Классификация текстовой информации с помощью SVM

Н.В.Учителев
Кафедра «Предсказательного моделирования и оптимизации» МФТИ, ИППИ РАН
uchitelev.nikita@gmail.com

Аннотация

Проводилась классификация большого числа текстов судебных решений арбитражных судов РФ по тематикам. Тематики были заранее известны, а обучающую и тестовую выборки составляли эксперты-юристы. Обсуждаются результаты классификации в зависимости от тематической окрашенности текстов, а также от размеров выборки и размерности пространства.

1. Введение

Ежедневно на индексацию и обработку в компании, занимающиеся разработкой и поддержкой справочно-правовых систем, поступают десятки тысяч текстов судебных решений со всей страны. Для облегчения поиска нужной информации в таких больших хранилищах данных удобно уметь автоматически разделять новые документы по тематикам на этапе их индексации. Тексты, составленные по итогам судебных разбирательств, имеют строгую регламентированную структуру, а юридический язык подразумевает частое использование одних и тех же терминов. Такие стилистические особенности позволяют достигать сравнительно высоких показателей точности классификации при использовании довольно простых подходов. Впрочем, у этой методики есть много других приложений [1].

В этой статье рассматривается классификация порядка 200 000 документов по 17 тематикам, при этом каждый текст может относиться к одной тематике, сразу к нескольким или ни к какой из них вообще. В качестве алгоритма классификации был выбран метод опорных векторов (Support Vector Machine), как хорошо зарекомендовавший себя [2] для работы в пространствах очень высокой размерности (>10 000) и имеющий большое количество свободно распространяемых высокопроизводительных реализаций. Расчет производился в среде MATLAB с использованием модуля SVM, входящим в состав Bioinformatics Toolbox.

2. Машинное представление текста

Одним из общепринятых подходов [2] является сопоставление тексту вектора признаков, каждая координата которого соответствует частоте встречаемости одного из слов всей коллекции в этом тексте (*tf*, term frequency). Объединение всех таких векторов в единую таблицу приводит нас к прямоугольной матрице размером $n \times p$, где p – количество слов в коллекции (размерность пространства), а n – число документов. Такое представление называется Vector Space Model (VSM). Эта матрица получилась очень разреженой – лишь примерно 1,5% ее элементов отличны от нуля.

Описанная выше величина *tf*, однако, несет не полную частотную информацию, характеризующую слово. Для увеличения информативности ее можно [2] [3] умножать на *idf* – обратную документную частоту, вычисляемую, например, по формуле:

$$idf = ln\frac{D}{d}, \qquad (1)$$

где D – общее число документов в коллекции, а d – число документов, в которых встречается данное слово. Кроме того, каждую *tf* документа можно нормировать на максимальное значение *tf* данного слова по всем документам, чтобы нивелировать влияние длины документа на этот параметр [4]. Однако обе эти поправки в данной работе не использовались, т.к. количество слов в разных судебных решениях и без того варьируется не столь значительно, а вместо умножения на *idf* применялся более практичный метод фильтрации слов, учитываемых при обучении.

3. Фильтрация слов

Общее число различных слов во всей коллекции из 200 000 документов превышает 250 000 – здесь присутствует также числовая информация, а различные морфологические формы одного и того же терма считаются разными словами. После приведения слов к нормальной форме и объединения их характеристик, это число составило чуть более 150 000, что все равно

слишком много для эффективного анализа, а также чрезмерно требовательно к оперативной памяти.

Таким образом, встает задача отсечения большого числа таких слов, которые в качестве признака документа несут в себе мало информации. В перечень подобных слов в первую очередь входят так называемые стоп-слова – служебные части речи, которые необходимы для грамотного построения предложений, однако смысловой нагрузки не несут (например: «и», «на», «чтобы» и т.п.), также исключались числа, т.к. не могут как бы то ни было относиться к тематике документа – в основном это даты, денежные суммы, адреса, номера судебных дел и различные коды.

Далее отбрасывались термы, которые присутствуют почти во всех документах или же в очень малом их числе. В первую группу были отнесены признаки, характерные более чем для 80% документов. Сюда попали такие слова, как «суд», «дело», «постановить», «статья» и другие, типичные для любого судебного решения лексические единицы, которые, тем не менее, не относятся к стоп-словам. Во вторую – слова, встречающиеся менее, чем в 10 документах: это в основном фамилии, опечатки и редкие англоязычные термины. Такой фильтрации оказалось достаточно, чтобы сократить количество значимых слов до 13498.

