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

Л. Н. Сандуляну

Научный руководитель: Ю. В. Чехович
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

2014

Цель работы

Предмет исследования: генеральная совокупность объектов Ω характеризуемых некоторым набором признаков.
Объект $\omega \in \Omega$ представлен точкой в линейном пространстве признаков $\mathbf{x}(\omega) = \left(x^1(\omega), \ldots, x^n(\omega)\right) \in \mathbb{R}^n$.

Цель: вероятностная постановка задачи одноклассовой классификации.

Сферический пороговый классификатор

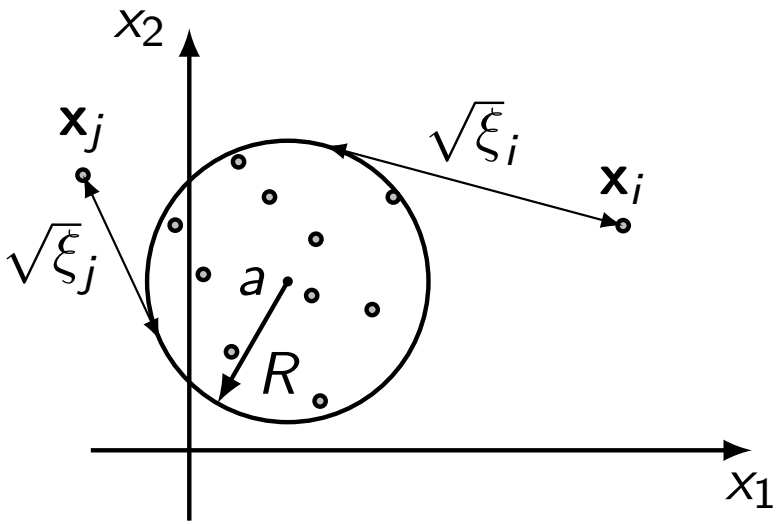
Сферический пороговый классификатор

$$z \leq 0, \text{ где } z(\mathbf{x}, \mathbf{a}, R) = \|\mathbf{x} - \mathbf{a}\| - R.$$

Вне шара значение величины $\|\mathbf{x} - \mathbf{a}\|^2 - R^2$ несёт смысл отступа ξ. Отступ объектов внутри шара равен 0.
Для подбора значений $\mathbf{a}, R$ решается задача

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

Пример описания объектов шаром:

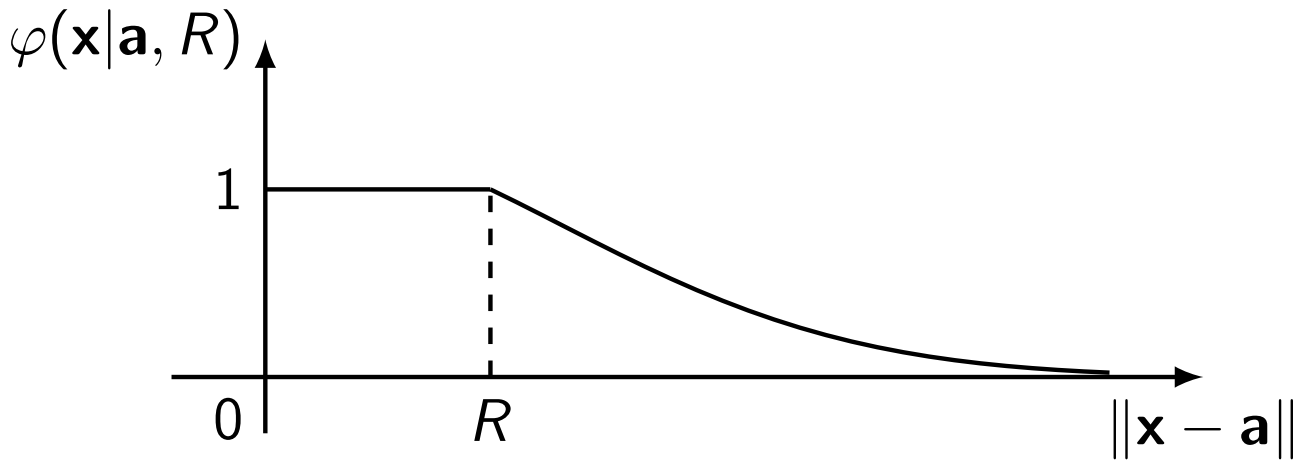


Параметрическое семейство условных плотностей распределения

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

где c — гиперпараметр.

