4П

УДК 519.4

Н.Б. Паклин[1], В.В. Афанасьев[2]
[1]Рязанский государственный университет имени С.А. Есенина, Россия
Россия, 390000, г. Рязань, ул. Свободы, 46
[2]Компания BaseGroup Labs, Россия
Россия, 390023, г. Рязань, ул. Новая, 53в

Оптимальное квантование для повышения качества бинарных классификаторов

N.B. Paklin[1], V.V. Afanasiev[2]
[1]Ryazan State University, Russia
Russia, 390000, c. Ryazan, st. Svobody, 46
[2]BaseGroup Labs Company
Russia, 390000, c. Ryazan, st. 53

Optimal Quantization to Improve the Quality of Binary Classifiers

М.Б. Паклін[1], В.В. Афанасьєв[2]
Рязанський державний університет імені С. Єсеніна, Росія
Росія, 390000, м. Рязань, вул. Свободи, 46
Компанія BaseGroup Labs, Росія
Росія, 390023, м. Рязань, вул. Нова, 53в

Оптимальне квантування для підвищення якості бінарних класифікаторів

В статье рассматривается задача оптимального квантования и ее более общий случай – формирование конечных классов с целью предобработки выборок в машинном обучении для повышения качества бинарных классификаторов. На примере решения открытой конкурсной задачи показано, что предварительное формирование конечных классов позволяет построить эффективные классификаторы даже с применением относительно простых средств интеллектуального анализа данных, таких как логистическая регрессия.
Ключевые слова: оптимальное квантование, машинное обучение, бинарный классификатор.

In the article the task of optimal quantization and its more general case – coarse classing for the purpose of sample data transforming in machine learning to improve the quality of binary classifiers. On the example of an open competition is shown that fine and coarse classing procedures allows build effective classifiers, even with a relatively simple data mining tools, such as logistic regression.
Key words: optimal quantization, optimal binning, coarse classing, machine learning, binary classifier.

У статті розглядається задача оптимального квантування і її більш загальний випадок - формування кінцевих класів з метою предобрабки вибірок у машинному навчанні для підвищення якості бінарних класифікаторів. На прикладі рішення відкритого конкурсного завдання показано, що попереднє формування кінцевих класів дозволяє побудувати ефективні класифікатори навіть із застосуванням відносно простих засобів інтелектуального аналізу даних, таких як логістична регресія.
Ключові слова: оптимальне квантування, машинне навчання, бінарний класифікатор.

Введение

В машинном обучении и предсказательной аналитике большое число задач сводится к бинарной классификации. Число прикладных проблем, которые могут быть сформулированы в терминах «событие» и обратном ему «не-событие», большое: наличие просрочки по кредиту, отклик на маркетинговую кампанию, отказ клиента от услуг компании, факт заболевания и так далее. Поэтому построение эффективных бинарных классификаторов является актуальной и важной практической задачей, и один из распространенных путей повышения их качества сводится к разработке новых алгоритмов и модификации существующих. Впоследствии часто оказывается, что наиболее эффективные алгоритмы (нейронные сети, леса деревьев решений) плохо интерпретируются исследователем, что согласуется с принципом неопределенности Бреймана [1]: чем выше точность, тем хуже интерпретируемость. Кроме того, проблема обработки выбросов и пропущенных значений нередко опускается из рассмотрения.

Существует ряд подходов, позволяющих провести предварительную обработку обучающих выборок с целью улучшения работы классификаторов, а также решить ряд сопутствующих задач: исследовать значимость входных переменных, и в той или иной форме проверить гипотезы о причинных связях между ними. Это особенно важно при использовании относительно простых классификаторов, например, на основе логит- и пробит-моделей.

Одним из таких подходов является *оптимальное квантование*, хотя данная формулировка, по нашему мнению, не отражает полностью его сути, поскольку он может применяться к переменным, измеренным не только в интервальной шкале. В англоязычной литературе [2] применяется термин «Fine&Coarse Classing», что можно перевести как «начальные и конечные классы».

