного скоринга на основе логических алгоритмов классификации

Д. А. Кочедыков., А. А. Ивахненко, К. В. Воронцов
(Москва)

Задача кредитного скоринга

Задача кредитного скоринга возникает в банках и других кредитных организациях при принятии решений о выдаче кредитов. Задача заключается в том, чтобы на основе некоторой информации о заявителе обоснованно принять решение — стоит ли ему выдавать кредит, и если да, то на каких условиях. Если речь идёт о физических лицах, то исходной информацией для классификации заёмщиков на «хороших» и «плохих» является заполняемая ими анкета, информация о прошлых кредитах заемщика (кредитная история), в некоторых случаях — информация о движении средств на счетах заемщика. В анкете указывается сумма и назначение кредита, возраст, пол, социальное положение, образование, профессия, и т. д. Сотрудники банка могут вносить дополнительные сведения, например, психологический портрет заёмщика или результаты проверки, проведённой службой безопасности банка. В итоге формируется описание заёмщика, содержащее от 20 до 200 признаков, в зависимости от используемой банком методики оценки кредитоспособности.

Перечислим некоторые специфические особенности данной прикладной задачи, которые накладывают существенные ограничения на выбор модели алгоритма классификации и способы её настройки.

1. Анкетные данные, как правило, являются неполными и разнородными. Кроме того, они характеризуются высоким уровнем шума, так как невозвраты кредитов чаще всего обусловлены случайными факторами. В некоторых случаях построить классификатор, допускающий менее 25% ошибок, не представляется возможным. Отчасти это компенсируется тем, что ошибки I и II рода неравноценны: по экспертным оценкам пропуск «плохого» заёмщика обходится банку в 3–10 раз дороже, чем потеря «хорошего» клиента.

2. Наряду с классификацией банк решает задачу определения риска, который несет в себе выдача запрошенного кредита в каждом конкретном случае и принятие решения о приемлемости такого риска для кредитного портфеля. При этом банк старается минимизировать риск и максимизировать прибыль путём варьирования условий выдаваемого кредита — ставки, суммы, срока, и т. д. Для решения этих задач алгоритм должен выдавать не только саму классификацию, но и оценку вероятности того, что данный заёмщик принадлежит классу «плохих». Кроме того, должна быть возможна более тонкая классификация рисков, чем «плохой»/«хороший»: кредит может возвращаться не полностью, либо с задержками (что влечёт операционные расходы), либо досрочно (что влечёт потерю части прибыли).

3. Одно из самых существенных ограничений состоит в том, что алгоритм классификации заёмщиков должен быть простым, понятным кредитному эксперту и должен допускать запись на естественном языке в терминах предметной области.

4. В настоящее время многие российские банки только приступают к кредитной деятельности и ещё не накопили значительного числа кредитных историй. Поэтому алгоритмы приходится настраивать по малому числу прецедентов. Используемые в западных банках на протяжении 40 лет статистические методы не могут уверенно работать на столь малых обучающих выборках. В частности, к этим методам относится широко применяемая логистическая регрессия, с помощью которой строят так называемые «скоринговые карты».

Логические алгоритмы классификации

Сформулированным выше требованиям удовлетворяют логические алгоритмы классификации, в частности, решающие списки, решающие деревья и алгоритмы типа «Кора», основанные на взвешенном голосовании конъюнкций. Все они строятся в виде композиции простых логических условий и хорошо интерпретируются. Вместе с тем, применение в области кредитного скоринга требует существенной доработки классических методов синтеза логических классификаторов — при любых условиях они должны оставаться интерпретируемыми, и их настройка должна занимать приемлемое время. При этом приходится искать компромисс между точностью настройки на обучающих данных и сложностью алгоритмов.

В данной работе рассматриваются алгоритмы синтеза классификаторов, основанные на поиске логических закономерностей. *Закономерность* — это предикат $\varphi_c : X \to \{0,1\}$, определённый на множестве объектов X. Если $\varphi_c(x) = 1$, то говорят, что закономерность φ_c относит объект x к классу c. Будем рассматривать закономерности, имеющие вид конъюнкции простых однопризнаковых условий

$$\varphi_c(x) = \beta_1(x) \wedge \ldots \wedge \beta_k(x),$$

где термы $\beta_i(x)$ имеют вид $[x_j < a]$, $[x_j > a]$, $[x_j = a]$ или $[a < x_j < b]$, $(x_1, \ldots, x_n)$ — признаковое описание объекта x. Синтез закономерности заключается в оптимизации числа термов k, опорного множества признаков, образующих термы, и параметров a, b в каждом терме. Оптимизация производится по критерию информативности на основе *точного теста Фишера* [1]. Это позволяет находить достаточно надёжные закономерности в условиях зашумлённых данных малого объёма. Синтез информативных конъюнкций производится методом усечённого поиска в ширину, аналогичным алгоритму ТЭМП [2] или многорядному итерационному алгоритму МГУА. При реализации алгоритма было введено несколько

десятков различных эвристик, нацеленных на повышение качества, интерпретируемости и разнообразия синтезируемых закономерностей.