4. Тематики

Как уже отмечалось, в данной статье рассматривается классификация документов по 17 тематикам, а каждый документ может входить сразу в несколько таких тематик. Подобное распределение обусловлено различной общностью каждого из классов, например, как видно из Таблицы 1, тематика «Правосудие» включает в себя почти все документы коллекции, в то время как «Труд и занятость населения», «Здравоохранение» и «Наука» весьма частные. Т.к. разделение на тематики в этом смысле независимо, то для каждого документа классификатор может выставлять бинарную оценку, относится он к данному классу или нет, по отдельности для каждого класса. Это позволяет оценивать точность классификатора в зависимости от степени общности класса, т.е. доли документов коллекции, которые он покрывает. Кроме того, необходимо отметить, что все документы коллекции уже экспертно разделены на тематики, но некоторое число ошибок, тем не менее, присутствует, что обусловлено историческими особенностями развития данной методики.

5. Машина опорных векторов

Задача бинарной классификации с помощью метода опорных векторов состоит в построении оптимальной разделяющей гиперплоскости в пространстве признаков высокой размерности. Оптимальность здесь понимается в смысле минимизации верхней оценки вероятности ошибки классификации [5].

Таблица 1. Распределение документов по тематикам

Название тематики	Доля
Правосудие	91,25%
Гражданское право	78,33%
Финансы	37,46%
Основы государственного управления	27,72%
Хозяйственная деятельность	15,04%
Окружающая среда и природные ресурсы	8,19%
Международные отношения, международное право	6,71%
Жилище	6,31%
Информация и информатизация	5,28%
Прокуратура, орагны юстиции, адвокатура, нотариат	4,98%
Внешнеэкономическая деятельность	4,31%
Конституционный строй	3,62%
Безопасность и охрана правопорядка	2,71%
Социальное обеспечение и социальное страхование	1,61%
Здравоохранение, спорт, туризм	1,06%
Труд и занятость населения	0,55%
Образование, наука, культура	0,54%

Пусть дана выборка:

$$S = \big((\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n)\big), \quad (2)$$

где $\vec{x}_i \in \mathcal{R}^p$ – набор признаков для i-го документа, а $y_i \in \{-1,1\}$ – бинарная оценка принадлежности этого документа к одному из классов. Тогда вектор $\vec{\omega}$, получаемый из выражения

$$(\vec{\omega} \cdot \vec{\omega}) = \sum_{i=1}^{p} \omega_i^2 \to \min \quad (3)$$

при наложении на него условия для гиперплоскости

$$y_i(\vec{\omega} \cdot \vec{x}_i) + b \geq 1 \quad (4)$$

назовем вектором весов.

5.1 Линейный случай

Обычно, заранее неизвестно, обладают ли документы свойством линейной разделимости, однако во многих задачах по классификации текстов считается [3], что они все же линейно разделимы. В этом случае линейный метод SVM сводится к поиску такой гиперплоскости, что сумма расстояний от ближайших к ней (сверху и снизу) точек максимальна среди всех разделяющих гиперплоскостей, расположенных на равных от них расстоянии. Это квадратичная задача оптимизации лагранжиана

$$L(\vec{\omega}, b, \vec{\lambda}) = \frac{1}{2}(\vec{\omega} \cdot \vec{\omega}) - \sum_{i=1}^{p} \lambda_i \big(y_i\big((\vec{\omega} \cdot \vec{x}_i) + b\big) - 1\big). \quad (5)$$

Решение уравнения $L(\vec{\omega}, b, \vec{\lambda}) = 0$ для $\vec{\omega}$ выглядит так:

$$\vec{\omega} = \sum_{i=1}^{p} \lambda_i \, y_i \vec{x}_i, \qquad \sum_{i=1}^{p} \lambda_i y_i = 0. \quad (6)$$

Пусть лагранжиан (5) достигает максимума по $\vec{\lambda}$ при в точке $\vec{\lambda}^0 = \left(\lambda_1^0, \lambda_2^0, \dots, \lambda_p^0\right)^T$, тогда решение задачи поиска оптимальной гиперплоскости имеет вид [5]:

$$\vec{\omega}_0 = \sum_{i=1}^{p} \lambda_i^0 \, y_i x_i, \tag{7}$$

при этом свободный член равен

$$b_0 = \frac{1}{2}\left(\min_{y_i=1}(\vec{\omega}_0 \cdot \vec{x}_i) + \max_{y_i=-1}(\vec{\omega}_0 \cdot \vec{x}_i)\right). \tag{8}$$