Процедура формирования конечных классов определяет, как будет представлена переменная в выборке с точки зрения числа ее уникальных значений. По сути, это есть сокращение числа разнообразных значений переменной, которое обычно связывают с изменением интервала дискретизации значений. Задача заключается в уменьшении числа значений исходного набора данных за счет их объединения в пределах некоторого интервала с использованием информации о целевой переменной. В результате такого преобразования число значений переменной должно уменьшиться без существенного ущерба для информативности данных.

Рассмотрим подробно постановку задачи оптимального квантования на примере интервальной переменной x, принимающей значения на диапазоне $[L, R]$ [3]. Опишем ее как n-мерный вектор начальных значений, или начальных классов, которые встречались в обучающей выборке:

$$\mathbf{a} = \{a_j \in Z : a_j < a_{j+1}, a_0 = L, a_{n-1} = R, j \in [0, n-1]\}.$$

Данный диапазон разбивается на m интервалов квантования, $m < n$, границами которых могут быть только начальные классы:

$$[L, R] = \bigcup_{k=0}^{m-2} (b_j, b_{j+1}],$$

$$b_k \in \{a_j\}, b_0 = L, b_{m-1} = R.$$

Конечные классы однозначно определяются m-мерным вектором границ интервалов квантования:

$$\mathbf{b} = \{b_k \in \{a_j\} : b_k < b_{k+1}, b_0 = L, b_{m-1} = R, k \in [0, m-1]\}.$$

Алгоритмы оптимального квантования

Способ поиска вектора **b** определяет тот или иной алгоритм формирования конечных классов. Разработано несколько таких алгоритмов, например, *MAPA* (Monotone Adjacent Pooling Algorithm) [2]. Заметим, что алгоритмы квантования являются квазиоптимальными, тем или иным образом сокращающие перебор всех возможных вариантов границ конечных классов. В их основе лежит метод *WoE*-анализа. Пусть имеется обучающая выборка, в которой каждому объекту, который описывается набором переменных, ставится в соответствие бинарная целевая переменная класса с двумя состояниями – событие и не-событие. Для произвольного интервала $(b_k, b_{k+1}]$ вычисляется коэффициент *WoE*, или вес доказательства:

$$WoE_k = \ln \frac{(N_k / N)}{(P_k / P)} = \ln \frac{F^-}{F^+}, \quad (1)$$

где k – индекс начального класса, N_k – число не-событий, попавших в интервал, N – общее число не-событий в исходном наборе данных, P_k – число событий, попавших в класс, P – общее число событий.

Интерпретация коэффициентов *WoE* следующая. В числителе отношения (1) под логарифмом стоит относительная частота появления не-событий в классе F^-, а знаменателе – относительная частота появления событий F^+. Если $F^- > F^+$, то логарифм их отношения также больше 0 (логарифмическая зависимость для вычисления коэффициентов *WoE* представлена на рис. 1).

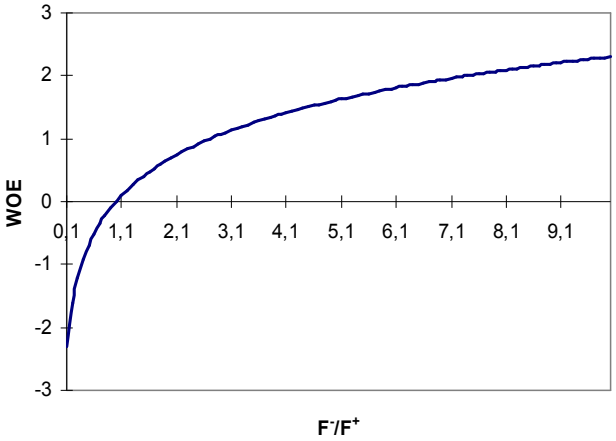


Рисунок 1 – Логарифмическая зависимость для *WoE*

Отрицательные значения коэффициентов *WoE* указывают на большую вероятность появления событий в интервале.

