Классификация нестационарного потока текстовых объявлений

Гринчук Олег

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. ВЦ РАН К. В. Воронцов

Москва,
2014 г.

Цель работы

Задача

Дана обучающая коллекция текстовых объявлений D, разбитая на 2 класса. Требуется классифицировать тестовую коллекцию.

Особенности задачи

- неоднородность текстов;
- соотношение классов 1:50;
- нестационарность;

Критерий качества

Площадь под ROC кривой (AUC)

Поля объявления

Текст

- заголовок объявления - *title*
- содержание объявления - *description*

Время

- Точное время создания объявления - t

Дополнительная информация

- Данные об авторе (имя, телефон, ...)
- Заявленная автором тема объявления
- Подключение платных услуг - *isVIP*

Блокировка

- Несоответствие темы, заявленной автором, настоящей теме - $y = \{-1, 1\}$

Переход к числовым признакам

Обработка текста

1. Объдинение полей *title* и *description*
2. Лемматизация
3. Удаление неправильных и бессмысленных слов
4. Замена других полей словами-триггерами ('isVIP','isCompany',...)

Переход к числовым признакам

Все слова коллекции нумеруются. Текстовое поле объявления d_j представляется в виде вектора $\mathbf{x}_j = (x_j^1, x_j^2, \ldots, x_j^n)$, где x_j^i - количество вхождений i-го слова в j-е объявление.

Новый вид объявления

$$d_j = (\mathbf{x}_j, y_j, t_j)$$

Стационарная задача

Дано: обучающая выборка $\{\mathbf{x}_i, y_i\}_{i=1}^{N} = (\mathbf{X}, y)$,
$\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$.
Требуется: построить алгоритм классификации объектов
$a(\mathbf{x}_i, \mathbf{w}) = \text{sign } f(\mathbf{x}_i, w)$, где $\mathbf{w}$ - вектор параметров.
Критерий качества: AUC для тестовой выборки $\{\mathbf{x}_i, y_i\}_{i=1}^{M}$.
Решение: Метод опорных векторов

$$\begin{cases} -\sum\limits_{i=1}^{N} \alpha_i + \frac{1}{2} \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \to \min\limits_{\alpha}, \\ 0 \leq \alpha_i \leq C, \; i = 1, ..., N, \\ \sum\limits_{i=1}^{N} \alpha_i y_i = 0, \end{cases} \tag{1}$$

$$f(\mathbf{x}_i, \mathbf{w}) = \sum_{j=1}^{N} \alpha_j y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + b$$

Нестационарность

Нестационарная выборка

Объекты выборки изменяются со временем: $d_j{=}(\mathbf{x}_j(t_j){,}y_j)$;

Значимость i-го признака

Статистика $S_i(\mathbf{X}^i)$.

- $S_i = \frac{1}{L}\sum\limits_{j=1}^{L} x_j^i$ - эмпирическое матожидание для выборки из некоторого распределения
- $S_i = \frac{n_w}{n}$ - оценка вероятности встретить слово w в коллекции документов

Условие стационарности

Для любых двух больших подвыборок выборки значимости признаков одинаковы.
$|S_i(\mathbf{X}_1)^i) - S_i(\mathbf{X}_2^i)| < \varepsilon_i\ \forall i \in [1, ..., n]$, где ε_i - уровень погрешности

Условия

- Обучающая (X_N, t) и тестовая (X_M, t) выборки.
- Тестовая выборка расположена дальше обучающей на временной оси.
- Объемы выборок довольно велики.

Первичный тест

1. Выбрать $S_i, \varepsilon_i \ \forall i \in [1, .., n]$
2. Инициализировать веса $\{w_i\}$ $(w_i = \frac{1}{n})$
3. Посчитать $U = \sum\limits_{i=1}^{n} w_i[|S_i(\mathbf{X}_N)^i) - S_i(\mathbf{X}_M^i)| < \varepsilon_i]$

Если $U < 1 =>$ выборка нестационарна.

Условия

- Известны ответы Y_M

Вторичный тест

1. Случайно разбить выборку X_{N+M} на подвыборки из N и M элементов
2. Обучить классификатор (SVM) на N-выборке
3. Получить критерий качества (AUC) на M-выборке
4. Повторить шаги (1)-(3) для большого количества итераций
5. Построить статистику критерия качества

Если критерий качества для исходных тестовой и обучающей выборок лежит в критической области статистики => выборка нестационарна.

