Министерство образования и науки Российской Федерации
Московский физико-технический институт
(государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Сандуляну Любовь Николаевна

# Байесовский подход к построению одноклассового классификатора в задачах обнаружения текстовых заимствований и фильтрации нежелательной почты

511656 - Математические и информационные технологии

Выпускная квалификационная работа магистра

**Научный руководитель:**
к.ф.-м.н. Чехович Юрий Викторович

Москва
2014

# Содержание

**Аннотация**

В работе вводится квазивероятностная модель для классической эмпирической постановки задачи одноклассовой классификации и производится сведение классического подхода к новой модели. Для получения улучшенной модели описания данных в работе предлагается использовать потенциальные функции. В случае радиальной базисной функции Гаусса предложена вероятностная интерпретация. В вычислительном эксперименте построенная модель была применена к двум задачам: задаче фильтрации электронных писем на предмет наличия в них спама и задаче выявления текстовых заимствований.

# 1 Введение

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 двум реальным задачам: задаче фильтрации электронных писем на предмет наличия в них спама и задаче выявления текстовых заимствований.

С широким развитием сети интернет и её проникновением в большую часть всех сфер жизни, у людей появилась возможность свободно обмениваться информацией. 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 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 [1, 2], 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,
- большое число шаблонных писем (уведомления от сервисов).

По этим причинам предлагается использовать методы одноклассовой классификации [3, 4], чтобы отказаться от требования к обучающей вы-

борке содержать достаточно широкое множество разнообразных представителей обоих классов. В описан [5] вероятностный подход к одноклассовой классификации.

Также решается задача классификации цитат в текстовых документах на правомерное и неправомерное цитирование. Решение данной задачи необходимо для повышения качества Интернет-сервиса AntiPlagiat.ru — первой в России системы для проверки текстовых документов на наличие заимствований из общедоступных сетевых источников. В работе [7] решалась задача классификации оформленных цитат, в нашей же работе решается задача классификации заимствованных блоков найденных системой Антиплагиат, но не оформленных как цитаты. Проблема заключается в том, что правомерное заимствование не должны считаться плагиатом и снижать общую оценку оригинальности проверяемого документа. Сложность данной задачи вызвана тем, что не существует чётких формальных критериев, позволяющих отличать правомерное цитирование от плагиата, тем не менее, человек в подавляющем большинстве случаев их отличает. Поэтому было решено использовать методы машинного обучения для построения алгоритма автоматической классификации цитат, сформировав обучающую выборку экспертным путём. Для решения данной задачи предлагается использовать одноклассовую классификацию, так как правомерные цитаты имеют некоторый шаблон, в то время как неправомерные цитаты более разнородны.

# 2 Байесовская постановка задачи

Рассмотрим одноклассовую классификацию объектов генеральной совокупности $\Omega$. 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 [3] предлагается строить сферический пороговый классификатор вида $[z \leq 0]$, где $z(\mathbf{x}, \mathbf{a}, R) = \|\mathbf{x} - \mathbf{a}\| - R$ без вероятностного обоснования такого подхода. При этом в области $z(\mathbf{x}, \mathbf{a}, R) \geq 0$ значение величины $\|\mathbf{x} - \mathbf{a}\|^2 - R^2$ несёт смысл отступа $\xi$, а для объектов внутри шара отступ полагается равным 0. Для подбора значений $\mathbf{a}, R$ решается

задача

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

где суммирование производится 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.

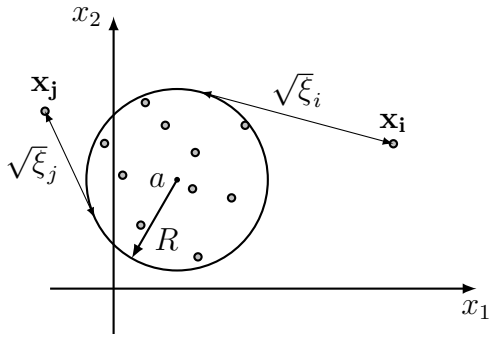


Рис. 1: Пример описания объектов шаром

Будем придерживаться вероятностной модели распределения объектов генеральной совокупности. Параметрическое семейство условных плотностей распределения в признаковом пространстве имеет вид