Значение плотности распределения вдоль радиуса:



Байесовское правило обучения

Пусть объекты выборки независимы, тогда

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

$\Psi(\mathbf{a},R)$ — априорная плотность совместного распределения, тогда для апостериорной плотности распределения получим

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

Из принципа максимума плотности апостериорного распределения $\left(\hat{\mathbf{a}},\hat{R}|\mathbf{X}\right)=\arg\max\limits_{\mathbf{a},R}p(\mathbf{a},R|\mathbf{X})$

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

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

Предположения о виде распределения

Пусть $\Psi(\mathbf{a}, R)$ обладает следующими свойствами:

- $\mathbf{a}$ и R — случайные независимые величины,
- R — нормально распределен с нулевым матожиданием и дисперсией σ^2,
- $\mathbf{a}$ равномерно распределено по всему пространству $\mathbb{R}^n$.

Тогда совместное распределение параметров имеет вид

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

Переход к классической постановке

Классическая постановка — частный случай вероятностной постановки

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

$$= -\frac{R^2}{2\sigma^2} + \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} \ln e^{-c\left(\|\mathbf{x}-\mathbf{a}\|^2-R^2\right)} =$$

$$= -\frac{R^2}{2\sigma^2} - \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R} c\left(\|\mathbf{x}_i-\mathbf{a}\|^2-R^2\right) =$$

$$= -\frac{1}{2\sigma^2}\left(R^2 + 2\sigma^2 c \sum_{i:\|\mathbf{x}_i-\mathbf{a}\|>R}\left(\|\mathbf{x}_i-\mathbf{a}\|^2-R^2\right)\right) \to \max_{\mathbf{a},R}.$$

При $C = 2\sigma^2 c$ получаем классическую постановку.

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

Функция Лагранжа этой задачи

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

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

где $\alpha_i \geq 0$ и $\gamma_i \geq 0$ — множители Лагранжа.
Приравняв все частные производные нулю и подставив в функцию Лагранжа, получим

$$\mathcal{L}(\mathbf{a}, R, \boldsymbol{\xi}, \alpha, \gamma) = \sum_i \alpha_i \mathbf{x}_i^\mathsf{T} \cdot \mathbf{x}_i - \sum_{i,j} \alpha_i \alpha_j \mathbf{x}_j^\mathsf{T} \cdot \mathbf{x}_i \to \max_\alpha .$$

Полученное выражение является квадратичной формой.

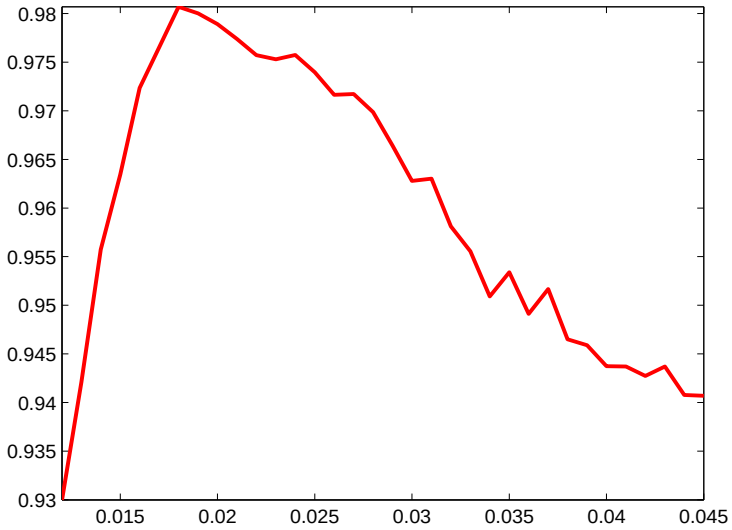
- $N = 400$ случайных точек $\{\mathbf{x}_i\}_{i=1}^{N}$ из распределения $\varphi(\mathbf{x}|\mathbf{a}, R, c)$ в пространстве $\mathbb{R}^2$,
- направления смещений случайные,
- $\mathbf{a} = (1, 2)^T$, $R = 3$, $c = 0{,}2$.