Если задача нестационарна, классификация большой тестовой выборки даст плохие результаты. Предлагается классифицировать объекты последовательных маленьких выборок, обучаясь на предыдущих

Нестационарный алгоритм

X_N - обучающая выборка X_{M1}, X_{M2} - тестовые выборки, поступающие потоково

1. Обучение классификатора на X_N
2. Классификация X_{M1}
3. Дообучение классификатора на X_{M1}
4. Классификация X_{M2}
5. ...

SVM для нестационарной задачи классификации

Рассмотрим задачу дообучения на новых M объектах SVM классификатора, обученного на N объектах ($M \geq 1$).

$$\min_{0 \leq \alpha_i \leq C} W = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i Q_{ij} \alpha_j - \sum_{i=1}^{N} \alpha_i + b \sum_{i=1}^{N} y_i \alpha_i, \tag{2}$$

где $Q_{ij} = y_i y_j \langle x_i, x_j \rangle$.

Условия KKT

$$g_i = \frac{\partial W}{\partial \alpha_i} = \sum_{j=1}^{N} Q_{ij} \alpha_j + y_i b - 1 \begin{cases} > 0, & \alpha_i = 0 \\ = 0, & 0 < \alpha_i < C \\ < 0, & \alpha_i = C. \end{cases} \tag{3}$$

$$h = \frac{\partial W}{\partial b} = \sum_{j=1}^{N} y_j \alpha_j = 0 \tag{4}$$

SVM для нестационарной задачи классификации

Категории объектов: $\mathcal{S}$ ($g_i = 0$); $\mathcal{E}$ ($g_i < 0$); $\mathcal{R}$ ($g_i > 0$); $\mathcal{U}$.

Условия ККТ должны сохраняться

Изменение частных производных

$$g_i = \sum_j Q_{ij}\alpha_j + \sum_{k\in\mathcal{S}} Q_{ik}\Delta\alpha_k + \sum_{l\in\mathcal{U}} Q_{il}\Delta\alpha_l + y_i(b+\Delta b) - 1 = 0 \quad \forall\ i \in \mathcal{S} \tag{5}$$

$$h = \sum_j y_j\alpha_j + \sum_{k\in\mathcal{S}} y_k\Delta\alpha_k + \sum_{l\in\mathcal{U}} y_l\Delta\alpha_l = 0 \tag{6}$$

В дифференциальной форме

$$\Delta g_i = \sum_{k\in\mathcal{S}} Q_{ik}\Delta\alpha_k + \sum_{l\in\mathcal{U}} Q_{il}\Delta\alpha_l + y_i\Delta b = 0 \quad \forall\ i \in \mathcal{S} \tag{7}$$

$$\Delta h = \sum_{k\in\mathcal{S}} y_k\Delta\alpha_k + \sum_{l\in\mathcal{U}} y_l\Delta\alpha_l = 0 \tag{8}$$

SVM для нестационарной задачи классификации

Incremental SVM обновляет коэффициенты в виде серии последовательных возмущений

Вводится параметр возмущения $p \in [0, 1]$

$\Delta\alpha_k = \beta_k \Delta p$ $(k \in \mathcal{S})$, $\Delta\alpha_l = \lambda_l \Delta p$ $(l \in \mathcal{U})$, $\Delta b = \beta \Delta p$

β_k, λ_l и β - *коэффициенты чувствительности*.

$$\gamma_i = \frac{\Delta g}{\Delta p} = \sum_{k \in S} Q_{ik}\beta_k + \sum_{l \in \mathcal{U}} Q_{il}\lambda_l + y_i\beta = 0 \ \ \forall \ i \in \mathcal{S} \qquad (9)$$

$$\frac{\Delta h}{\Delta p} = \sum_{k \in \mathcal{S}} y_k\beta_k + \sum_{l \in \mathcal{U}} y_l\lambda_l = 0 \qquad (10)$$

Определяем $\{\lambda_l = C : \forall l \in \mathcal{U}\}$

Находим $\{\beta_k, \beta : \forall k \in \mathcal{S}\}$ из (9),(10)

Находим *граничную чувствительность* γ_i для векторов из $\mathcal{U}$, $\mathcal{E}$ и $\mathcal{R}$

Нужно найти минимальный шаг возмущения $\Delta p_{\min}$.