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

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

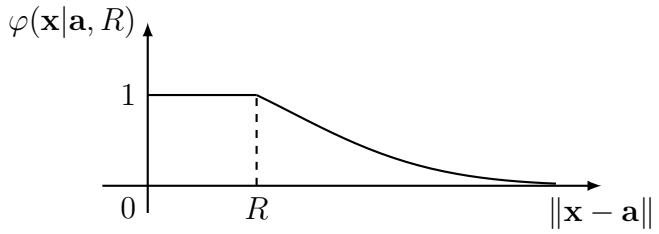


Рис. 2: Значение плотности распределения вдоль радиуса

Совместную плотность распределения случайной обучающей совокупности будем понимать как плотность распределения выборки независимых реализаций

$$\Phi(\mathbf{X}|\mathbf{a},R)=\prod_{j=1}^{N}\varphi(\mathbf{x}_j|\mathbf{a},R),$$

где $\mathbf{X}=\{\mathbf{x}\}_{j=1}^{N}$. Пусть, далее, выбрана априорная плотность совместного распределения вероятностей $\Psi(\mathbf{a},R)$ для параметров распределения $\varphi\left(\mathbf{x}|\mathbf{a},R;c\right)$. Тогда апостериорная плотность распределения параметров $\mathbf{a}$ и $R$ относительно обучающей совокупности определяется формулой Байеса

$$p(\mathbf{a},R|\mathbf{X})=\frac{\Psi(\mathbf{a},R)\Phi(\mathbf{X}\mathbf{a},R)}{\int\Psi(\mathbf{a}',R')\Phi(\mathbf{X}|\mathbf{a}',R')d\mathbf{a}'dR'}. \tag{3}$$

Из принципа максимума плотности апостериорного распределения в пространстве параметров модели генеральной совокупности получим байесовское правило обучения

$$\left(\hat{\mathbf{a}},\hat{R}|\mathbf{X}\right)=\arg\max_{\mathbf{a},R}p(\mathbf{a},R|\mathbf{X}). \tag{4}$$

При этом решающее правило принимает вид

$$f(\mathbf{x})=\left[\|\mathbf{x}-\hat{\mathbf{a}}\|\leq\hat{R}\right]. \tag{5}$$

Поскольку знаменатель в выражении (3) не зависит от целевых переменных

$$p(\mathbf{a},R|\mathbf{X})\propto\Psi(\mathbf{a},R)\Phi(\mathbf{X}|\mathbf{a},R)=\Psi(\mathbf{a},R)\prod_{j=1}^{N}\varphi(\mathbf{x}_j|\mathbf{a},R),$$

то в задаче максимизации (4) достаточно рассматривать только числитель

$$\left(\hat{\mathbf{a}},\hat{R}|\mathbf{X}\right)=\arg\max_{\mathbf{a},R}p(\mathbf{a},R|\mathbf{X})=\arg\max_{\mathbf{a},R}\left(\ln\Psi(\mathbf{a},R)+\sum_{j=1}^{N}\ln\varphi(\mathbf{x}_j|\mathbf{a},R)\right).$$

Теперь покажем, что задача в такой постановке обобщает задачу (1). Положим, что априорное распределение параметров $\Psi(\mathbf{a},R)$ обладает следующими свойствами:

- $\mathbf{a}$ и $R$ — случайные независимые величины,
- $|R|$ — нормально распределенная случайная величина с нулевым математическим ожиданием и дисперсией $\sigma^2$,
- $\mathbf{a}$ равномерно распределено по всему пространству $\mathbb{R}^n$ (такое распределение будет несобственным [8]).

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)

$$\begin{aligned}
\ln p(\mathbf{a}, R|\mathbf{X}) &= \ln \Psi(\mathbf{a}, R) + \sum_{j=1}^{N} \ln \varphi(\mathbf{x}_j|\mathbf{a}, R) = \\
&= -\frac{R^2}{2\sigma^2} + \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|\le R} \ln 1 + \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} \ln e^{-c\left(\|\mathbf{x}-\mathbf{a}\|^2-R^2\right)} = \\
&= -\frac{R^2}{2\sigma^2} - \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} c\left(\|\mathbf{x}_i-\mathbf{a}\|^2 - R^2\right) = \\
&= -\frac{1}{2\sigma^2}\left(R^2 + 2\sigma^2 c \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} \left(\|\mathbf{x}_i-\mathbf{a}\|^2 - R^2\right)\right) \to \max_{\mathbf{a},R}. \qquad (6)
\end{aligned}$$

Очевидно, задачи (6) и (1) эквивалентны при $C = 2\sigma^2 c$.

