Комбинированный порождающий и разделяющий подход в задачах классификации с малой выборкой*

Животовский Н. К.
nikita.zhivotovskiy@phystech.edu
Московский физико-технический институт

В работе рассмотрены два статистических подхода к решению задачи классификации и способ их комбинации, предназначенный для оценки параметров классификатора по выборкам различной мощности. Для случая, когда объекты в классах имеют многомерное нормальное распределение, построена комбинированная модель, сочетающая в себе порождающий и разделяющие подходы к задачам классификации. В серии экспериментов показано, что при ограничениях на длину обучающей выборки использование этой модели может приводить к уменьшению вероятности ошибки получаемого классификатора по сравнению с чисто порождающими или разделяющими моделями.

Ключевые слова: *классификация, порождающая и разделяющая модели, логистическая регрессия.*

Combined generative and descriminative approach for calssification with a small learning set*

Zhivotovskiy N. K.
Moscow Institute of Physics and Tecnology

This paper deals with two statistical approaches to solving classification problems and way of their combination designed to evaluate the parameters of a classifier for samples of different cardinality. The combined descriminative and generative model was built for the case of the multivariate normal distribution of objects within classes. This model shows lower probability of error of classificator as compared with one obtained purely from generative or descriminative model when restrictions are put on the size of the learning set

Keywords: *classification, generative and descriminative approaches, logistic regression.*

Введение

Исследуется комбинированный порождающий и разделяющий подход в задачах классификации, описанный в [1]. Задача классификации заключается в нахождении оптимального, например, с точки зрения вероятности ошибки зрения правила, которое относит объекты, представленные точками конечномерного действительного векторного пространства к одному из конечного числа классов. Построенное правило называется *классификатором*. Классификатор выбирается из множества правил, которые называются *моделью*. Модель описывается в виде параметрически заданного семейства функций, отображающих множество объектов во множество классов. При оценке параметров модели с целью выбора оптимального классификатора используется заранее известная конечная выборка объектов, называемая *обучающей*, для которой уже известен класс каждого из входящих в неё объектов.

Научный руководитель В.В. Стрижов
Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00709.

Используемые модели подразделяется на *разделяющие* и *порождающие* [1, 2]. Оценка параметров в разделяющих моделях можно рассматривать как подбор таких значений параметров модели, которые максимизирует правдоподобие обучающей выборки по отношению к вероятности класса [2]. Классификатор в таком случае относит объект к его наиболее вероятному классу. Альтернативный подход, называющийся *порождающим* [1, 2], заключается в максимизации правдоподобия совместного распределения объектов и классов, а затем в использовании формулы Байеса для нахождения вероятности отношения объекта к классу.

В работе [6] производится сравнение обоих подходов на примере логистической регрессии [2], для которой параметры оцениваются исходя из разделяющего подхода, и наивного Байесовского классификатора [2], для которого оценка параметров максимизирует функцию совместного правдоподобия объектов и классов. Результаты теоретического и экспериментального исследований подтверждают, что с ростом длины обучающей выборки разделяющий подход приводит к меньшей вероятности ошибки, то есть к лучшему качеству классификации.

Однако, во-первых, для получения меньшей вероятности ошибки разделяющий подход требует большую длину обучающей выборки в то время как порождающий подход достигает своего асимптотического по длине обучающей выборки минимума вероятности ошибки гораздо быстрее. Во-вторых, для малых длин выборок на 14 из 15 рассмотренных в статье задачах из репозитория UCI порождающий подход даёт в среднем лучшее качество классификации.

Таким образом, ни разделяющий, ни порождающий подход не является строго предпочтительным для всех длин выборок и для всех задач. Поэтому в целях улучшения качества классификации в ряде работ исследуется идея комбинированного подхода. В [1] и [4] для модельных данных, а также для задач классификации изображений, показано, что с помощью выбора подходящей модели комбинация двух подходов может улучшать качество классификации по сравнению с каждым из подходов по отдельности. В частности, в [1] комбинированный подход к оценке параметров модели заключается в замене правдоподобия обучающей выборки на выпуклую комбинацию логарифмов правдоподобий, относящихся соответственно к разделяющему и порождающему подходам. Альтернативные подходы к построению комбинированных моделей для задач классификации изображений приводятся, например, в [3].

