

КОНСТРУИРОВАНИЕ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ ДЛЯ АНАЛИЗА ЭФФЕКТИВНОСТИ МЕТОДОВ КЛАССИФИКАЦИИ

В. М. Неделько^{1,2}

¹ *Институт математики им. С. Л. Соболева СО РАН, 630090, Новосибирск*

² *Новосибирский государственный технический университет, 630090, Новосибирск*

УДК 519.246

В работе рассматривается проблема конструирования синтетических данных для анализа эффективности методов классификации. Предлагается семейство вероятностных моделей, которые позволяют отражать такие свойства реальных данных, как сложность закономерностей, степень независимости переменных, степень зашумлённости.

Моделирование с использованием специальным образом сконструированных синтетических данных позволяет не только оценить, какой из исследуемых методов более эффективен, но и понять за счёт чего эта эффективность достигается.

Ключевые слова: вероятностная модель, распознавание образов, машинное обучение, вероятность ошибочной классификации, бустинг, риск.

Введение

В настоящее время boosting является одним из методов классификации, наиболее часто используемых для решения практических задач интеллектуального анализа данных (машинного обучения). Если проанализировать решения задач на портале kaggle.com, то можно заметить, что в большинстве лучших решений используется бустинг, а именно, алгоритм XGBoost. При этом причины такой высокой эффективности бустинга в настоящее время до конца не исследованы [1], [2].

Как известно, почти для любого метода классификации можно найти (или придумать) такие данные, на которых он будет работать хорошо [3]. Если метод хорошо работает на реальных данных, значит он «угадывает» какие-то существенные свойства этих данных. Однако реальные данные обычно обладают одновременно большим числом свойств, поэтому крайне затруднительно понять, какие из них благоприятны для заданного метода, а какие — нет. Эту проблему можно решить исследованием методов на синтетических данных, свойства которых задаёт исследователь.

Задачей настоящего исследования является в том числе выявление таких особенностей данных, которые обеспечивают преимущество бустинга над другими методами. Для этого предлагается семейство вероятностных моделей, которые позволяют отражать такие свойства данных, как сложность закономерностей, степень независимости переменных, степень зашумлённости.

1 Задача построения решающей функции

Пусть X — пространство значений переменных, используемых для прогноза, а $Y = \{-1, 1\}$ — пространство значений прогнозируемых переменных, и пусть \mathcal{C} — множество всех вероятностных мер на заданной σ -алгебре подмножеств множества $D = X \times Y$. При каждом $c \in \mathcal{C}$ имеем вероятностное пространство $\langle D, \mathcal{B}, P_c \rangle$, где \mathcal{B} — σ -алгебра, P_c — вероятностная мера.

Решающей функцией называется соответствие $\lambda: X \rightarrow Y$.

Качество принятого решения оценивается заданной функцией потерь $\mathcal{L}: Y^2 \rightarrow [0, \infty)$. Под риском будем понимать средние потери:

$$R(c, \lambda) = E\mathcal{L}(y, \lambda(x)) = \int_D \mathcal{L}(y, \lambda(x)) P_c(dx, dy), \quad x \in X, y \in Y. \quad (1)$$

При $\mathcal{L}(y, \tilde{y}) = I(y \neq \tilde{y})$, где $I(\cdot)$ — индикаторная функция (равна 1, если условие истинно, и 0 — иначе), риск есть вероятность ошибочной классификации.

Задача классификации заключается в построении решающей функции, которая бы минимизировала риск.

Заметим, что значение риска зависит от c — распределения, которое неизвестно. Поэтому в приведённой формулировке задача некорректна. Однако полностью строгой общей постановки задачи распознавания в настоящее время не существует. На практике либо решаются более частные строго поставленные задачи, либо разрабатываются эвристические методы, например, методы, минимизирующие выборочную оценку риска.

Пусть $V = ((x^i, y^i) \in D \mid i = 1, \dots, N)$, $V \in D^N$ — случайная независимая выборка из распределения P_c . Метод построения решающих функций есть отображение

$$Q: D^N \rightarrow \Lambda,$$

где Λ — заданный класс решающих функций, а $\lambda_{Q,V}$ — функция, построенная по выборке V методом Q .

Для оценивания риска обычно используются эмпирические функционалы качества, т. е. точечные оценки риска, такие как эмпирический риск, оценка скользящего экзамена, оценка bootstrap и т.п.

Эмпирический риск определяется как средние потери на выборке:

$$\tilde{R}(V, \lambda) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y^i, \lambda(x^i)).$$

Как правило, методы классификации в той или иной степени минимизируют эмпирический риск. Однако сама по себе минимизация эмпирического риска не гарантирует минимизацию риска, и тот факт, что некоторый алгоритм эффективно минимизирует эмпирический риск, ещё не объясняет эффективность этого алгоритма на реальных задачах.