# 3 Решение оптимизационной задачи

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

$$\begin{cases} R^2 + C\sum\limits_i \xi_i \to \min\limits_{\mathbf{a},R,\xi}, \\ \|\mathbf{x}_i - \mathbf{a}\|^2 \le R^2 + \xi_i, \quad \xi_i \ge 0, \quad i = 1, \ldots, N. \end{cases} \qquad (7)$$

Функция Лагранжа этой задачи имеет вид

$$\mathcal{L}(\mathbf{a}, R, \xi, \alpha, \gamma) = R^2 + C\sum_i \xi_i - \sum_i \gamma_i \xi_i -$$

$$- \sum_i \alpha_i \left( R^2 + \xi_i - \left( \mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{x}_i - 2\mathbf{a}^{\mathrm{T}} \cdot \mathbf{x}_i + \mathbf{a}^{\mathrm{T}} \cdot \mathbf{a} \right) \right),$$

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

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial R} = 0: & \sum\limits_i \alpha_i = 1 \text{ (случай } R = 0 \text{ рассмотрим отдельно,)} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{a}} = 0: & \mathbf{a} = \dfrac{\sum\limits_i \alpha_i \mathbf{x}_i}{\sum\limits_i \alpha_i} = \sum\limits_i \alpha_i \mathbf{x}_i, \\ \frac{\partial \mathcal{L}}{\partial \xi_i} = 0: & \gamma_i = C - \alpha_i, \quad i = 1, \ldots, N. \end{cases} \tag{8}$$

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

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

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

Тогда для функции Лагранжа получим выражение

$$\mathcal{L}(\mathbf{a}, R, \xi, \alpha, \gamma) = R^2 - \sum_i \alpha_i R^2 + C \sum_i \xi_i - \sum_i \alpha_i \xi_i +$$

$$+ \sum_i \alpha_i \mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{x}_i - 2 \sum_i \alpha_i \mathbf{a}^{\mathrm{T}} \cdot \mathbf{x}_i + \sum_i \alpha_i \mathbf{a}^{\mathrm{T}} \cdot \mathbf{a} - \sum_i \gamma_i \xi_i =$$

$$= R^{\mathrm{T}} R^{\mathrm{T}} \left( 1 - \sum_i \alpha_i \right) + \sum_i \xi_i \left( C - \alpha_i - \gamma_i \right) +$$

$$+ \sum_i \alpha_i \mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{x}_i - 2 \sum_i \alpha_i \sum_j \alpha_j \mathbf{x}_j^{\mathrm{T}} \cdot \mathbf{x}_i + \sum_{i,j} \alpha_i \alpha_j \mathbf{x}_j^{\mathrm{T}} \cdot \mathbf{x}_i =$$

$$= \sum_i \alpha_i \mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{x}_i - \sum_{i,j} \alpha_i \alpha_j \mathbf{x}_j^{\mathrm{T}} \cdot \mathbf{x}_i \to \max_{\alpha}.$$

Полученное выражение является квадратичной формой. Тогда его максимум находится по известным алгоритмам решения задач квадратичного программирования. По оптимальным значениям $\alpha$ мы сможем

найти оптимальное значение центра гиперсферы $\mathbf{a}$ и отступов $\xi$ используя соотношения (8).

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

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

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

Если же $R = 0$, то задача (7) имеет вид

$$\begin{cases} C \sum\limits_i \xi_i \to \min\limits_{\mathbf{a},\xi}, \\ \|\mathbf{x}_i - \mathbf{a}\|^2 \le \xi_i, \quad \xi_i \ge 0, \quad i = 1, \ldots, N. \end{cases} \tag{9}$$

т.е.

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

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 $\alpha$ несовместны.

Для возможности описания данных более гибкой формой, нежели сфера, в работе [3] предлагается использовать потенциальные функции [9]. Наиболее часто используемыми потенциальными функциями являются полиномиальная

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

и радиальная базисная функция Гаусса (которую мы и будем в дальнейшем рассматривать)

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

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.