Скользящим контролем подбираем параметр C и вычисляем F_1-метрику при каждом его значении

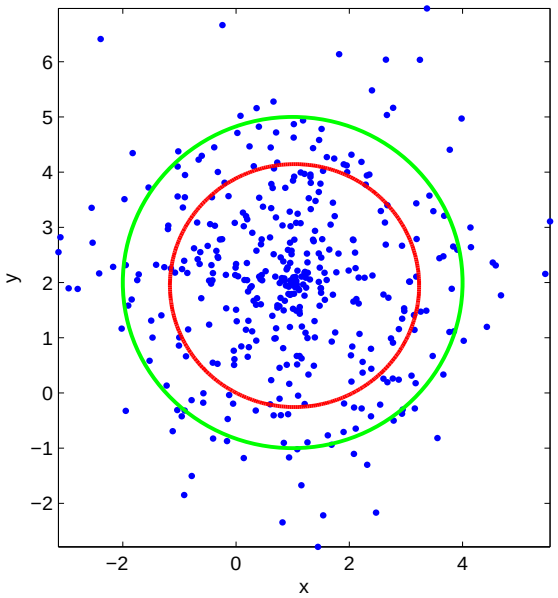
$$F_1 = \frac{2PR}{P + R}.$$

P — доля верно классифицированных объектов тестовой выборки среди объектов, отнесенных алгоритмом к классу.
R — доля верно классифицированных объектов тестовой выборки среди всех объектов, принадлежащих к классу.

Зависимость F_1-метрики от параметра регуляризации C



Пример результата работы алгоритма при $C = 0,007$

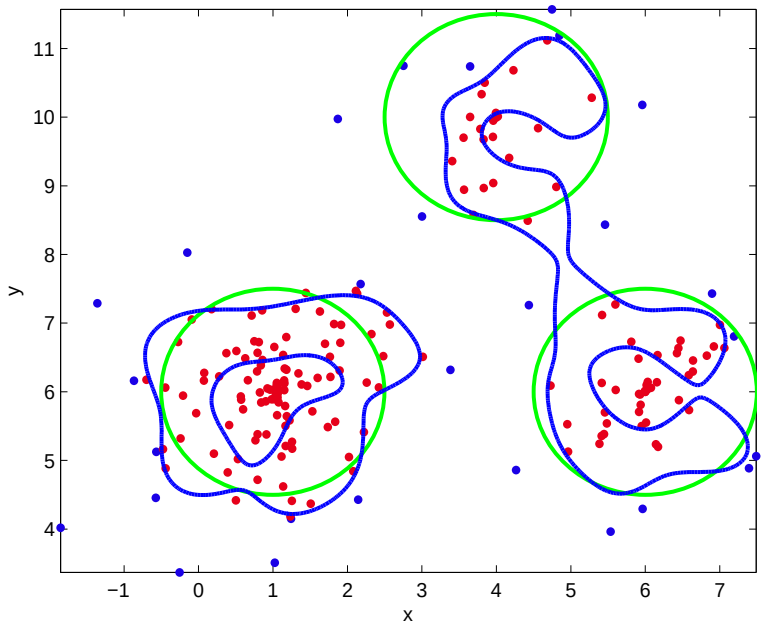


Зеленым изображена граница истинного распределения, красным — построенного.