Приведённые понятия являются базовым набором понятий в задаче распознавания. Однако для дальнейших рассуждений нам потребуется ввести некоторые их обобщения.

2 Вероятностная модель

Основная проблема при подборе вероятностных моделей заключается в том, что разнообразие возможных моделей слишком велико, поэтому сложно обосновать произвол в выборе конкретной модели. В связи с этим на практике обычно ограничиваются самыми простыми моделями, например нормальными распределениями.

В данной работе предлагается семейство моделей, достаточно простое по структуре, но позволяющее моделировать достаточно сложные закономерности в данных, а также свойства данных, которые существенно влияют на точность классификации.

2.1 Базовая модель

Предлагаемое семейство моделей основано на комбинировании простых моделей, которые назовём «базовыми».

В качестве базовых естественно взять модели, основанные на смеси нормальных распределений

$$\mathcal{N}(x_\theta, \Sigma_\theta), \quad \theta = 1, \dots, \Theta, \quad (2)$$

где Θ — число компонент смеси, x_θ , Σ_θ — параметры распределений.

Чтобы полностью определить модель, осталось задать вероятности компонент $P(\theta)$ и отображение компонент в классы $y(\theta)$.

Рассмотренный вид модели позволяет отражать разнообразные конфигурации классов. На рис. 1 приведён пример конфигурации, известной как модель «шахматной доски» или «исключающего или».

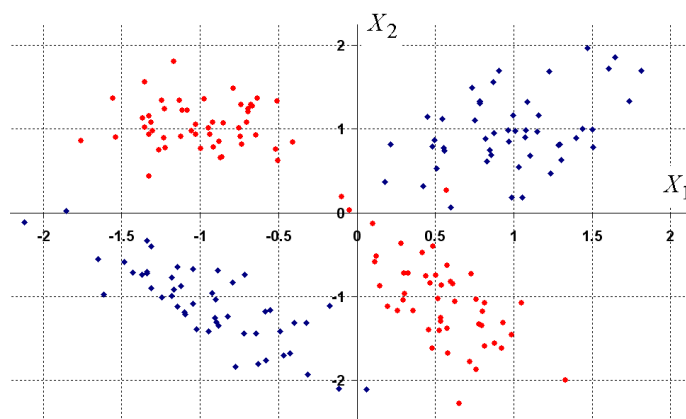


Рис. 1: Пример распределения классов для базовой вероятностной модели.

2.2 Неинформативные переменные

Неинформативные или «шумовые переменные» встречаются в большинстве реальных задач анализа данных, поэтому при исследовании эффективности методов классификации такие переменные необходимо включать в модель.

Шумовые переменные — это те, для которых распределения обоих классов совпадают. В рамках базовой модели это означает совпадение числа компонент и параметров.

2.3 Независимые переменные

Независимость переменных — это одно из основных свойств данных, которые позволяют повышать точность классификации. Если с ростом числа зависимых переменных точность решения обычно уменьшается, то с добавлением независимых переменных — растёт. При этом под независимостью понимается условная независимость (переменные независимы для каждого класса по отдельности).

Заметим, что в «чистом виде» независимость в реальных задачах встречается редко. Как правило, переменные являются частично зависимыми, а сама независимость имеет место не между отдельными переменными, а между подмножествами переменных.

Чтобы построить вероятностную модель с учётом эффекта независимости, достаточно взять несколько базовых моделей, каждую на своём подмножестве переменных.

Выборка в соответствии с такой моделью генерируется следующим образом. Сначала «разыгрывается» (выбирается случайно в соответствии с заданными безусловными вероятностями) номер класса. После этого для каждой базовой модели «разыгрывается» номер компоненты (из тех, что соответствуют выбранному классу). Наконец, для каждой выбранной компоненты генерируются значения переменных.

2.4 Обобщённая модель

Номер компоненты θ_l для базовой модели l можно рассматривать как дискретную случайную величину.

До сих пор мы номер класса рассматривали как функцию отдельных компонент.

В общем случае номер класса может быть функцией $y(\theta_1, \dots, \theta_k)$ всего вектора номеров компонент. При этом на θ_l может быть задано произвольное совместное распределение.

Таким способом можно задавать достаточно сложные модели, например, можно задать многомерную модель «исключающего или» как $y(\theta_1, \dots, \theta_k) = ((\theta_1 + \dots + \theta_k) \bmod 2) \cdot 2 - 1$.

Дальнейшее обобщение модели возможно путём создания смеси моделей, т.е. объединения в одной выборке данных, сгенерированных на основе различных моделей.

3 Случай независимых переменных

Методы построения решающих функций, использующие гипотезу независимости переменных, в зарубежной терминологии принято называть наивными байесовскими классификаторами.