При таком обобщении решающее решающее правило 5 принимает вид

$$f(\mathbf{x}) = [\|\mathbf{x}-\mathbf{a}\| \leq R] = \left[\mathbf{x}\cdot\mathbf{x} - 2\mathbf{x}\cdot\mathbf{a} + \mathbf{a}\cdot\mathbf{a} \leq R^2\right] =$$

$$= \left[\mathbf{x}\cdot\mathbf{x} - 2\mathbf{x}\cdot\left(\sum_i \alpha_i \mathbf{x}_i\right) + \left(\sum_i \alpha_i \mathbf{x}_i\right)\cdot\left(\sum_i \alpha_i \mathbf{x}_i\right) \leq R^2\right] =$$

$$= \left[K(\mathbf{x},\mathbf{x}) - 2\sum_i \alpha_i K(\mathbf{x},\mathbf{x}_i) + \sum_{i,j} \alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j) \leq R^2\right].$$

Здесь оптимальное значение $R$ определяется как значение выражения

$$K(\mathbf{x},\mathbf{x}) - 2\sum_i \alpha_i K(\mathbf{x},\mathbf{x}_i) + \sum_{i,j} \alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j)$$

для граничных объектов ($0 < \alpha_i < C$). После такого преобразования становится неясен вероятностный смысл предложенного подхода. Для того, чтобы исправить это воспользуемся интерпретацией, изложенной в [10].

Сперва сформулируем, что мы будем называть ядром. Пусть имеем некоторое пространство $\mathcal{X}$. Тогда для любого гильбертова пространства $\mathcal{H}$ и любого отображения $\phi : \mathcal{X} \to \mathcal{H}$ будем называть функцию $\mathcal{K} : \mathcal{X}\times\mathcal{X} \to R$ ядром, если $\forall x, y \in \mathcal{X} \qquad \mathcal{K}(x,y) = \langle\phi(x),\phi(y)\rangle$. При этом пространство $\mathcal{H}$ называют спрямляющим для ядра $\mathcal{K}$.

В [11] приводится доказательство того, что РБФ Гаусса действительно является ядром, а также явно приводится вид спрямляющего пространства.

Таким образом, вероятностная интерпретация следующая. Существует некоторая генеральная совокупность объектов $\Omega$, их признаковое описание в пространстве $\mathcal{X}$ и отображение $\phi$ из $\mathcal{X}$ в гильбертово пространство $\mathcal{H}$. В пространстве $\mathcal{H}$ над $\phi(\mathcal{X})$ случайно порождается распределение согласно 2

$$\varphi(\mathbf{h}|\mathbf{a},R;c) \propto \begin{cases} 1, & z(\mathbf{h},\mathbf{a},R) < 0, \\ e^{-c\left(\|\mathbf{h}-\mathbf{a}\|^2 - R^2\right)}, & z(\mathbf{h},\mathbf{a},R) \geq 0. \end{cases} \tag{10}$$

где величины $\mathbf{a}$ и $R$ случайно выбираются из априорного распределения параметров $\Psi(\mathbf{a}, R)$ аналогично случаю без использования ядер. После чего вероятность пронаблюдать объект с признаковым описанием $\mathbf{x} \in \mathcal{X}$ пропорциональна $\varphi(\phi(\mathbf{x})|\mathbf{a}, R; c)$.

# 4 Экспериментальное исследование

## 4.1 Модельные данные

Для оценки качества работы алгоритма предлагается ввести метрику. Будем измерять качество одноклассовой классификации 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 [12]:

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

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 $a = (1,2)^{\mathrm{T}}, R = 3, c = 0{,}2$. 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.

На рисунке 3 изображена полученная зависимость значения $F_1$-метрики от параметра $C$. 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.

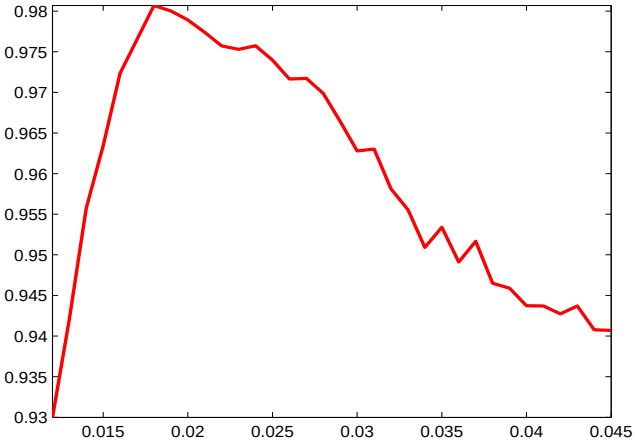


Рис. 3: Зависимость $F_1$-метрики от параметра регуляризации $C$

Пример работы алгоритма приведен на рисунке 4 при параметрах $n = 2, N = 400, R = 3, a = (1; 2)^T, c = 0,6, C = 0,007$. Зеленым изображена граница истинного распределения, красным — построенного. Видно, что здесь $C$ слишком мало и полученная сфера описывает лишь небольшую часть объектов.