SVM для нестационарной задачи классификации

Таблица: Возможные адиабатические переходы

До	После	Δp	Условие
$\mathcal{S}$	$\mathcal{R}$	$\frac{-\alpha_i}{\beta i}$	$\beta_i < 0$
$\mathcal{E}$	$\mathcal{S}$	$\frac{-g_i}{\gamma i}$	$\gamma_i > 0$
$\mathcal{R}$	$\mathcal{S}$	$\frac{-g_i}{\gamma i}$	$\gamma_i < 0$
$\mathcal{S}$	$\mathcal{E}$	$\frac{C-\alpha_i}{\beta i}$	$\beta_i > 0$
$\mathcal{U}(g_i < 0)$	$\mathcal{S}$	$\frac{-g_i}{\gamma i}$	$\gamma_i > 0$
$\mathcal{U}(g_i < 0)$	$\mathcal{E}$	$\frac{C-\alpha_i}{\lambda i}$	$\lambda_i > 0$

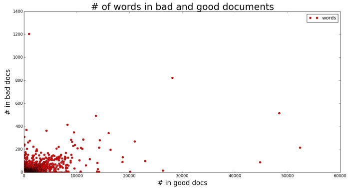
$\mathcal{C}=\{\Delta p\}$. Минимальный шаг возмущения: $\Delta p_{\min} = \min\limits_{c\in\mathcal{C}} \Delta p^c_{\min}$.

Изменяем коэффициенты и категории объектов

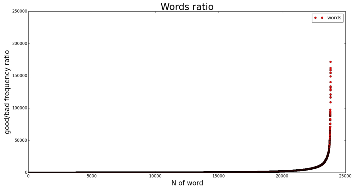
$(\alpha_k \to \alpha_k + \beta_k \Delta p_{\min} : \forall k \in \mathcal{S})$, $(\alpha_l \to \alpha_l + \lambda_l \Delta p_{\min} : \forall l \in \mathcal{U})$

Эксперимент

Частота встречания слов-признаков в $\{1\}$ и $\{-1\}$ классах



(a) $(n_{w\{1\}}, n_{w\{-1\}})$



(b) $\left(w, \dfrac{n_{w\{1\}}}{n_{w\{-1\}}}\right)$

Эксперимент

Статистика $S_i(\mathbf{X}) = \frac{n_{w\{1\}}}{n_{\{1\}}} / \frac{n_{w\{-1\}}}{n_{\{-1\}}}$ - отношение вероятностей встретить слово в $\{1\}$ и $\{-1\}$ классах.
Показана зависимость $S_i(X_{train})/S_i(X_{test})$ от i

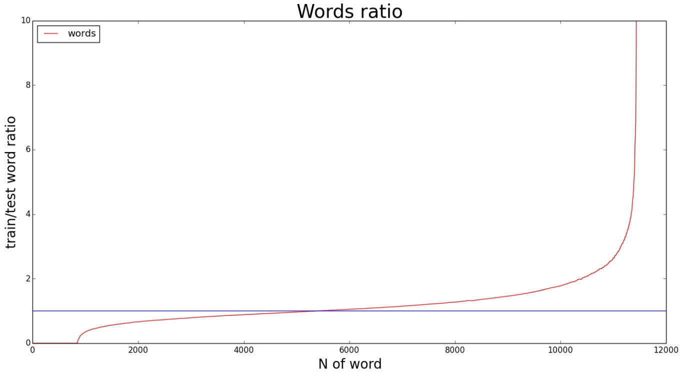


Рис.: Первичный тест нестационарности

Распределение статистики для 1000 итераций алгоритма

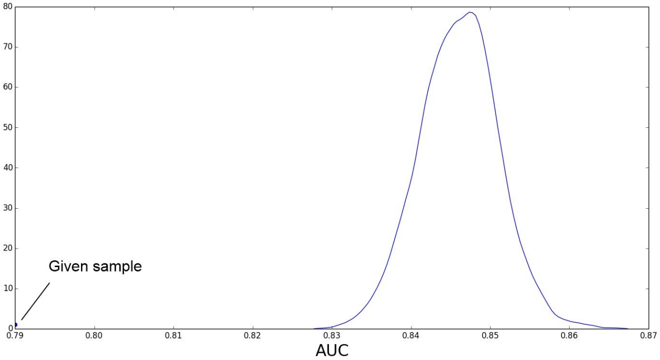


Рис.: Вторичный тест нестационарности

Эксперимент

Таблица: Сравнение SVM и Incremental SVM

Алгоритм	Обучающая выборка	Тестовая выборка	AUC
SVM	5000	10000	0.779
Inc.SVM	4000+1000	10000	**0.791**

Таблица: Сравнение SVM и SVM с отбором признаков

Алгоритм	Обучающая выборка	Тестовая выборка	AUC
SVM	150000	50000	0.782
SVM+	150000	50000	**0.836**

Заключение

- Проведена предварительная обработка данных и выделение численных признаков
- Построен SVM классификатор для стационарной задачи, проведен частичный отбор признаков
- Предложены тесты для определения нестационарности выборки
- Предложен Incremental SVM классификатор для нестационарной задачи
- Получены частичные результаты, демонстрирующие превосходство предложенного алгоритма над базовым