Постановка задачи

Пусть $\{\mathbf{x}_i, y_i\}_{i=1}^{\ell}$ — конечная обучающая выборка, выбранная независимо из некоторого неизвестного совместного распределения объектов и классов, а $\{1, -1\}$ — множество классов. Будем считать, что $P(y|\mathbf{x}, \boldsymbol{\theta})$ — вероятность принадлежности объекта $\mathbf{x}$ классу y и $p(y, \mathbf{x}|\boldsymbol{\theta})$ — совместная плотность распределения объектов и классов задаются общим набором параметров $\boldsymbol{\theta}$, априорно неизвестных.

Согласно [1] предполагается, что искомые значения параметров максимизируют выпуклую оболочку логарифма разделяющего правдоподобия обучающей выборки

$$L_D = \sum_{i=1}^{\ell} \log\bigl(p(y_i|\mathbf{x}_i, \boldsymbol{\theta})\bigr) \tag{1}$$

зависящего от вероятности принадлежности объекта выборки к классу и логарифма её порождающего правдоподобия

$$L_G = \sum_{i=1}^{\ell} \log\bigl(p(y_i, \mathbf{x}_i|\boldsymbol{\theta})\bigr) \tag{2}$$

зависящего от совместной плотности объектов и классов. Общая формула имеет следующий вид:

$$\lambda L_D + (1-\lambda)L_G = \lambda \sum_{i=1}^{\ell} \log\bigl(p(y_i|\mathbf{x}_i, \boldsymbol{\theta})\bigr) + (1-\lambda) \sum_{i=1}^{\ell} \log\bigl(p(y_i, \mathbf{x}_i|\boldsymbol{\theta})\bigr), \quad \lambda \in [0,1] \tag{3}$$

Тогда поиск оптимальных значений параметров заключается в максимизации этой взвешенной суммы правдоподобий:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} (\lambda L_G + (1-\lambda)L_D)$$

Выбор параметра λ определяет значимость каждого из подходов.

Построение комбинации правдоподобий

В качестве основной модели, с помощью которой будет иллюстрироваться комбинированный подход в данной работе принята логистическая регрессия [2]. В этой модели предполагается, что вероятность принадлежности объекта $\mathbf{x}$ к классу 1 задаётся в виде формулы:

$$P(1|\mathbf{x}) = \sigma((\mathbf{w}, \mathbf{x})),$$

где $\sigma(z) = \frac{1}{1+\exp(-z)}$ — сигмоидная функция, определённая для всех действительных z. Оценка параметров в случае логистической регрессии заключается в поиске такого значения вектора параметров $\mathbf{w}$, которое максимизирует логарифм правдоподобия обучающей выборки:

$$L = \sum_{i=1}^{\ell} \log\bigl(y_i \sigma(\mathbf{w}, \mathbf{x}_i)\bigr) \to \max_{\mathbf{w}}$$

Оценка параметров этой модели соответствует разделяющему подходу (1). В данной работе чисто разделяющая функция правдоподобия (1) будет заменена на выпуклую комбинацию разделяющего и порождающего (3).

Предполагается, что объекты $\mathbf{x} \in \mathbb{R}^n$, а функции правдоподобия $p_y(\mathbf{x})$ (плотности распределения объектов при фиксированном классе обозначаются соответственно $p_1(\mathbf{x})$ и $p_{-1}(\mathbf{x})$) имеют многомерное нормальное распределение со средними $\boldsymbol{\mu}_y$ и ковариационной матрицей $\boldsymbol{\Sigma}$, общей для обоих классов. Пусть P_1 — априорная вероятность класса $+1$, тогда вероятность класса -1 равна $1 - P_1$.

В таком случае можно рассчитать апостериорные вероятности классов. Заметим, что функция правдоподобия классов в нашем случае имеет вид:

$$p_y(\mathbf{x}) = \exp\bigl(\boldsymbol{\mu}_y^T \boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_y^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y - \tfrac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{n}{2}\ln(2\pi) - \tfrac{1}{2}\ln(|\boldsymbol{\Sigma}|)\bigr).$$

Таким образом,

$$\frac{P(+1|\mathbf{x})}{P(-1|\mathbf{x})} = \frac{P_1 p_1(\mathbf{x})}{(1-P_1)p_{-1}(\mathbf{x})} = \frac{P_1}{(1-P_1)} \exp\bigl((\mathbf{w}, \mathbf{x}) + c\bigr),$$