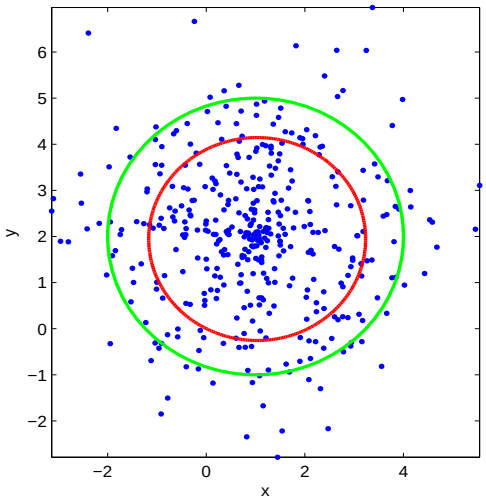


Рис. 4: Пример результата работы алгоритма при $C = 0,007$

Примеры работы алгоритма при использовании ядер приведены на рисунках 5, 6. Зеленые круги — смесь трёх распределений с общим $R = 1,5, c = 0,6$ и различными центрами:

- $a_1 = (1; 6)^T, N_1 = 150,$
- $a_1 = (5; 0)^T, N_1 = 75,$
- $a_1 = (3; 4)^T, N_1 = 50.$

Значение гиперпараметра положено равным $C = 0,015$. Синим отмечена построенная разделяющая поверхность. Существенной проблемой подхода является вопрос связности построенных областей: например, встречаются области, не отнесенные алгоритмом к классу, полностью содержащиеся в областях, которые алгоритм принял за относящиеся.

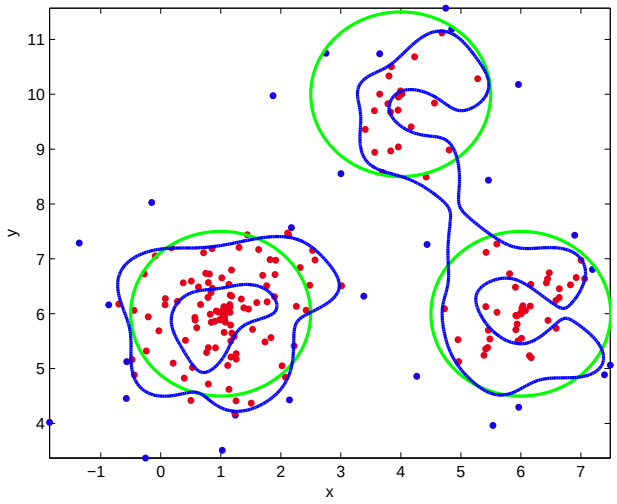


Рис. 5: Пример результата работы алгоритма при $C = 0,015, s = 1$

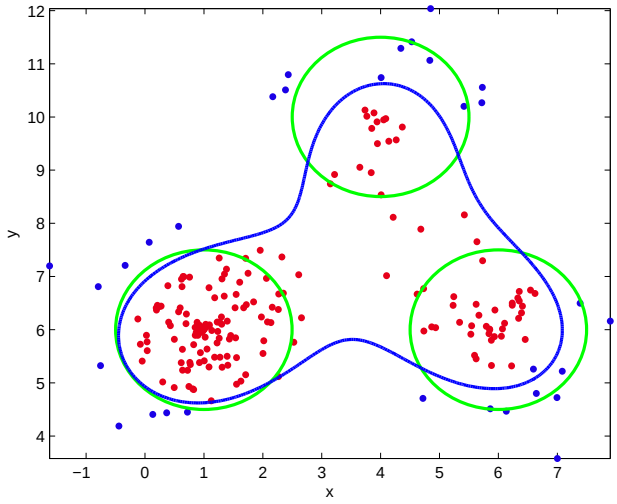


Рис. 6: Пример результата работы алгоритма при $C = 0,015, s = 10$

## 4.2 Задача фильтрации электронных писем на предмет наличия в них спама

В настоящее время одним из наиболее распространенных способов общения людей через интернет является использование электронной почты. Однако, в связи с тем, что этот канал связи открыт с точки зрения возможности передачи любого сообщения произвольному пользователю, он активно используется мошенниками 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 является весьма актуальной. Примеры писем относящихся и не относящихся к спаму приведены на рисунках 7, 8.

