Вероятностная модель одноклассовой классификации*

Бурмистров М. О., Сандуляну Л. Н.
burmisha@gmail.com, liubov.sanduleanu@gmail.com
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

Решается задача одноклассовой классификации электронных писем на предмет наличия в них спама. В работе вводится квазивероятностная модель для классической эмпирической постановки задачи одноклассовой классификации и производится сведение классического подхода к новой модели. Построенные методы классификации проверяются вычислительными экспериментами на модельных и реальных данных.

Ключевые слова: *одноклассовая классификация, вероятностная модель, байесовский подход, ядерные функции.*

Probabilistic model for one-class classification problem*

Burmistrov M. O., Sanduleanu L. N.
Moscow Institute of Physics and Technology

One-class classification methods are used to test e-mails for spam. Quasi-probabilistic model is introduced for traditional empirical approach to problem. The old model is shown to be a reduction of the new one. Built approaches to classification are numerically tested on model and real data.

Keywords: *one-class classification, probabilistic model, Bayesian approach, kernel functions.*

Введение

С широким развитием сети интернет и её проникновением в большую часть всех сфер жизни, у людей появилась возможность свободно обмениваться информацией и получать доступ к разнообразным ресурсам. Одним из наиболее распространенных способов общения людей через интернет является использование электронной почты. В силу большой открытости этого канала связи с точки зрения возможности передачи любого сообщения произвольному пользователю, он активно используется мошенниками, злоумышленниками и распространителями рекламных материалов. При этом создается не только повышенная нагрузка на техническую инфраструктуру, но и тратится время людей, которым приходится отделять полезную информацию от всей остальной. Поэтому задача автоматизации фильтрации электронной почты будет оставаться актуальной в течение всего времени её существования.

Задача фильтрации спама уже решалась различными методами [3, 4], однако они в большой степени являлись эвристическими и не имели под собой четкой вероятностной модели. Также проблемой является корректное составление обучающей выборки. Дело в том, что спам-письма зачастую шаблонны и имеют много общего в своей структуре, к тому же они широко доступны. Составить же обучающую выборку, содержащую письма, полезные для пользователей, гораздо сложнее по следующим причинам:

— меньшая доступность,
— высокая разнородность,

Научный руководитель О. В. Красоткина

— большое число шаблонных писем (разнообразные уведомления от сервисов).

По этим причинам предлагается использовать методы одноклассовой классификации [1, 2], чтобы отказаться от требования к обучающей выборке содержать достаточно широкое множество разнообразных представителей обоих классов.

В работе предложена квазивероятностная постановка задачи одноклассовой классификации. Такой подход позволяет уточнить область применимости построенной модели и предъявляемые требования к данным. На основе полученной вероятностной постановки задачи, строится новая вероятностная модель порождения объектов, в ходе оптимизации которой происходит построение классификатора.

Полученные методы построения одноклассовых классификаторов применяются к модельным и реальным данным.

Байесовская постановка задачи

Объектом исследования является множество электронных сообщений характеризуемых некоторым набором признаков. Рассмотрим одноклассовую классификацию объектов генеральной совокупности Ω. Пусть каждый объект $\omega \in \Omega$ представлен точкой в линейном пространстве признаков $\mathbf{x}(\omega) = (x^1(\omega), \ldots, x^n(\omega)) \in \mathbb{R}^n$. При этом мы изучаем лишь объекты одного класса, поэтому меткой класса объект существенно не обладает. Тем не менее нашей задачей будет построение классификатора, который будет давать ответ 1, если предъявленный объект лежит в множестве, и 0 иначе.

В работе [1] предлагается строить сферический пороговый классификатор вида $[z \leqslant 0]$, где $z(\mathbf{x}, \mathbf{a}, R) = \|\mathbf{x} - \mathbf{a}\| - R$ без вероятностного обоснования такого подхода. При этом в области $z(\mathbf{x}, \mathbf{a}, R) \geqslant 0$ значение величины $\|\mathbf{x} - \mathbf{a}\|^2 - R^2$ несёт смысл отступа ξ, а для объектов внутри шара отступ полагается равным 0. Для подбора значений $\mathbf{a}, R$ решается задача

$$F(R, \mathbf{a}, \boldsymbol{\xi}) = R^2 + C \sum_i \xi_i \to \min_{\mathbf{a}, R, \boldsymbol{\xi}}, \tag{1}$$

при этом здесь и далее мы полагаем, что суммирование по индексу i (а в дальнейшем и j) означает суммирование по всем объектам обучающей выборки.