где $\mathbf{w} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})^T \boldsymbol{\Sigma}^{-1},\ c = (\mathbf{w}, \boldsymbol{\mu}_1 + \boldsymbol{\mu}_{-1})$. Так как классов всего два, то

$$P(+1|\mathbf{x}) + P(-1|\mathbf{x}) = 1$$

Отсюда с учётом полученного равенства получаем:

$$\mathsf{P}(1|\mathbf{x}) = \frac{1}{1 + \frac{1-P_1}{P_1}\exp(-(\mathbf{w},\mathbf{x}) - c)}, \quad \mathsf{P}(-1|\mathbf{x}) = \frac{1}{1 + \frac{P_1}{1-P_1}\exp((\mathbf{w},\mathbf{x}) + c)}.$$

Формулой Байеса позволяется получить совместную плотность распределения $p(y, \mathbf{x})$:

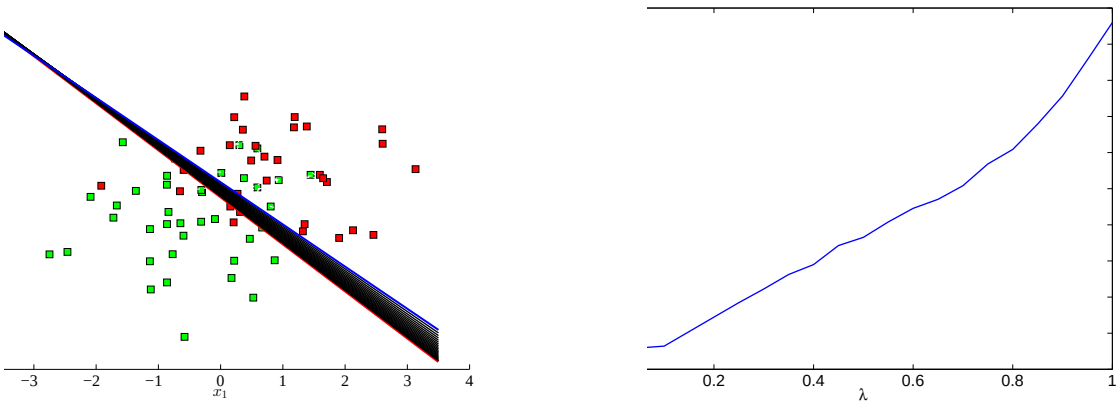
$$p(y, \mathbf{x}) = p(y|\mathbf{x})\big(P_1 p_1(\mathbf{x}) + (1 - P_1)p_{-1}(\mathbf{x})\big).$$

Подстановка полученных результатов в формулу для выпуклой комбинации правдоподобий даёт:

$$L_\lambda = -\sum_{i=1}^{\ell}\Big(\log\big(1 + (\tfrac{1-P_1}{P_1})^{y_i}\exp(-y_i((\mathbf{w},\mathbf{x}_i) + c))\big) + (1-\lambda)\log\big(P_1 p_1(\mathbf{x}_i) + (1 - P_1)p_{-1}(\mathbf{x}_i)\big)\Big).$$

Стоит отметить, что L_λ зависит лишь от $\boldsymbol{\mu}_1, \boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}, P_1$. Таким образом, получена оптимизационная задача:

$$L_\lambda \ \to \ \max_{\boldsymbol{\mu_1},\boldsymbol{\mu_{-1}},\boldsymbol{\Sigma},P_1}.$$



(a) Обучающая выборка и разделяющие прямые, соответствующие разным значениям λ

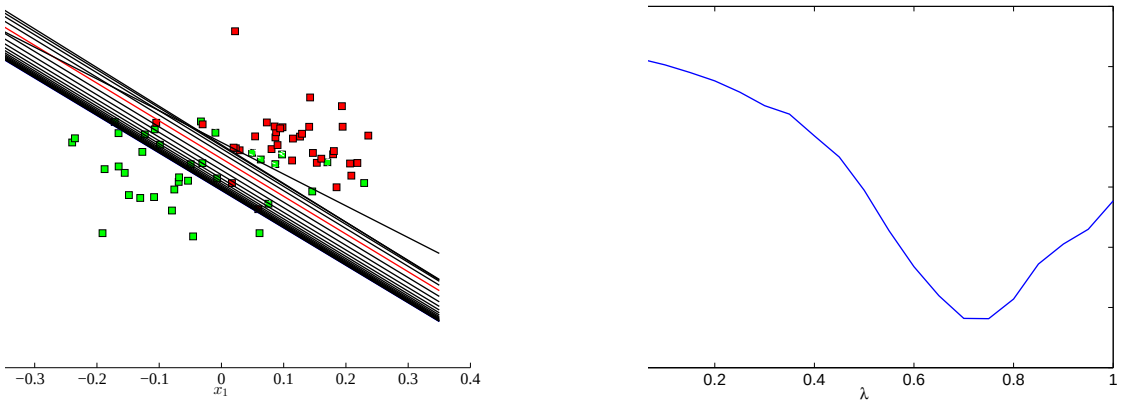
(b) Частота ошибок классификатора в зависимости от значения параметра λ

Рис. 1: Случай разреженных классов.