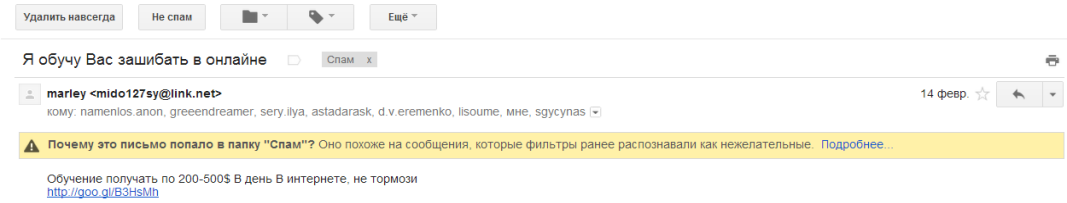


Рис. 7: Пример письма относящегося к спаму

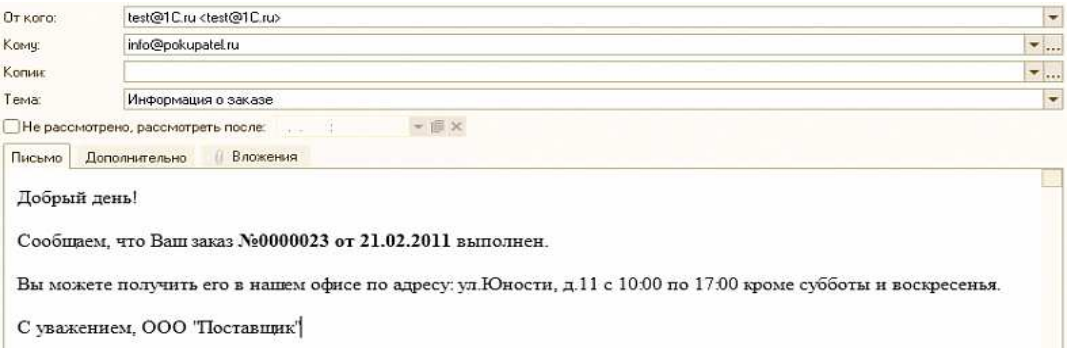


Рис. 8: Пример письма не относящегося к спаму

Эксперимент на реальных данных был проведен с использованием данных, находящихся в открытом доступе (UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Spambase). Эти данные содержат уже вычисленные $n = 57$ признаков сообщений. Использованная база содержит в себе как объекты, относящиеся к спаму, так и не относящиеся.

Задача фильтрации спама неоднократно решалась различными методами, 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, большое число шаблонных писем (уведомления от сервисов). По этим причинам предлагается использовать методы одноклассовой классификации, чтобы отказаться от требования к обучающей выборке содержать достаточно широкое множество разнообразных представителей обоих классов.

В вычислительном эксперименте каждый из $n = 57$ признаков документов линейно отображался в отрезок $[0, 1]$, чтобы учесть различие в их масштабах. Для обучения бралась небольшая часть спам-документов (200 из 1800). Затем по новым точкам строилась сфера (в 57-мерном пространстве). В эксперименте контрольная выборка содержала все доступные объекты (в том числе объекты из обучения). Для каждого из объектов контрольной выборки проверялось попадание в построенную сферу и далее вычислялась $F_1$-метрика. Отдельно стоит отметить, что в контроле участвуют как объекты из исследуемого класса (спам-сообщений), так и не из него, хотя обучение происходило только на объектах целевого класса (спама). Результаты подбора параметра $C$ изображены на рисунке 9. Данные усреднены по 20 случайным выборкам без повторений по 200 объектов из 1800.