Здесь величина C задает баланс между минимальным объёмом шара и наименьшим числом объектов обучающей выборки вне сферы. Пример описания объектов шаром приведен на рисунке 1.

x_1

ковом пространстве имеет вид

$$\varphi\left(\mathbf{x}|\mathbf{a},R;c\right)\propto\begin{cases}1, & z(\mathbf{x},\mathbf{a},R)<0,\\ e^{-c\left(\|\mathbf{x}-\mathbf{a}\|^2-R^2\right)}, & z(\mathbf{x},\mathbf{a},R)\geqslant 0.\end{cases}\tag{2}$$

Здесь величина c является гиперпараметром. График данной функции плотности изображен на рисунке 2.

Совместную п... нимать как плотн...

Р...

$\|\mathbf{x}-\mathbf{a}\|$

— $\mathbf{a}$ и R — случайные независимые величины,
— $|R|$ — нормально распределенная случайная величина с нулевым математическим ожиданием и дисперсией σ^2,
— $\mathbf{a}$ равномерно распределено по всему пространству $\mathbb{R}^n$ (такое распределение будет несобственным [7]).

Тогда совместное распределение параметров также будет несобственным

$$\Psi(\mathbf{a}, R) \propto e^{-\frac{1}{2\sigma^2}R^2}.$$

Подставим это выражение и функцию распределения из (2)

$$\begin{aligned}
\ln p(\mathbf{a}, R|\mathbf{X}) &= \ln \Psi(\mathbf{a}, R) + \sum_{j=1}^{N} \ln \varphi(\mathbf{x}_j|\mathbf{a}, R) = \\
&= -\frac{R^2}{2\sigma^2} + \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|\leqslant R} \ln 1 + \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} \ln e^{-c\left(\|\mathbf{x}-\mathbf{a}\|^2-R^2\right)} = \\
&= -\frac{R^2}{2\sigma^2} - \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} c\left(\|\mathbf{x}_i-\mathbf{a}\|^2 - R^2\right) = \\
&= -\frac{1}{2\sigma^2}\left(R^2 + 2\sigma^2 c \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} \left(\|\mathbf{x}_i-\mathbf{a}\|^2 - R^2\right)\right) \to \max_{\mathbf{a},R}. \qquad (5)
\end{aligned}$$

Очевидно, задачи (5) и (1) эквивалентны при $C = 2\sigma^2 c$.

Решение оптимизационной задачи

Итак, для нахождения значений $\mathbf{a}$ и R необходимо решить следующую задачу

$$\begin{cases} R^2 + C\sum\limits_i \xi_i \to \min\limits_{\mathbf{a},R,\boldsymbol{\xi}}, \\ \|\mathbf{x}_i - \mathbf{a}\|^2 \leqslant R^2 + \xi_i, \quad \xi_i \geqslant 0, \quad i = 1, \ldots, N. \end{cases} \qquad (6)$$

Функция Лагранжа этой задачи имеет вид

$$\mathcal{L}(\mathbf{a}, R, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = R^2 + C\sum_i \xi_i - \sum_i \gamma_i\xi_i - \sum_i \alpha_i\left(R^2 + \xi_i - (\mathbf{x}_i^\intercal\mathbf{x}_i - 2\mathbf{a}^\intercal\mathbf{x}_i + \mathbf{a}^\intercal\mathbf{a})\right),$$

где $\alpha_i \geqslant 0$ и $\gamma_i \geqslant 0$ — множители Лагранжа. Необходимым условием минимума является равенство нулю частных производных функции Лагранжа по всем переменным

$$\begin{aligned}
&\tfrac{\partial\mathcal{L}}{\partial R} = 0: \quad \sum_i \alpha_i = 1 \text{ (случай } R = 0 \text{ рассмотрим отдельно,)} \\
&\tfrac{\partial\mathcal{L}}{\partial \mathbf{a}} = 0: \quad \mathbf{a} = \frac{\sum\limits_i \alpha_i\mathbf{x}_i}{\sum\limits_i \alpha_i} = \sum_i \alpha_i\mathbf{x}_i, \qquad (7) \\
&\tfrac{\partial\mathcal{L}}{\partial \xi_i} = 0: \quad \gamma_i = C - \alpha_i, \quad i = 1, \ldots, N.
\end{aligned}$$