Нужные нам решения должны дополнительно удовлетворять условиям Каруша-Куна-Таккера [5], которые являются необходимым условием оптимальности задачи нелинейного программирования:

$$\lambda_i^0\left(y_i\left((\vec{\omega}_0 \cdot \vec{x}_i) + b\right) - 1\right) = 0, \tag{9}$$

из которых следует, что $\lambda_i^0 > 0$ может выполняться только для тех i, для которых $y_i\left((\vec{\omega}_0 \cdot \vec{x}_i) + b\right) - 1 = 0$, т.е. для тех векторов, которые лежат на гиперплоскостях $(\vec{\omega}_0 \cdot \vec{x}_i) + b_0 = \pm 1$. Эти векторы называются опорными.

При вычислении параметров оптимальной гиперплоскости остальные неопорные векторы можно отбросить, т.к. их значения никак на нее не влияют.

5.2 Нелинейный случай

Если предполагать, что тексты все же линейно неразделимы, необходимо воспользоваться несколько иной формой метода опорных векторов. Отличие заключается в том, что теперь мы предполагаем невозможность построения оптимальной разделяющей гиперплоскости в пространстве признаков типа VSM, зато в некотором пространстве большей размерности разделение документов на классы возможно таким же линейным методом [6]. Для этого необходимо перевести наблюдения (документы) в пространство признаков с помощью некоторого нелинейного отображения:

$$\vec{x} = (x_1, \dots, x_n)^T \to \vec{\varphi}(\vec{x}) = \left(\varphi_1(\vec{x}), \dots, \varphi_m(\vec{x})\right)^T, \tag{10}$$

где $m > n$. Тогда очевидно, что прообразом в $\mathcal{R}^n$ оптимальной в $\mathcal{R}^m$ гиперплоскости будет в общем случае нелинейная поверхность, задаваемая уравнением:

$$\sum_{j=1}^{m} \omega_j \varphi_j(\vec{x}) + b = \sum_{i=1}^{n} \lambda_i^0 y_i K(\vec{x}, \vec{x}_i). \tag{11}$$

Здесь функция $K(\vec{x}, \vec{x}_i) = \left(\vec{\varphi}(\vec{x}) \cdot \vec{\varphi}(\vec{x}_i)\right)$ называется ядром. В этой работе для сравнения с линейным SVM приводятся результаты классификации с квадратичным ядром:

$$K(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y}^T)^2, \tag{12}$$

которое имеет смысл квадрата скалярного произведения двух векторов.

Бинарная классификация проводилась многократно для каждой тематики отдельно.

6. Генерация выборок

Все тексты, участвующие в экспериментах, предварительно были разделены экспертами-юристами на 17 тематик, другими словами, они сопоставили каждой тематике список номеров документов, которые к ней относятся. Из этих списков случайным образом выбиралось равное количество различных номеров для обучающей и для тестовой выборки, затем они дополнялись тем же объемом документов, которые не относятся к выбранному классу [4]. При этом для каждой тематики проводилось обучение на нескольких объемах выборки, чтобы оценить зависимость точности классификации от этой величины. Кроме того, в виду случайности выбора номеров, для каждого такого объема оказалось оправданным генерировать по 5 наборов документов, чтобы затем усреднять ошибку.

После составления выборок, как правило, оказывалось так, что каких-то слов нет ни в одном из документов, таким образом, на каждом этапе обучения можно было дополнительно снизить размерность, не учитывая такие слова. Количество признаков, которые отличны от нуля хоть в одном наблюдении, в данной работе называется эффективной размерностью. Она в является характеристикой каждой конкретной выборки; зависимость ошибки классификации от этого параметра в данной работе также отслеживалась.

7. Результаты экспериментов

Основными результатами экспериментов являются величины ошибок, полученные при разных условиях. Ошибки классификации бывают двух типов:

- ошибка первого рода равна доле случаев, когда классификатор не отнес нужный документ к текущему классу
- ошибка второго рода равна доле случаев, когда классификатор отнес документ к текущему классу, однако на самом деле данный документ в него не входит.