В большинстве реальных задач предполагать независимость переменных нет оснований, однако с теоретической точки зрения этот случай очень важен. Если в общем случае чем больше переменных, тем больший объём выборки требуется (при прочих равных) для построения решающей функции, то при независимых переменных ситуация противоположная: чем больше переменных, тем меньший объём выборки требуется для достижения той же точности распознавания.

Гипотеза независимости переменных является очень сильной, и естественно желание иметь промежуточные варианты.

Для пространства бинарных переменных такая модель известна: это ряд Бахадура, который позволяет последовательно усложнять модель (ослаблять гипотезу) — от полной независимости, через учёт только парных зависимостей, зависимостей в тройках и т.д. до общего случая любых зависимостей.

К сожалению, обобщения ряда Бахадура для случая переменных произвольных типов автору неизвестны.

Более полную информацию, по сравнению с решающей функцией, несёт функция условной вероятности $g(x) = P(y = 1 | x)$, которая определяет вероятность заданного класса в каждой точке пространства переменных.

Выведем формулу для вычисления условной вероятности в случае независимых переменных.

Из формулы Байеса можем записать

$$g(x) = P(y = 1 | x) = \frac{P(dx, y = 1)}{P(dx, y = 1) + P(dx, y = -1)} = \frac{1}{1 + \frac{1-p}{p} \cdot \frac{P(dx|y=-1)}{P(dx|y=1)}},$$

где $p = P(y = 1)$.

Пусть условные распределения всех переменных X_j при условии обоих классов независимы, т.е. $P(dx | y) = \prod_{j=1}^n P(dx_j | y)$.

Подставив это произведение в предыдущее выражение, после преобразований имеем

$$\frac{p}{1-p} \cdot \left(\frac{1}{g(x)} - 1 \right) = \prod_{j=1}^n \frac{p}{1-p} \cdot \left(\frac{1}{g_j(x_j)} - 1 \right),$$

где $g_j(x_j) = P(y = 1 | x_j) = \frac{P(dx_j, y=1)}{P(dx_j)}$.

Логарифмируем последнее выражение и получаем

$$\sigma^{-1}(g(x)) = (n-1)(\ln p - \ln(1-p)) + \sum_{j=1}^n \sigma^{-1}(g_j(x_j)),$$

где $\sigma^{-1}(\cdot)$ — функция, обратная сигмоиду $\sigma(z) = \frac{1}{1+e^{-z}}$.

Заметим, что полученное выражение имеет вид логистической регрессии, а именно

$$g(x) = \sigma \left(u_0 + \sum_{j=1}^n u_j \sigma^{-1}(g_j(x_j)) \right),$$

при $u_0 = (n-1)(\ln p - \ln(1-p))$, $u_j = 1$.

Обычно логистическую кривую получают, исходя из предположений о виде распределения, однако сейчас мы предположили независимость переменных, но не ограничивали вид распределений.

Данное выражение справедливо не только при независимых переменных, а в несколько более общем случае, поскольку здесь мы имеем лишь одно соотношение, а независимость переменных требует выполнения мультипликативности условного распределения для каждого класса, что даёт число соотношений по числу классов.

Ещё более расширить область применимости можно, если считать веса свободными параметрами.

Дальнейшее обобщение возможно, если допустить произвольные оценочные функции

$$g(x) = \sigma \left(u_0 + \sum_{j=1}^n u_j s_j(x_j) \right). \quad (3)$$

Мы получили метод, который можно считать разновидностью метода логистической регрессии, а также разновидностью наивного байесовского классификатора (хоть он и не является частным случаем последнего).

В рамках данной статьи будем называть метод, дающий решения, представимые в форме (3), обобщённым наивным байесовским классификатором.

Формулу (3) можно естественным образом обобщить по аналогии с рядом Бахадура, включив возможность учитывать зависимости между переменными, последовательно добавляя парные зависимости, зависимости в тройках и т.д.

$$g(x) = \sigma \left(u_0 + \sum_{j=1}^n u_j s_j(x_j) + \sum_{j,k} u_{jk} s_{jk}(x_j, x_k) + \sum_{j,k,l} u_{jkl} s_{jkl}(x_j, x_k, x_l) + \dots \right). \quad (4)$$

Оказывается, что бустинг как раз и позволяет строить подобные решения. Например, метод AdaBoost на деревьях с числом конечных вершин не более M соответствует (4) с ограничением суммирования вплоть до слагаемых, соответствующих всевозможным сочетаниям из n по M переменных.

4 Эффективность методов классификации

Хотя эффект независимости переменных позволяет значительно повысить точность решения, разные методы в разной степени способны его использовать.

Пусть вероятностная модель представлена k независимыми базовыми моделями, и пусть $1 - R(\lambda_1^*)$ — точность классификации, которую заданный метод достигает на каждой базовой модели отдельно. Тогда максимально достижимая точность на полной модели есть

$$1 - R(\lambda_k^*) = \sigma(k * \sigma^{-1}(1 - R(\lambda_1^*))).$$

Эта величина изображена на рис. 2.