Вычислительный эксперимент

Для иллюстрации комбинированного подхода к обучению производится серия экспериментов. Для двух классов генерируются обучающая выборка, выбранная из двумерного мерного нормального распределения. Ради упрощения вычислений в эксперименте предполагается, что ковариационная матрица $\boldsymbol{\Sigma}$ известна и является диагональной с σ^2 на диагонали. Параметр σ^2 будет изменяться в эксперименте и позволит задавать дисперсии объектов в классах. Предполагается, что априорные вероятности классов равны. Координаты средних значений правдоподобий классов $\boldsymbol{\mu}_y$ при этом равны $(-0.7\sigma, -0.7\sigma)$ для

класса $y = -1$ и $(0.7\sigma, 0.7\sigma)$ для класса $y = 1$. Выбор таких средних позволяет с одной стороны достичь некоторого смешения классов, а с другой стороны производит их кластеризацию вокруг удалённых друг от друга средних. При этом чем меньше дисперсия σ^2, тем меньше и расстояние между классами, то есть выборка не будет линейно разделима практически при всех значениях σ^2. Множитель 0.7 задаёт соотношение между дисперсией классов и их средним и характеризует степень смешения классов. Чем меньше его значение, тем меньше расстояние между классами и тем хуже они разделяются прямой.



(a) Обучающая выборка и разделяющие прямые, соответствующие разным значениям λ

(b) Частота ошибок классификатора в зависимости от значения параметра λ

Рис. 2: Случай классов с малой дисперсией.

Тем не менее, параметры $\boldsymbol{\mu}_y$ считаются неизвестными и оцениваются при максимизации правдоподобия. Для эксперимента создавалась генеральная выборка из $N = 10000$ прецедентов из описанного распределения. Из неё случайным образом выбиралась обучающая выборка из ℓ элементов. Пусть $\mathbf{x}_i$ – i-ый объект выборки, а выбранные ℓ объектов имеют соответственно номера $1, \ldots, \ell$. Индикатор ошибки классификатора на $\mathbf{x}_i$ обозначается как $I(\mathbf{x}_i)$. Тогда критерием качества классификатора будет минимум частоты ошибок на объектах, не попавших в обучающую выборку:

$$\frac{\sum\limits_{i=\ell+1}^{N} I(\mathbf{x}_i)}{N - \ell} \to \min$$

Аналогично результату из [6] на больших обучающих выборках разделяющий подход, соответствующий $\lambda = 0$, показывает лучшее качество классификации, то есть меньшую частоту ошибок на оставшихся $N - \ell$ объектах.

Выборки, длина которых меньше $30 - 40$ объектов, не позволяют понять важность каждого из подходов, так как дисперсия частоты ошибок получаемого классификатора слишком велика. Поэтому в качестве промежуточного значения выбрано $\ell = 70$.

Таким образом, $L_\lambda = L_\lambda(\boldsymbol{\mu}_1, \boldsymbol{\mu}_{-1})$. Для поиска параметров распределения, максимизирующих введённое правдоподобие, был использован оптимизационный toolbox yalmip языка Matlab. В случае комбинированного правдоподобия как и в случае простой логистической регрессии градиентные методы успешно находят локальные максимумы.

Случай разреженных классов.

Сначала рассматривается случай $\sigma = 1$.

На рис. 81 изображено множество разделяющих прямых, которые получаются при различных значениях параметра λ. По осям отложены координаты объектов. Можно заметить, что в данном случае происходит движение разделяющей прямой при изменении λ. Верхняя прямая в пучке при этом соответствует разделяющему подходу, а нижняя — порождающему. Красным цветом выделена прямая, которая доставляет наименьшую частоту ошибок на генеральной выборке. Синим цветом — доставляющая наибольшую частоту ошибок.

Частота ошибок на генеральной выборке, отложенная по оси ординат, в зависимости от λ изображена на рис. 82.