В этой работе сравниваются результаты классификации с линейным и квадратичным ядром, кроме того отмечается зависимость количества опорных векторов от параметров выборки.

7.1 Зависимость ошибки от объема класса

Важной особенностью распределения текстов по тематикам в данной работе является большой разброс долей коллекции, объединяющих их в различные классы. Здесь представлены как большие классы, покрывающие более 90% всех документов, так и маленькие порядка 0,5% от общего числа. Интересно проанализировать зависимость точности классификации от этого показателя в первую очередь по тому, что доля документов, занимаемая каждой из тематик, наглядно характеризует общность соответствующей темы. При этом очевидно, что чем

выше общность тематики, тем сложнее классифицировать документы по набору входящих в них слов. Рис.1 экспериментально подтверждает эту гипотезу:

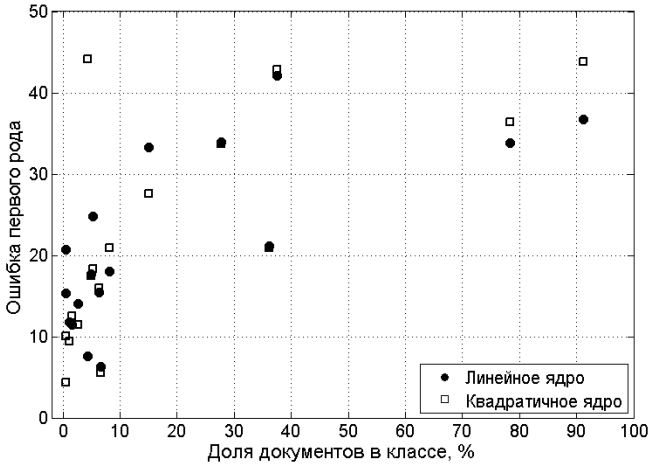

Рис.1 - Зависимость ошибки 1-го рода от объема класса

Из рис.1 видно, что наименьшая ошибка классификации первого рода достигается на тематиках, которые занимают не более 10% выборки. Самый лучший результат был показан на тематике «Международные отношения», состоящей из 6,71% документов, что равно примерно 13000 документов. Скорее всего, словами-маркерами здесь оказались названия стран, которые встречаются в каждом документе этого класса. Самый высокая ошибка в 35% был показана на самой объемной тематике «Правосудие», что объясняется отсутствием характерных черт, объединяющих столь большое количество судебных решений. Аналогичная ситуация наблюдается и для ошибок второго рода с той лишь разницей, что для квадратичного ядра эта ошибка как правило в 2-3 раза больше аналогичной для линейного классификатора (рис.2).

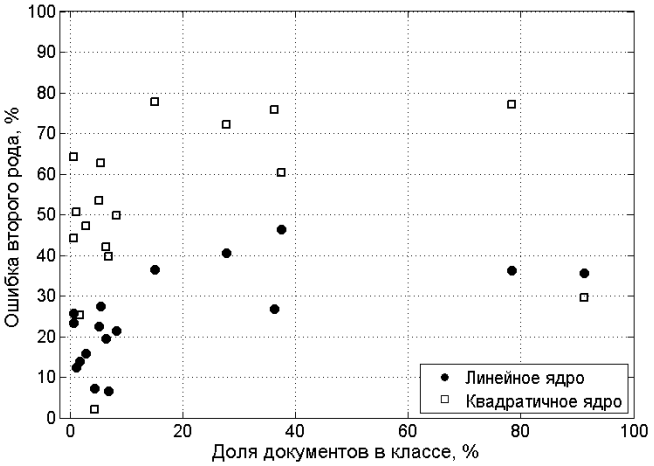

Рис.2 - Зависимость ошибки 2-го рода от объема класса

Вообще классификация с квадратичном ядром в большей части случаев показывает сравнимый с линейным алгоритмом уровень ошибок первого рода, при этом ошибки второго рода, как правило, неприемлемо высоки.