Из последнего уравнения получаем, что $\alpha_i = C - \gamma_i$. Таким образом, мы получаем новые ограничения на α_i

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, N.$$

Если это ограничение выполнено, то мы можем вычислить γ_i по формуле $\gamma_i = C - \alpha_i$, и при этом автоматически будет выполнено условие $\gamma_i \geqslant 0$.

Тогда для функции Лагранжа получим выражение

$$\begin{aligned}\mathcal{L}(\mathbf{a}, R, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= R^2 - \sum_i \alpha_i R^2 + C\sum_i \xi_i - \sum_i \alpha_i \xi_i + \\ &+ \sum_i \alpha_i \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_i - 2\sum_i \alpha_i \mathbf{a}^{\mathsf{T}}\mathbf{x}_i + \sum_i \alpha_i \mathbf{a}^{\mathsf{T}}\mathbf{a} - \sum_i \gamma_i\xi_i = \\ &= R^{\mathsf{T}}R^{\mathsf{T}}\left(1 - \sum_i \alpha_i\right) + \sum_i \xi_i\,(C - \alpha_i - \gamma_i) + \\ &+ \sum_i \alpha_i \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_i - 2\sum_i \alpha_i \sum_j \alpha_j \mathbf{x}_j^{\mathsf{T}}\mathbf{x}_i + \sum_{i,j} \alpha_i\alpha_j \mathbf{x}_j^{\mathsf{T}}\mathbf{x}_i = \\ &= \sum_i \alpha_i \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_i - \sum_{i,j} \alpha_i\alpha_j \mathbf{x}_j^{\mathsf{T}}\mathbf{x}_i \to \max_{\boldsymbol{\alpha}}.\end{aligned}$$

Полученное выражение является квадратичной формой. Тогда его максимум находится по известным алгоритмам решения задач квадратичного программирования. По оптимальным значениям $\boldsymbol{\alpha}$ мы сможем найти оптимальное значение центра гипершара $\mathbf{a}$ и отступов $\boldsymbol{\xi}$, используя соотношения (7).

Для каждого объекта $\mathbf{x}_i$ оптимальное значение α_i (или же $\gamma_i = C - \alpha_i$) задает тип принадлежности объекта построенному гипершару:

— $\alpha_i = 0 \Rightarrow$ объект $\mathbf{x}_i$ лежит внутри гипершара, имеет нулевой отступ;
— $0 < \alpha_i < C \Rightarrow$ объект $\mathbf{x}_i$ лежит на границе гипершара, имеет нулевой отступ;
— $\alpha_i = C \Rightarrow$ объект $\mathbf{x}_i$ лежит вне гипершара, имеет ненулевой отступ.

Радиус R определяется как расстояние от центра гипершара $\mathbf{a}$ до опорных векторов, лежащих на границе гипершара.

Если же $R = 0$, то задача (6) имеет вид

$$\begin{cases} C\sum\limits_i \xi_i \to \min\limits_{\mathbf{a},\boldsymbol{\xi}}, \\ \|\mathbf{x}_i - \mathbf{a}\|^2 \leqslant \xi_i, \quad \xi_i \geqslant 0, \quad i = 1, \ldots, N. \end{cases} \tag{8}$$

т.е.

$$C\sum_i \|\mathbf{x}_i - \mathbf{a}\|^2 \to \min_{\mathbf{a}}, \tag{9}$$

а эта задача соответствует методу наименьших квадратов. Тогда $\mathbf{a} = \frac{\sum_i \mathbf{x}_i}{N}$. При этом следует понимать, что значение $R = 0$ обнуляет обобщающую способность нашего классификатора, поэтому следует отказываться от такого решения, если есть выбор. Здесь же стоит отметить, что $R = 0$ обязательно, если $C < \frac{1}{N}$, где N — число объектов в обучающей выборке, поскольку в этом случае условия на $\boldsymbol{\alpha}$ несовместны.