На основе коэффициентов *WoE* вычисляется величина, определяющая значимость переменной, называемая информационным индексом (от англ. Information Value – *IV*), по формуле:

$$IV = \sum_{k=0}^{m-1} \left\{ \left(\frac{N_k}{N} - \frac{P_k}{P} \right) \cdot WoE_k \right\}. \quad (2)$$

Информационный индекс пропорционален разности относительных частот появления событий и не-событий в каждом интервале, просуммированной по всем интервалам. Информационный индекс всегда является положительной величиной. На основе

него определяется значимость признака по следующей шкале: $IV < 0{,}02$ – отсутствует; $0{,}02 \le IV < 0{,}1$ – низкая; $0{,}1 \le IV < 0{,}3$ – средняя; $IV \ge 0{,}3$ – высокая.

Формирование конечных классов производится путем установки границы между ними таким образом, чтобы максимизировать информационный индекс *IV*:

$$IV(\mathbf{b}) \to \max .$$

Потеря информации при оптимальном квантовании характеризуется как разница информационных индексов начальных и конечных классов: $IV(\mathbf{a}) - IV(\mathbf{b})$.

Предварительные экспериментальные исследования авторов показали, что принцип жадного выбора, заложенный в алгоритме *MAPA*, часто приводит к неоптимальным конечным решениям и сильной потере информационной насыщенности переменной.

Предлагаемый алгоритм разработан авторами на основе работы [3] и реализован в программном продукте «Аналитическая платформа Deductor 5.3» (http://www.base-group.ru/deductor). Он состоит из двух главных этапов: формирование начальных классов и формирование конечных классов.

Формирование начальных классов

Для каждого начального класса a_j подсчитывается число «событий» и «не-событий», попавших в данный класс, и вычисляется вес доказательства WoE_j. Если имеются пропуски значений переменной, то для таких наблюдений формируется отдельный начальный класс с номером $j = -1$.

Формирование конечных классов

При формировании первых двух конечных классов, представляемых вектором $\mathbf{b} = \{b_0, b_1, b_2\}$, единственная граница b_1 последовательно проходит по всем начальным классам, исключая крайние: $b_1(j) = a_j \forall\, j \in (0; n-1)$. Для этого разбиения вычисляется информационный индекс, согласно выражению (2) с использованием коэффициентов WoE_j, рассчитанных на этапе формирования начальных классов. Среди возможной границы выбирается та, которая максимизирует *IV*: $b_1^* = \arg\max_j IV(b_1(j))$ при выполнении условия соблюдения минимального веса конечного класса (обычно не менее 0,05). Затем процедура повторяется внутри каждого найденного конечного класса. Алгоритм формирует конечные классы до тех пор, пока не будет достигнуто максимально установленное число классов, либо станет невозможно сформировать новый конечный класс с весом, меньше заданного.

В *MAPA* граница конечного класса выбирается каждый раз при срабатывании условия $IV(a_j) > IV(a_{j+1})$. Это снижает вычислительные затраты (для формирования границ требуется один проход), но делает, как уже отмечалось, результаты квантования неудовлетворительными. В рассматриваемом алгоритме за основу взят принцип рекурсивного разбиения, который, в частности, применяется при построении деревьев классификационных правил.

В случае переменной, измеренной в шкале наименований или отношений, алгоритм формирования конечных классов ничем не отличается от приведенного, только делается предварительный этап: метки начальных классов преобразовываются в уникальные числовые коды и упорядочиваются по возрастанию *WoE*. Таким образом, объединению будут подлежать только соседние начальные классы с близкими коэффициентами *WoE*.