Таблица 2. Ошибки классификации для разных тематик (линейный алгоритм)

Название тематики	Ошибка 1-го рода	Ошибка 2-го рода
Правосудие.	36,75%	35,60%
Гражданское право.	33,87%	36,33%
Финансы.	42,13%	46,29%
Основы государственного управления.	33,95%	40,53%
Хозяйственная деятельность.	33,25%	36,46%
Окружающая среда и природные ресурсы.	18,04%	21,40%
Международные отношения. Международное право.	6,32%	6,52%
Жилище.	15,41%	19,53%
Информация и информатизация.	24,77%	27,41%
Прокуратура. Орагны юстиции. Адвокатура. Нотариат.	17,68%	22,39%
Внешнеэкономическая деятельность.	7,53%	7,22%
Конституционный строй.	21,16%	26,72%
Безопасность и охрана правопорядка.	13,98%	15,84%
Социальное обеспечение и социальное страхование.	11,41%	13,89%
Здравоохранение. Спорт. Туризм.	11,82%	12,46%
Труд и занятость населения.	15,30%	25,77%
Образование. Наука. Культура.	20,73%	23,43%

7.2 Зависимость ошибки от размера обучающей выборки

Как показала практика, величина ошибки практически не зависит от размера выборки, начиная с определенного момента (когда документов становится достаточно), и, как правило, ее динамика напоминает случайные колебания вокруг среднего значения. В ходе экспериментов алгоритм автоматически старался сгладить эти колебания и выбрать оптимальный размер выборки для каждой тематики, на котором и были получены результаты. При этом у небольших классов (не более 7% выборки) эта точка делила всю выборку примерно пополам, а у больших классов составляла сравнимую с ней величину, таким образом, в среднем размер обучающей выборки составил порядка 4000 – 6000 документов для разных классов.

Более интересной оказалась зависимость ошибки классификации от доли документов тематики (рис.3 и рис.4), которую составляла обучающая выборка. Вспомним, что половину обучающей выборки мы брали из класса, а вторую половину извне. Облако точек справа на рисунке соответствует маленьким классам, точки в левой части рисунка – большим.

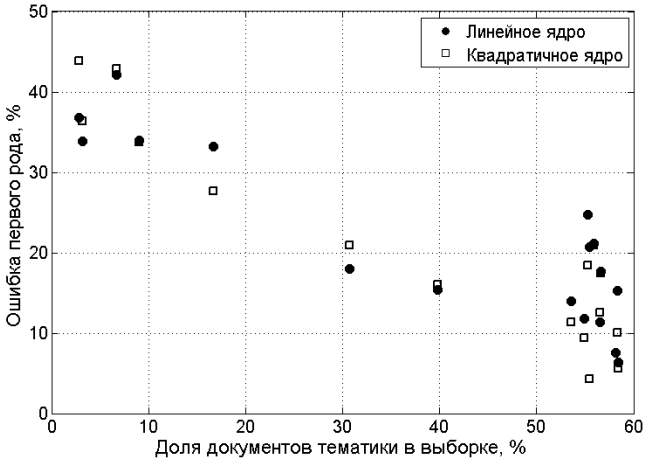

Рис.3 – Зависимость ошибки 1-го рода от доли обучающей выборки в классе.

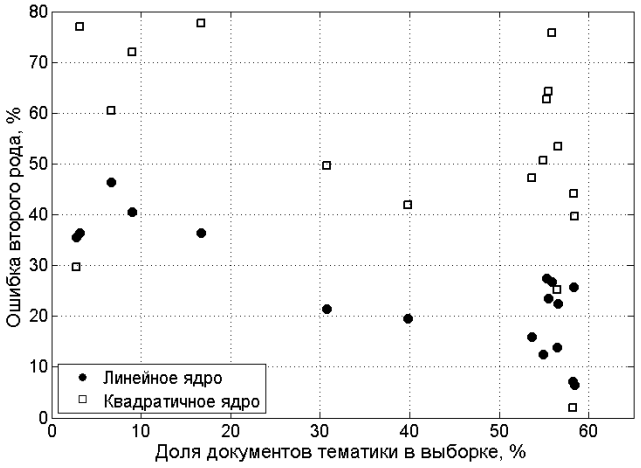

Рис.4 - Зависимость ошибки 2-го рода от доли обучающей выборки в классе.

7.3 Зависимость ошибки от количества опорных векторов

Как отмечалось выше, не все решения вида (7) оказывают влияние на оптимальную разделяющую гиперплоскость, а только те, которые удовлетворяют условию (9). Зависимость ошибки классификации от этого числа для двух ядер представлена на рис.5 и рис.6 и интересна тем, что по логике метода, чем меньше найдено опорных векторов, тем ниже размерность этой гиперплоскости.