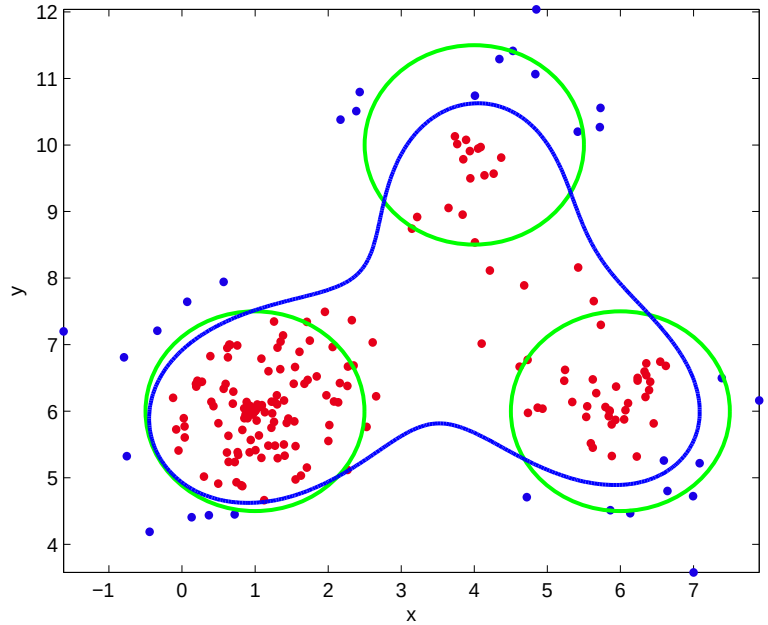
Радиальная базисная функция Гаусса

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

Пример результата работы алгоритма:



$C = 0,015, s = 1$



$C = 0,015, s = 10$

Задача фильтрации электронных писем на предмет наличия в них спама

Одноклассовая классификация предпочтительна, так как:

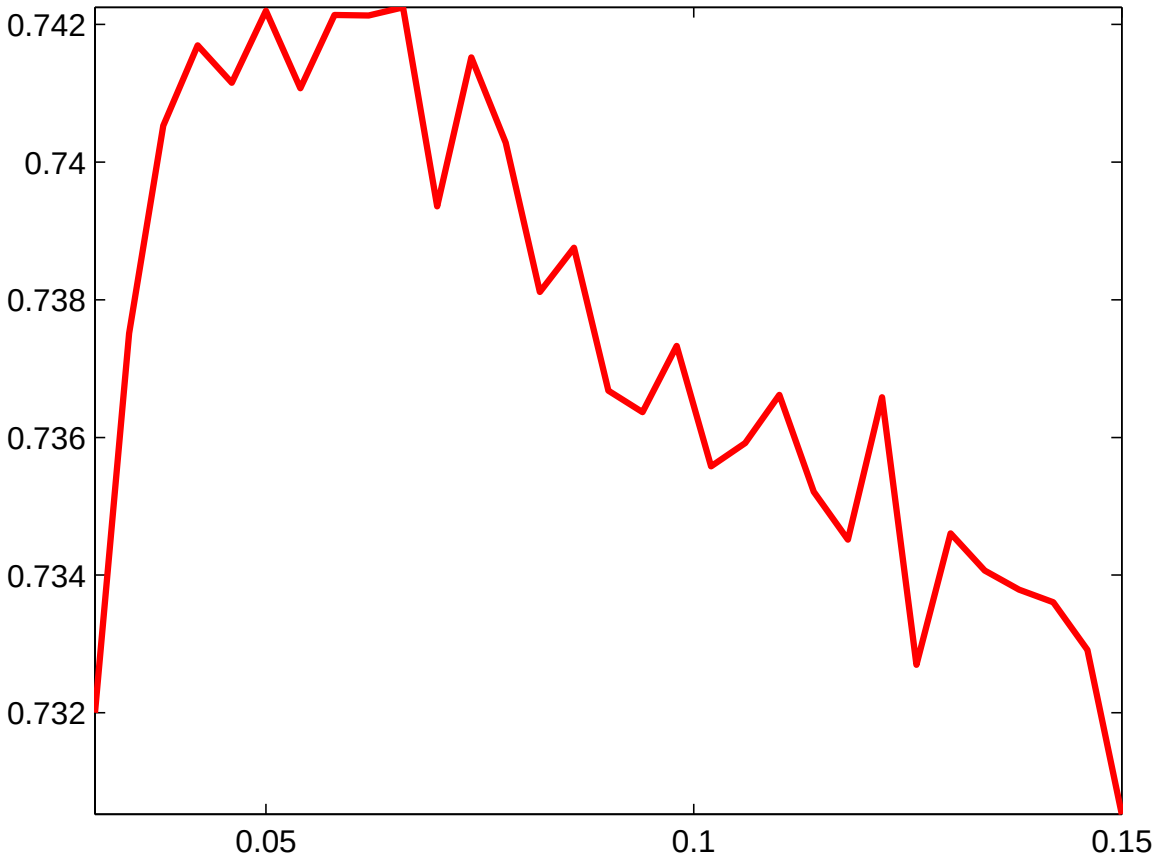
- полезные письма менее доступны и сильно разнородны,
- спам-письма доступны, шаблонны и имеют много общего в своей структуре