Получаемые закономерности $\varphi^1_{c_1}(x),\ldots,\varphi^T_{c_T}(x)$ объединяются в алгоритм классификации одним из двух способов.

1. *Решающий список* закономерностей (decision list, DL) или комитет старшинства. При классификации объекта x закономерности проверяются последовательно для всех $t=1,\ldots,T$, пока для некоторого t не выполнится условие $\varphi^t_{c_t}(x)=1$. Тогда объект x относится к классу c_t. На практике оказалось, что решающие списки хорошо интерпретируются экспертами только в том случае, когда для каждого класса c строится отдельный список $\varphi^1_c(x),\ldots,\varphi^{T_c}_c(x)$ и закономерности разных классов не перемешиваются.

2. *Взвешенное голосование* закономерностей (weighted voting, WV) или комитет взвешенного большинства. В этом алгоритме объект x относится к тому классу c, для которого максимальна доля закономерностей $\Gamma_c(x)$, относящих данный объект к данному классу:

$$\Gamma_c(x)=\frac{1}{T_c}\sum_{t=1}^{T_c} w_t\varphi^t_c(x).$$

Для настройки весов закономерностей w_t используется вариант алгоритма бустинга [3]. Преимущества бустинга в том, что он позволяет вычислять веса по явным формулам и последовательно строить закономерности таким образом, чтобы каждая следующая существенно отличалась от предыдущих. Кроме того, бустинг позволяет эффективно выделять из обучающей выборки шумовые объекты.

Экспериментальное сравнение качества алгоритмов

Был проведен эксперимент по сравнению качества работы предложенных алгоритмов DL и WV с эталонными алгоритмами на 6 наборах реальных данных из репозитория UCI [4], причём german, crx и australian — задачи о выдаче кредитов по анкетным данным. В таблице 1 приведены доли ошибок на скользящем контроле в процентах от длины обучающей выборки. Алгоритмы DL и WV показали качество классификации, сравнимое со стандартными алгоритмами. Отметим, что оба алгоритма существенно превосходят остальные на выборке малого размера (hepatits).

Программный комплекс ScoringAce®

На основе разработанных алгоритмов классификации создан программный комплекс ScoringAce для поиска логических закономерностей в данных, автоматического и полуавтоматического построения хорошо интерпретируемых алгоритмов классификации и автоматизации принятия кредитных решений. Программный комплекс внедрен в одном из крупных

российских банков и в настоящее время используется при массовом кредитовании физических лиц. Архитектура программного комплекса складывается из следующих компонент.

Решающий сервер предназначен для обработки поступающих кредитных заявок и принятия кредитных решений в режиме онлайн. База данных сервера накапливает кредитные истории, хранит все действующие в настоящее время и действовавшие в прошлом логические классификаторы (скоринговые модели), а также протоколирует всю информацию, проходящую через сервер.

АРМ аналитика рисков используется для статистического анализа данных, поиска закономерностей, построения логических классификаторов, оценивания риска отдельных кредитов и текущего кредитного портфеля в целом, построения аналитических отчетов.

АРМ кредитного аналитика служит для online-верификации решений, принимаемых Решающим сервером по поступающим заявкам на выдачу кредитов. Используется, в частности, при выходе на новые рынки для оперативного контроля и корректировки принимаемых решений.

АРМ аудита применяется для отслеживания процессов кредитования и построения статистических отчетов по работе Решающего сервера.

АРМ подписания скоринг-моделей предназначено для обновления списка текущих скоринговых моделей на Решающем сервере.

Задача	RIPPER		C4.5		C5.0	SLIPPER	**DL**	**VW**
	-opt	+opt	Trees	Rules				
german	28.6	28.7	27.5	**27.0**	28.3	27.2	28.2	27.9
hepatits	20.7	23.2	18.8	18.8	20.1	17.4	**7.7**	8.4
crx	15.5	15.2	14.2	15.5	14.0	15.7	**13.9**	14.2
liver	32.7	31.3	37.7	37.5	31.9	32.2	30.6	**28.4**
horse-colic	17.0	16.3	16.3	16.0	15.3	**15.0**	21.7	16.7
australian	14.7	—	15.2	—	14.6	—	15.1	**12.8**

Таб.1 Сравнение качества логических алгоритмов классификации. Процент ошибок на тестовых данных при 10-кратном скользящем контроле (10-fold cross-validation).

Литература

1. Martin J. K. An exact probability metric for decision tree splitting and stopping // Machine Learning. — 1997. — Vol. 28, no. 2-3. — Pp. 257–291.
2. Лбов Г. С. Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981.
3. Cohen W. W., Singer Y. A simple, fast and effective rule learner // Proc. of the 16 National Conference on Artificial Intelligence. — 1999. — P. 335.
4. Blake C., Merz C. UCI repository of machine learning databases: Tech. rep.: University of California, Irvine, CA, 1998.