Из рисунков видно, что результат классификации текстов тем лучше, чем ниже размерность пространства, в котором возможно их разделить. Оно и понятно, ведь можно считать, что число найденных опорных векторов приблизительно соответствует количеству слов-маркеров. В вырожденном случае, если наличие в тексте всего одного слова является точным признаком того, что он относится к какому-то классу, то вне зависимости от остальных слов, этот класс можно отделить одномерной плоскостью.

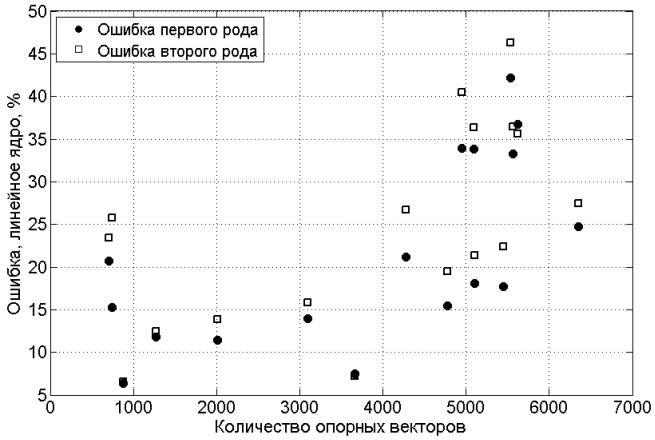

Рис.5 – Зависимость ошибки от числа опорных векторов для линейного ядра

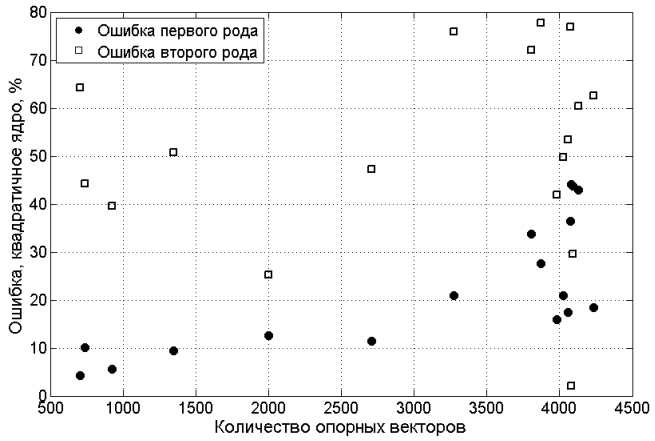

Рис.6 - Зависимость ошибки от числа опорных векторов для квадратичного ядра

7.4 Зависимость числа опорных векторов от других параметров

Естественным образом встает вопрос, от чего зависит число опорных векторов. В данной работе рассмотрены зависимости их числа от эффективной размерности (числа признаков, отличных от нуля хотя бы в одном наблюдении, попавшем в обучающую и тестовую выборки) и от объема класса, т.е. доли всех документов, которые входят в каждую тематику.

На рис.7 и рис.8 видно, что минимальное число опорных векторов достигается при уменьшении объема тематики и максимальной информативности признаков. Здесь нужно понимать, что количество нулевых факторов зависит от числа документов в обучающей выборке – чем больше документов в выборке, тем больше признаков нужно задействовать, чтобы научиться разделять две группы наблюдений. Поэтому здесь важно соблюдать баланс.

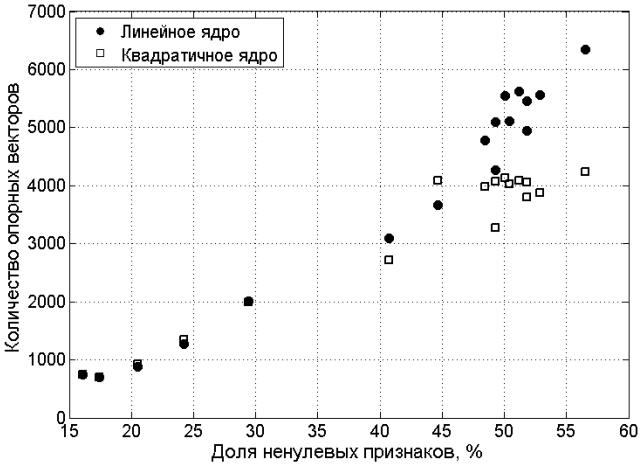

Рис.7 – Зависимость количества опорных векторов от эффективной размерности выборки