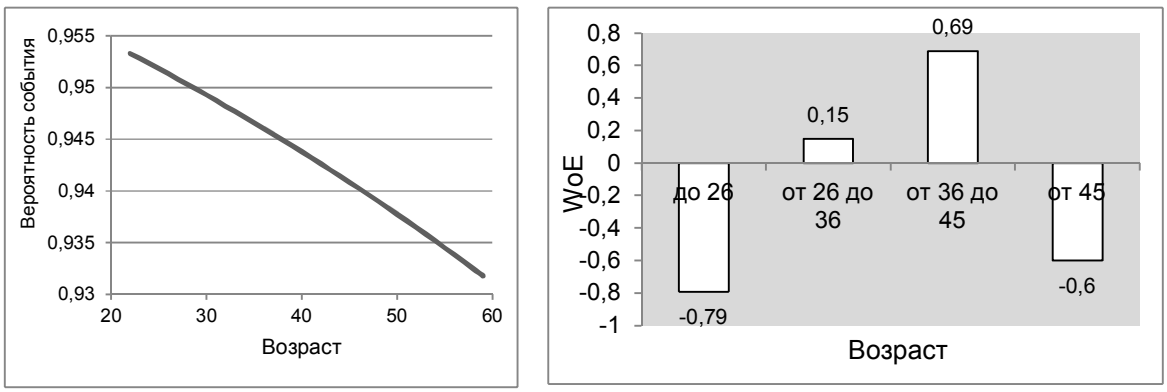
Заметим, что результаты оптимального квантования могут не устроить исследователя, и тогда добавляется третий этап – ручная корректировка границ или состава конечных классов.

Оптимальное квантование в логистической регрессии

Несмотря на разнообразие алгоритмов построения бинарных классификаторов, логистическая регрессия сегодня остается востребованным инструментом в прикладных задачах, так как позволяет получать хорошо интерпретируемые балльные скоринговые карты и вероятностные оценки наступления события [2], [4].

Существует много причин, по которым процедуры оптимального квантования активно используются при подготовке выборок к моделированию методом логистической регрессии. Основная заключается в том, что взаимосвязи между непрерывной переменной и событием не всегда линейные. Уравнение логистической регрессии, несмотря на то, что ее выходное значение подвергается нелинейному преобразованию путем логита, все равно моделирует линейные зависимости между входами и выходами. Она линейна по параметрам, а сами параметры представлены непосредственно, а не как функции.

Проиллюстрируем это на несложном примере с нелинейной зависимостью между возрастом и некоторым событием. Пусть рассчитанный регрессионный коэффициент в уравнении простой логистической регрессии получился отрицательным. Это значит, что вероятность наступления события с возрастом уменьшается (рис. 2*а*).



а *б*

Рисунок 2 – Пример зависимостей между возрастом и событием

Предположим, что была проведена процедура оптимального квантования для признака «Возраст», число конечных классов 4 с границами, проходящими через точки 26, 36 и 45. Диаграмма индексов *WoE* для сформированных конечных классов приведена на рис. 2*б*. Отчетливо видно, что зависимость между возрастом и событием нелинейная: первый (молодые) и последний (старший возраст) возрастной сегмент хорошую реакцию на событие: в данных группах доля событий больше доли не-событий. Промежуточные сегменты, наоборот, не склонны реагировать на событие, о чем говорит положительный индекс *WoE*. Также отметим, что в первом случае уравнение логистической регрессии оказалось статистически незначимо, а во втором – значимо (*P*-значение менее 0,0011).

Экспериментальная часть

Продемонстрируем полезность оптимального квантования в логистической регресии на примере решения реальной открытой задачи по предсказанию отклика

клиентов ОТП банка, предложенной на конкурсе в рамках всероссийской конференции в 2011 г. [5]. Задание заключалось в том, чтобы сформировать алгоритм, который будет выдавать оценку склонности клиента к положительному отклику по его признаковому описанию. Исходная выборка содержит записи о 15 223 клиентах, классифицированных на два класса: 1 – отклик был (1812 клиентов), 0 – отклика не было (13 411 клиентов). Ещё 14 910 записей отложены в качестве тестовых и используются для оценки качества работы алгоритма (индекс *AUC* – площадь под *ROC*-кривой [4]). Записи (признаковые описания) клиентов состоят из 50 признаков, в состав которых входит, в частности, возраст, пол, социальный статус относительно работы, социальный статус относительно пенсии, количество детей, количество иждивенцев, образование, семейное положение, отрасль работы и другие. В данных присутствуют пропуски (иногда – значительные), выбросы, мультиколлинеарность.