Задача фильтрации электронных писем на предмет наличия в них спама

- Данные содержат уже вычисленные $n = 57$ признаков сообщений.
- Каждый из $n = 57$ признаков линейно отображался в отрезок $[0, 1]$.
- Для обучения бралась небольшая часть спам-документов (200 из 1800).
- Контрольная выборка содержала все доступные объекты.
- Данные усреднены по 20 случайным выборкам без повторений по 200 объектов из 1800.

Задача фильтрации электронных писем на предмет наличия в них спама

Зависимость F_1-метрики от параметра регуляризации C



Метод опорных векторов (SVM): $F_1 = 62.43$.

Задача выявления текстовых заимствований

Решается задача классификации цитат в текстовых документах, требуется определить является ли данная цитата плагиатом или нет.

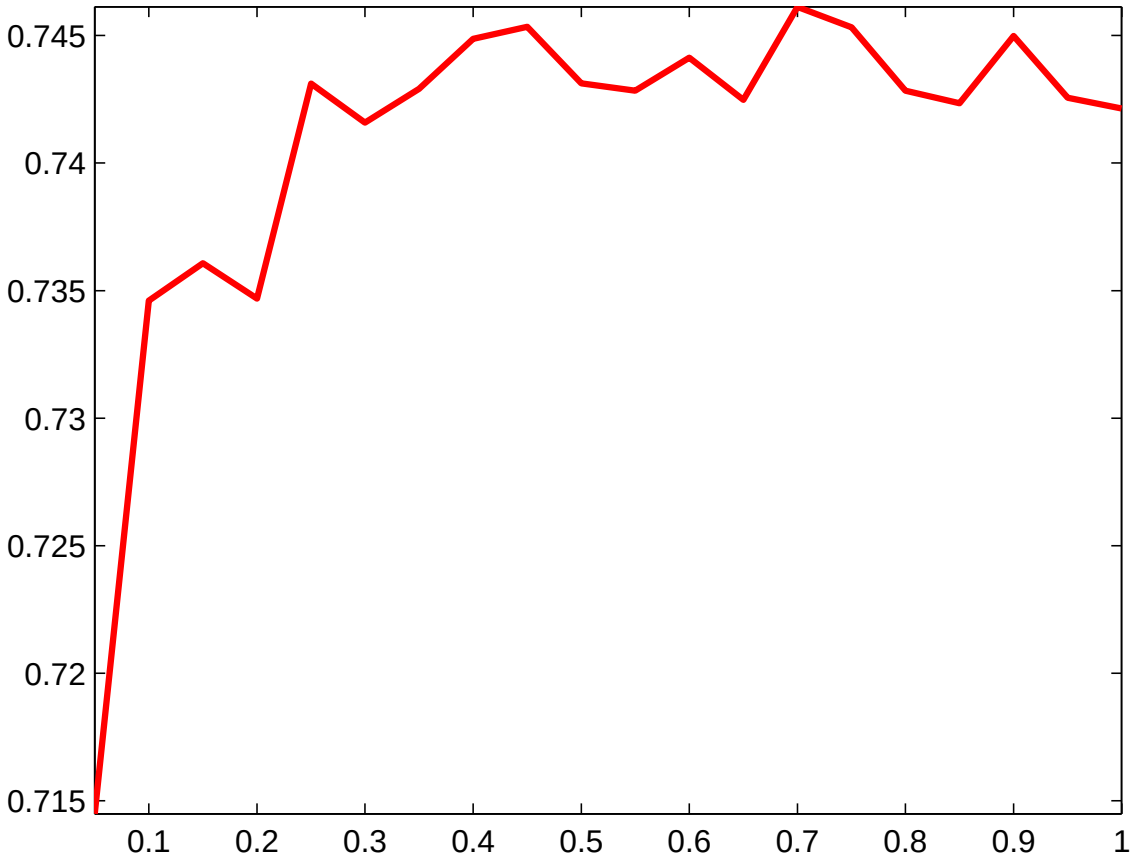
Цитата — блок текста, найденный антиплагиатом.