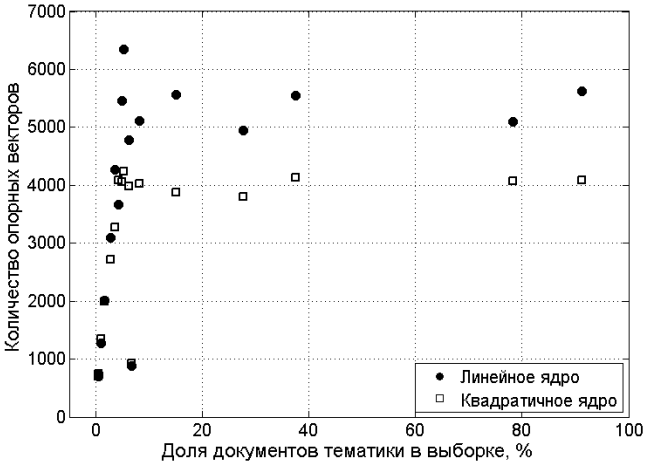

Рис.8 – Зависимость количества опорных векторов от объема класса

8. Заключение

В этой работе представлены результаты серии экспериментов по классификации большого числа текстовых документов на тематики, заранее определенные экспертами. В качестве алгоритма классификации используется метод опорных векторов (SVM), при этом классификация происходит на «чистых» данных, т.е. без дополнительных алгоритмов повышения точности. По итогам исследований определены зависимости между ошибками, допускаемыми после машинного обучения, и различными параметрами задачи, из которых можно сделать дальнейшие предположения о том, в каком направлении нужно развивать алгоритм, чтобы повысить точность.

К числу таких предложений можно отнести различные алгоритмы снижения размерности входных данных, которые в состоянии повысить плотность информации, содержащейся в простой VSM-модели машинного представления текста. Однако, на первый взгляд неочевидно, какой из многочисленных подходов будет наиболее эффективным, т.к. топология входных данных неясна. Кроме того, можно использовать один из алгоритмов бустинга, что вполне естественно для такого рода задач. Это позволит повысить точность за счет применения целого ансамбля однородных классификаторов, обучаемых на различных выборках. Отдельно следует уделить внимание отбору наблюдений для составления обучающей выборки, т.к. от этого может заметно увеличиться точность, по сравнению со случайной генерацией.

Также в данной работе сравнивались результаты классификации с применением линейного и квадратичного ядер. По итогам этого сравнения, предположение о том, что тексты в первом приближении можно считать линейно разделимыми оправдалось, т.к. средняя ошибка линейного классификатора оказалось выше, чем квадратичного, который проводит линейную классификацию в искусственном пространстве более высокой размерности. Однако, полная уверенность в возможности такой разделимости отсутствует, т.к. квадратичное ядро может быть не самым оптимальным в классе нелинейных ядер. В связи с этим рекомендуется использовать линейный алгоритм, дополненный учетом ошибок, обусловленных линейной неразделимостью, в виде вектора переменных мягкого отступа [5].

В среднем описанный в статье подход позволяет классифицировать тексты с ошибкой 20%, однако, по нескольким особенным тематикам удалось снизить это значение до 4-5%, что является довольно хорошим результатом для столь простой методики. Компанию, заказавшую данное исследование, удовлетворил классификатор с ошибкой порядка 16%. Такая точность была принята за промышленный стандарт и в данный момент используется в продуктах, а обширные тематики, покрывающие более 30% документов были рекомендованы к упразднению. К сожалению, данный технический отчет недоступен для свободного чтения.

Список литературы

[1] F.Sebastiani, Machine Learning in Automated Text Categorization, New York, USA: Journal "ACM Computing Surveys" (CSUR), 2002, pp. 1-47.

[2] J.Shawe-Taylor, N.Cristianini, Kernel Methods for Pattern Analysis, New York: Cambrige University Press, 2004.

[3] T.Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Dortmund: Springer Berlin Heidelberg, 1998, pp. 137-142.

[4] Y.Yang, J.O.Pedersen, A Comparative Study of Feature Selection in Text Categorization, Machine Learning-International Workshop: Morgan Kaufman Publishers, INC, 1997, pp. 412-420.

[5] В.В.Вьюгин, Элементы математической теории машинного обучения, Москва: МФТИ, 2010.

[6] T.Hastie, R.Tibshirani, J.Friedman, The Elements of Statistical Learning, Springer, 2003.