Построение модели логистической регрессии на этом наборе данных без применения алгоритмов формирования конечных классов не приводит даже к результату так называемого «случайного классификатора», когда *AUC* = 0,5. Проблема заключается не столько в нелинейных зависимостях, сколько в разнообразии уникальных значений признаков (например, поле «Почтовый адрес» содержит 78 уникальных значений), которые, превращаясь в фиктивные переменные, становятся в модели статистически незначимыми.

Поэтому 45 признаков были поданы на предварительную обработку конечными классами (5 признаков содержали два и менее уникальных значений и не подвергались обработке). Максимальное количество конечных классов везде было ограничено 5. После проведения *WoE*-анализа и расчета *IV* 11 признаков получили среднюю значимость, 15 – низкую. Остальные имели очень низкий информационный индекс и были исключены из дальнейшего моделирования.

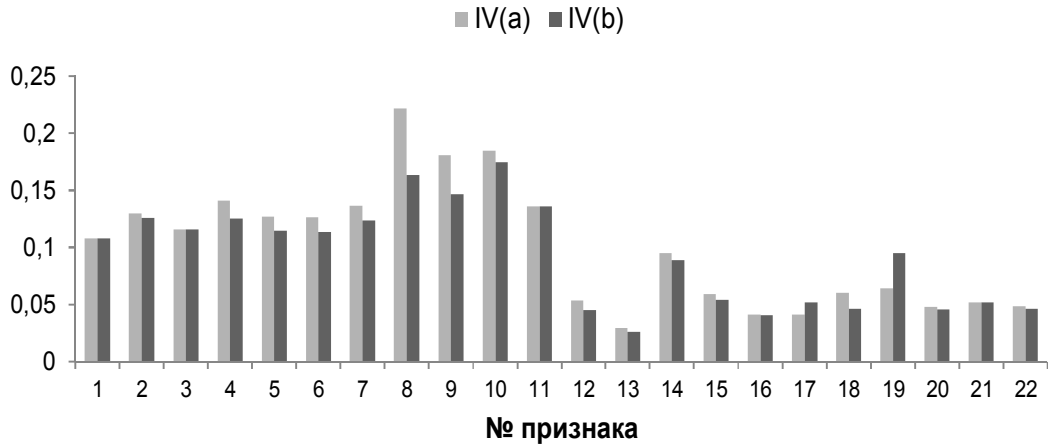


Рисунок 3 – Информационные индексы признаков до и после квантования

На рис. 3 приведена диаграмма с информационными индексами для начальных *IV*(**a**) и конечных *IV*(**b**) классов. Видно, что нигде потери значений информационных индексов не стали большими, и ни один признак не мигрировал из одной категории значимости в другую.

Впоследствии при построении уравнения логистической регрессии были исключены еще 4 признака по причине сильной мультиколлинеарности. В итоге была построена модель прогнозирования вероятности отклика по 22 признакам с числом степеней

свободы $df = 49$ и значимостью $P < 0{,}00001$. Диаграммы *ROC*-кривых для обучающей и тестовой выборок представлены на рис. 4.