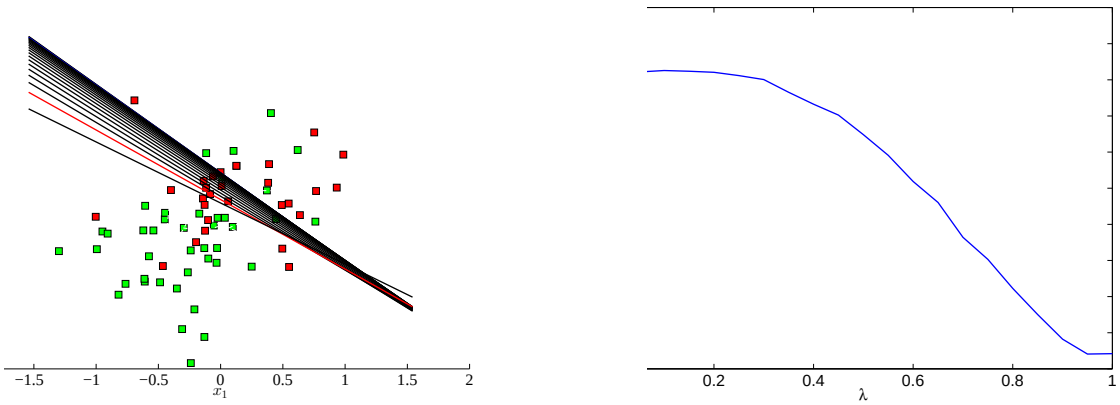
В случае, когда дисперсия классов σ^2 велика, большинство объектов удалено от всего множества разделяющих прямых. Таким образом, использование взвешенной функции правдоподобия практически не изменяет частоту ошибок получаемого классификатора.

Случай классов с малой дисперсией.

Пусть теперь $\sigma = 0.1$. Как показывают рис. 83 и рис. 84 в этом случае комбинированный подход позволяет существенно улучшить качество классификации.

Однако дальнейшее уменьшение параметра σ не позволяет получить уменьшение частоты ошибки при использовании комбинированного подхода. Более того, прямые, соответствующие значениям λ не равным нулю или единице, даже ухудшают качество классификации.

Случай средней разреженности. В качестве промежуточного случая рассматривается $\sigma = 0.44$. Соответствующие этому случаю иллюстрации изображены на рисунках 85 и 86.



(a) Обучающая выборка и разделяющие прямые, соответствующие разным значениям λ

(b) Частота ошибок классификатора в зависимости от значения параметра λ.

Рис. 3: Случай средней разреженности.

В этом случае комбинированный подход позволяет получить некоторое улучшение качества классификации.

Заключение

Серия экспериментов показывает, что для малых обучающих выборок комбинированный подход к оценке параметров позволяет в некоторых случаях улучшить качество клас-

сификации. В отличие от логистической регрессии результат классификации при комбинированном подходе существенно зависит от масштаба вектора. Действительно, выбор параметра дисперсии просто изменяет координаты объекта, являющегося действительным вектором, в одно и то же число раз.

В случае больших значений дисперсий малая часть объектов попадает в окрестность тех прямых, которые разделяют классы при разных значениях параметра, регулирующего вклад каждого из подходов. Поэтому качество классификации практически не меняется, если использовать комбинированный подход.

В то же время при меньших значениях дисперсии удаётся существенно улучшить качество классификации, используя комбинированный подход.

Литература

[1] *C. M. Bishop, J. Lasserre* Generative or Discriminative? Getting the Best of Both Worlds (2007) // Bayesian Statistics 8, — С. 3–24.

[2] *C. M. Bishop* Pattern Recognition and Machine Learning (2006) // Springer, Series: Information Science and Statistics 8, — С. 740 pp.

[3] *Perina, A. , Cristani, M. , Castellani, U. , Murino, V. , Jojic, N.* A hybrid generative/discriminative classification framework based on free-energy terms (2009) // Computer Vision, 2009 IEEE 12th International Conference, — С. 2058 – 2065 .

[4] *J. Lasserre, C. M. Bishop, T. Minka* Principled Hybrids of Generative and Discriminative Models (2006) // Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1, — С. 87–94.

[5] *Liang, P. , Jordan M. I.* An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators (2008) // Proceedings of the 25th international conference on Machine learning, — С. 584-591.

[6] *Ng, A. Y., Jordan, M. I.* On discriminative vs. generative: A comparison of logistic regression and naive Bayes (2002) // Advances in Neural Information Processing Systems 14, Cambridge, MA: The MIT Press, — С. 841–848.