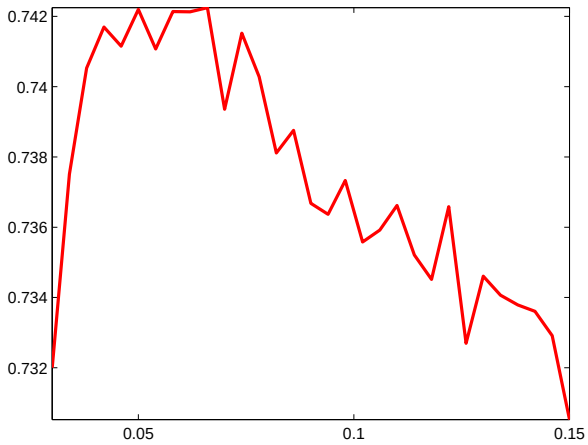


Рис. 9: Зависимость $F_1$-метрики от параметра регуляризации $C$

Отчетливо прослеживается максимум метрики, что свидетельствует о наличии оптимального значения параметра $C$. Результат, полученный при использовании ядра Гаусса, изображен на рисунке 11. Ещё раз отметим, что отличие заключается лишь в записи задачи квадратичной оптимизации: теперь вместо скалярного произведения следует вычислять значения ядерной функции. Использовался параметр $s = 1$, при этом при его вариации значения $F_1$-метрики менялись незначительно.

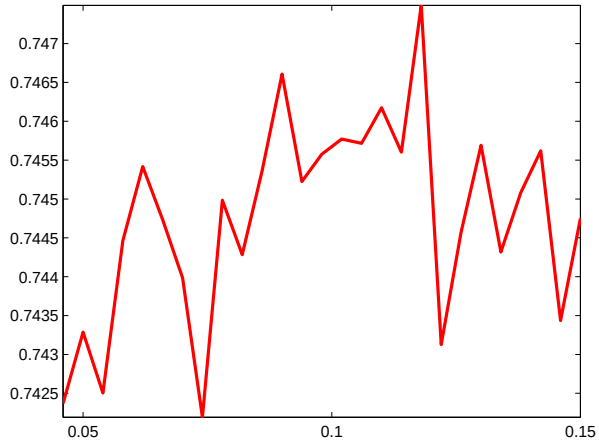


Рис. 10: Зависимость $F_1$-метрики от $C$ при РБФ Гаусса с $s = 1$

Видно, что результаты не столь устойчивы, однако они стабильно выше результатов обычного алгоритма в достаточно широком диапазоне параметра регуляризации.

Для сравнения был применен так же метод опорных векторов (SVM), при этом обучающая выборка так же содержала 200 объектов, а контроль проводился по всем доступным объектам. После усреднения по 20 случайным выборкам без повторений по 200 объектов из 1800, получен результат $F_1 = 62.43$, что значительно ниже результатов полученных предложенным методом основанным на одноклассовой классификации.

## 4.3 Задача выявления текстовых заимствований

Решается задача классификации цитат в текстовых документах, требуется определить является ли данная цитата плагиатом или нет. Решение данной задачи необходимо для повышения качества Интернет-сервиса AntiPlagiat.ru — первой в России системы для проверки текстовых документов на наличие заимствований из общедоступных сетевых источников. Под цитатой здесь понимается не правильно оформленная цитата, а просто блок текста, найденный антиплагиатом. Проблема заключается в том, что правомерные цитаты не должны считаться плагиатом и снижать общую оценку оригинальности проверяемого документа.

Сложность данной задачи вызвана тем, что не существует чётких формальных критериев, позволяющих отличать правомерное цитирование от плагиата, тем не менее, человек в подавляющем большинстве случаев их отличает. Поэтому было решено использовать методы машинного обучения для построения алгоритма автоматической классификации цитат, сформировав обучающую выборку экспертным путём. Для решения данной задачи предлагается использовать одноклассовую классификацию, так как правомерные цитаты имеют некоторый шаблон, в то время как неправомерные цитаты более разнородны.

Вычислительный эксперимент был проведен с использованием текстов росийской государственной библиотеки, размеченных экспертами. Все объекты, то есть цитат-блоки гарантированно есть в каких-то других источниках. Эксперты смотрели на них и решали, что из этого является настоящим плагиатом, а что нет, настоящая цитата не должна быть заимствованием. Помимо этого есть еще ряд цитат-блоков, которые эксперты по тем или иным причинам не отнесли к плагиату. Например, длинное название учреждения "московский государственный университет ордена красного знамени имени и т. д."это не является оформленной цитатой однако считать это плагиатом нельзя или "в своей работе Перельман доказал, что всякое односвязное компактное многообразие без края гомеоморфно сфере"никаких признаков цитаты нет, но плагиатом это тоже не является. Имются 364 текста с выделенными в них цитатами про которые уже известно плагиат это или нет. Нами был сформирован набор признаков цитат для построения классификатора. Ниже приведен список используемых признаков.