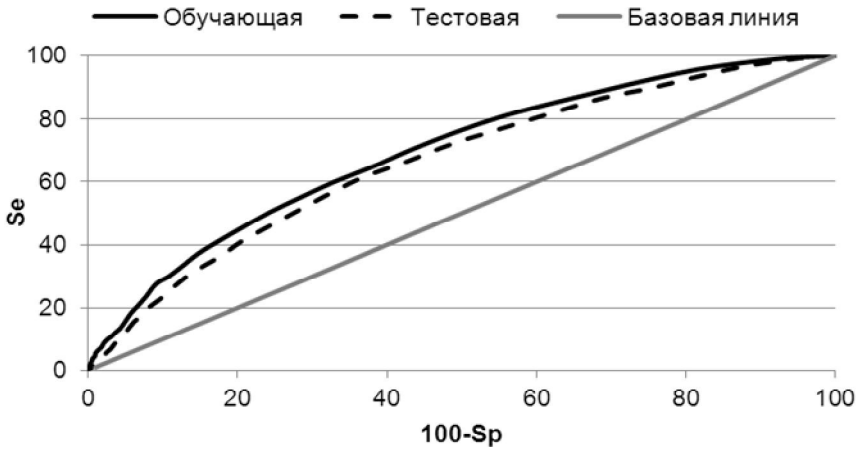


Рисунок 4 – Диаграммы ROC-кривых для задачи клиентов ОТП банка

Расчетное значение индекса *AUC* на тестовой выборке составило 0,6644 (лучший показатель по итогам конкурса равен 0,6935). Учитывая аддитивность связей логистической регрессии и то, что эффекты взаимодействия переменных в уравнении не учитывались, это можно признать отличным результатом. Эффективность оптимального квантования для других классификаторов, таких как деревья решений, нейронные сети, машины опорных векторов, стоит под вопросом и является темой для отдельного исследования.

Литература

1. Breiman L. Combining Predictors, in A.J.C. Sharkey (Ed.) Combining Artificial Neural Nets / L. Breiman // Springer Perspectives in Neural Computing series. – 1999. – P. 30-50.
2. Anderson R. The Credit Scoring Toolkit. Theory and Practice for Retail Credit Risk Management and Decision Automation / Anderson R. – OXFORD University Press, 2007.
3. Гашников М.В. Оптимальное квантование в задаче компрессии цифровых сигналов / М.В. Гашников // Компьютерная оптика. – 2001. – № 21. – С. 179-185.
4. Hosmer D.W. Applied Logistic Regression (Second Edition) / D.W. Hosmer, S. Lemeshow. – Wiley Publishing, Inc., 2000.
5. Конференция ММРО-15. Конкурс по анализу данных. Задача предсказания отклика клиентов ОТП банка. [Электронный ресурс]. – Режим доступа : http://www.machinelearning.ru/wiki/images/d/d5/Mmro-15_contest.pdf.

Literatura

1. Breiman L. Combining Predictors, in A.J.C. Sharkey (Ed.) Combining Artificial Neural Nets, Springer Perspectives in Neural Computing series, 30-50, 1999.
2. Anderson R. The Credit Scoring Toolkit. Theory and Practice for Retail Credit Risk Management and Decision Automation. – OXFORD University Press, 2007.
3. Gashnikov M.V. The optimal quantization in the problem of compression of digital signals // Computer Optics. – 2001. – № 21. – S. 179-185.
4. Hosmer D.W., Lemeshow S. Applied Logistic Regression (Second Edition). – Wiley Publishing, Inc., 2000.
5. Conference MMPR-15. KDD Competition. The task of predicting the response of OTP Bank customers. http://www.machinelearning.ru/wiki/images/d/d5/Mmro-15_contest.pdf.

RESUME

N.B. Paklin, V.V. Afanasiev

Optimal Quantization to Improve the Quality of Binary Classifiers

In the article discusses the optimal quantization algorithms to improve quality of binary classifiers.

Categorization of continuous variables and reducing the number of distinct values is based on the calculation of WoE and information values (also called «fine and coarse classing» or «optimal binning»). Based on the idea of a recursive partition optimal binning algorithm is proposed with the objective function that maximizes the variable information index. On the open competition prediction customer responses task we show that the optimal quantization can significantly improve the quality and statistical significance logistic regression model, especially if the input variables are measured on a scale with many gradations and when relationships between input variables and output field (called event) are nonlinear.

The developed optimal quantization algorithms are became part of software «Analytical Platform Deductor 5.3» (BaseGroup Labs Company, Russia).

Статья поступила в редакцию 08.07.2013.