Для возможности описания данных более гибкой формой, нежели сфера, в работе [1] предлагается использовать потенциальные функции [8]. Наиболее часто используемыми потенциальными функциями являются полиномиальная

$$K_p(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j)^p$$

и радиальная базисная функция Гаусса

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2s^2}\right).$$

Таким образом, чтобы получить улучшенную модель описания данных, необходимо заменить в функции Лагранжа операцию вычисления скалярного произведения двух векторов вычислением значения потенциальной функции двух аргументов.

Численный эксперимент

Для оценки качества работы алгоритма предлагается ввести метрику. Следуя работе [5], будем измерять качество одноклассовой класификации в терминах точности и полноты. В нашем случае точность (precision) — доля верно классифицированных объектов тестовой выборки среди всех объектов, отнесенных алгоритмом к единственному классу. Полнота (recall) — доля верно классифицированных объектов тестовой выборки среди всех объектов, принадлежащих к единственному классу. Более высокие значения точности и полноты соответствуют лучшему качеству классификации. В качестве агрегированного показателя, объединяющего точность P и полноту R используем F_1-меру [6]:

$$F_1 = \frac{2PR}{P+R}.$$

Для проведения вычислительного эксперимента сгенерируем $N = 400$ случайных точек $\{\mathbf{x}_i\}_{i=1}^N$ из распределения (2) при размерности пространства 2 (для наглядности), положив направления смещений случайными и придав параметрам значения $a = (1,2)^\intercal, R = 3, c = 0{,}2$. После этого проведем $t\times q$-fold кросс-валидацию с $t = 10, q = 3$, скользящим контролем подбирая параметр C и вычисляя F_1-метрику при каждом его значении. При этом всё, что лежит вне сферы мы считаем не принадлежащим классу, а всё, что внутри, — считаем. В результате получим следующую зависимость значения метрики от C (см. рисунок 4). Из графика видно, что при $C \to 0$ обобщающая способность также стремится к нулю, поскольку практически отсутствует штраф за непопадание в класс при обучении. При этом большие штрафы заставляют необоснованно увеличивать сферу, снижая точность.

Пример работы алгоритма приведен на рисунке 3 при параметре $C = 0{,}007$. Зеленым изображена граница истинного распределения, красным — построенного. Видно, что здесь C слишком мало и сфера получилось слишком маленькой.

Для проведения эксперимента на реальных данных были выбраны доступные в открытом доступе уже вычисленные признаки сообщений[1]. Здесь для обучения бралась небольшая часть спам-документов (200 из 1800). Сперва они линейно отображались в куб $[0,1]^k$ ($k = 57$ — размерность пространства), а затем по ним строилась сфера в этом 57-мерном пространстве. Для контроля все остальные данные преобразовывались по тому же правилу (что не гарантирует их попадание в этот же куб), после чего проверялось попадание в построенную сферу и вычислялась F_1-метрика. Здесь в контроле уже участвуют объекты как объекты из исследуемого класса (спам-сообщений), так и не из него, хотя обучение происходило только на объектах целевого класса. Результаты подбора параметра C изображены на рисунке (4). Данные усреднены по 20 случайным выборкам по 200 объектов из 1800.

[1]UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Spambase

Литература

[1] Tax D. *One-class classification; Concept-learning in the absence of counterexamples* // Ph.D thesis. 2001.

[2] Khan S., Madden G. *A Survey of Recent Trends in One Class Classiffcation* // College of Engineering and Informatics, National University of Ireland Galway. Ireland. 2006.

[3] Islam R., Chowdhury R. *Spam filtering using ML algorithms* // Universitetets Okonomiske Institute. IADIS International Conference on WWW/Internet. 2007.

[4] *Research of Spam Filtering system based on LSA and SHA* / Sun J. [et al.] // Advances in neural networks. ISNN. 2008.

[5] Романенко А. А. *Категоризация текстов на основе монотонного классификатора ближайшего соседа* // Выпускная квалификационная работа бакалавра. 2012.

[6] van Rijsbergen C. J. *Information Retrieval* // Butterworth. 2nd ed. 1979.

[7] Де Гроот М. *Оптимальные статистические решения* // М.: Мир, 1974.

[8] Айзерман М. А., Браверман Э. М., Розоноэр Л. И. *Метод потенциальных функций в теории обучения машин* // М.: Наука, 1970.