1. Начало в процентах от длины текста
2. Число символов в блоке
3. Число слов в блоке
4. Число предложений в блоке
5. Число символов во всем тексте
6. Средняя длина предложений в блоке в символах
7. Средняя длина предложений в блоке в словах

8. Средняя длина слов в блоке

9. Индекс удобочитаемости по блоку (мера определения сложности восприятия текста читателем)

10. отношение средних длин предложений в словах во всем тексте к средней длине предложений в блоке

11. отношение средних длин предложений в символах во всем тексте к средней длине предложений в блоке

12. отношение средних длин слов во всем тексте к средней длине слов в блоке

13. отношение индексов удобочитаемости по всему тексту и по блоку

14. процент заимствований из источника

15. количество блоков из этого источника

Каждый из $n = 15$ признаков линейно отображался в отрезок $[0, 1]$, чтобы учесть различие в их масштабах. Для обучения брались случайные 200 правомерных цитат. Затем по новым точкам строилась сфера. В эксперименте контрольная выборка содержала случайную 1000 из всех имеющихся цитат. Для каждого из объектов контрольной выборки проверялось попадание в построенную сферу и далее вычислялась $F_1$-метрика. Результаты подбора параметра $C$ изображены на рисунке 11. Данные усреднены по 10 случайным выборкам.

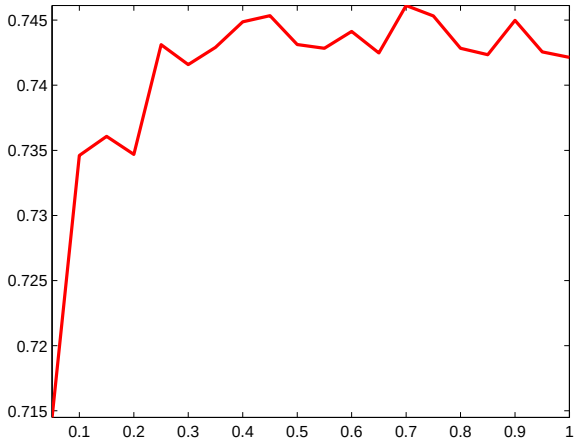


Рис. 11: Зависимость $F_1$-метрики от параметра регуляризации $C$

Для сравнения был применен так же метод опорных векторов (SVM), при этом обучающая выборка так же содержала 200 объектов, а контроль проводился по случайным 1000 объектам. После усреднения по 20 случайным выборкам без повторений по 200 объектов, получен результат $F_1 = 57.24$, что значительно ниже результатов полученных предложенным методом основанным на одноклассовой классификации.

# 5 Заключение

- В работе предложен вероятностный подход к задаче одноклассовой классификации. Такой подход более удобен, чем классический подход, основанный на эвристических соображениях, и с теоретической, и с практической точек зрения, поскольку несет в себе ясную возможность модификаций.

- Доказано, что классический подход является частным случаем предложенного подхода.

- Произведено обобщение алгоритма на случай ядерных функций. В случае радиальной базисной функции Гаусса предложена вероятностная интерпретация.

- Проведены вычислительные эксперименты на модельных и реальных данных. Построенная модель была применена к двум задачам: задаче фильтрации электронных писем на предмет наличия в них спама и задаче выявления текстовых заимствований. Проведена серия численных экспериментов, результаты которых позволяют говорить об актуальности использования предложенного метода при решении данных прикладных задач.

# Список литературы


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

[2] Sun J. *Research of Spam Filtering system based on LSA and SHA*, Advances in neural networks. ISNN, 2008.

[3] Tax D. *One-class classification; Concept-learning in the absence of counterexamples*, Ph.D thesis, 2001, 190 p.

[4] Khan S., Madden G. *A Survey of Recent Trends in One Class Classiffcation*, National University of Ireland Galway. Ireland, 2006.

[5] Бурмистров М. О., Сандуляну Л. Н. *Байесовский подход к построению одноклассового классификатора в задаче фильтрации нежелательной почты*, Известия Тульского государственного университета, 2013.

[6] Антиплагиат: http://www.antiplagiat.ru

[7] Куренной А. С. *Распознавание цитат в текстовых фрагментах* Выпускная квалификационная работа бакалавра, 2009.

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

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

[10] Мерков А.Б. *О статистическом обучении*, Лаборатория распознавания образов, М.: МЦНМО, 2013.

[11] An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels // IEEE Transactions on Information Theory, 2006.

[12] C. J. van Rijsbergen, Information Retrieval (2nd ed.), Butterworth, 1979.