Правомерные цитаты не должны считаться плагиатом и снижать общую оценку оригинальности проверяемого документа.

Задача выявления текстовых заимствований

- начало в процентах от длины текста
- число символов в блоке
- число слов в блоке
- число предложений в блоке
- число символов во всем тексте
- средняя длина предложений в блоке в символах
- средняя длина предложений в блоке в словах
- средняя длина слов в блоке
- индекс удобочитаемости по блоку
- отношение средних длин предложений в словах во всем тексте к средней длине предложений в блоке
- отношение средних длин предложений в символах во всем тексте к средней длине предложений в блоке
- отношение средних длин слов во всем тексте к средней длине слов в блоке
- отношение индексов удобочитаемости по всему тексту и по блоку
- процент заимствований из источника
- количество блоков из этого источника

Зависимость F_1-метрики от параметра регуляризации C



Метод опорных векторов (SVM): $F_1 = 57.24$.

Результаты:

- Предложен вероятностный подход к задаче одноклассовой классификации.
- Доказано, что классический подход является частным случаем предложенного.
- Произведено обобщение алгоритма на случай ядерных функций.
- 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.

Публикации

- **М.О.Бурмистров, Л.Н.Сандуляну *Байесовский подход к построению одноклассового классификатора в задаче фильтрации нежелательной почты*, Известия Тульского государственного университета (принята в печать).**
- ***Вероятностная модель одноклассовой классификации*, Ломоносов-2013.**
- **М.О.Бурмистров, Л.Н.Сандуляну *Вероятностная модель одноклассовой классификации*, Машинное обучение и анализ данных. — 2012. — № 4.**
- Л.Н.Сандуляну, В.В.Стрижов *Выбор признаков в авторегрессионных задачах прогнозирования*, Информационные технологии. — 2012. — № 6.
- Л.Н.Леонтьева *Последовательный выбор признаков при восстановлении регрессии*, Машинное обучение и анализ данных. — 2012. — № 3. — С. 63-74.
- Л.Н.Леонтьева *Выбор моделей прогнозирования цен на электроэнергии*, Машинное обучение и анализ данных. — 2011. — № 2. — С. 129-139.
- Л.Н.Леонтьева *Многомерная гусеница, выбор длины и числа компонент*, Машинное обучение и анализ данных. — 2011. — № 1. — С. 2-10.