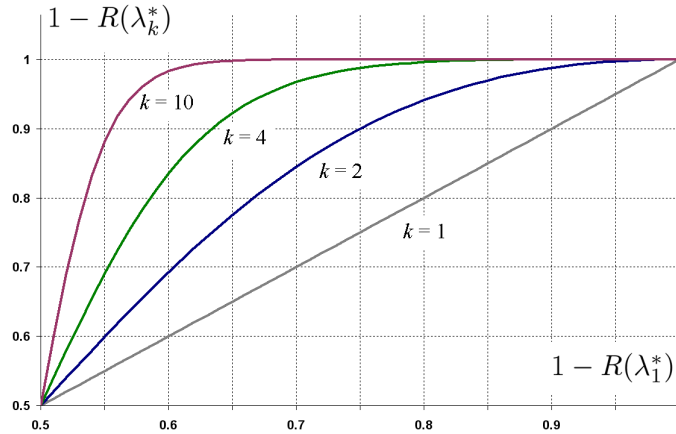


Рис. 2: Зависимость максимальной точности классификации от числа независимых переменных.

Заметим, что такое повышение точности действительно достижимо, но только в частных случаях, например, когда все переменные бинарные.

Более подходящую для практических целей оценку (см. рис. 3) можно получить в предположении нормальности распределений

$$1 - R(\lambda_k^N) = F_N \left(\sqrt{k} * F_N^{-1}(1 - R(\lambda_1^N)) \right),$$

где $F_N(\cdot)$ — функция стандартного нормального распределения.

Приведённые оценки эффекта независимости могут быть использованы для оценки того, насколько хорошо тот или иной метод классификации способен использовать независимость переменных.

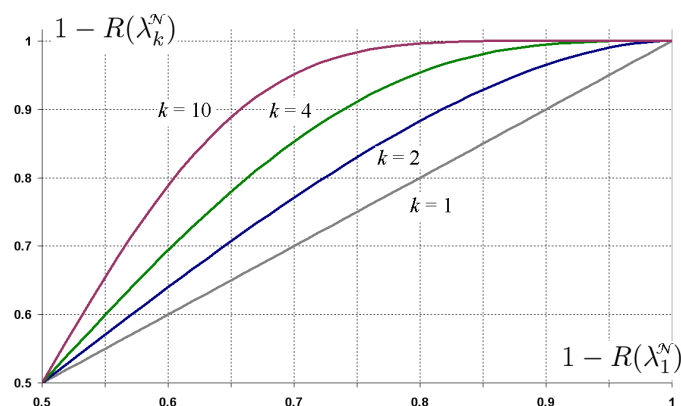


Рис. 3: Зависимость точности классификации от числа независимых переменных для нормальной модели.

Предложенное семейство вероятностных моделей было использовано в работе [4] для исследования эффективности ансамблевых методов классификации, таких как бустинг и нейронные сети (прямого распространения).

В результате удалось выявить некоторые важные свойства исследуемых методов, в частности, выявлено различие в устойчивости к шумам, а также различие в способности использовать эффект независимости (бустинг оказался лучшим по обоим характеристикам).

Заключение

Предложенное в работе семейство вероятностных моделей, несмотря на простоту, предоставляет широкие возможности по имитации реальных данных с заданными свойствами, в частности, позволяет варьировать сложность вероятностных зависимостей, моделировать независимость переменных, а также «шумовые» переменные.

Всестороннее изучение возможностей данного семейства требует большого объёма исследований в рамках целого цикла работ, однако уже предварительные исследования показывают его полезность.

Моделирование с использованием специальным образом сконструированных синтетических данных позволяет не только оценить, какой из исследуемых методов более эффективен, но и понять за счёт чего эта эффективность достигается.

Список литературы

- [1] David Mease and Abraham Wyner. Evidence contrary to the statistical view of boosting // Journal of Machine Learning Research, 9:131–156, 2008.
- [2] Неделько В.М. К вопросу об эффективности бустинга в задаче классификации // Вестник Новосибирского государственного университета. Математика, механика, информатика. 2015. Т. 15, вып. 2. С. 72–89.
- [3] Неделько В.М. Исследование эффективности некоторых линейных методов классификации на модельных распределениях // Машинное обучение и анализ данных, 2016. Т. 2. № 3. С. 305–328.
- [4] Бутаев Р.С. Сравнение эффективности методов построения решающих функций на основе бустинга и нейронных сетей. ВКР. 2007. 36 с.

Виктор Михайлович Неделько — к.ф.-м.н., ст. науч.сотр. Института математики им. С. Л. Соболева СО РАН;
e-mail: nedelko@math.nsc.ru.

Дата поступления — 3 июня 2017 г.