

**ПРИМЕНЕНИЕ В ЗАДАЧЕ КЛАССИФИКАЦИИ SMS СООБЩЕНИЙ  
ОПТИМИЗИРОВАННОГО НАИВНОГО БАЙЕСОВСКОГО КЛАССИФИКАТОРА**

© 2016 М.Е. Бурлаков

Самарский национальный исследовательский университет им. С.П. Королёва

Статья поступила в редакцию 11.11.2016

В статье рассматривается процесс оптимизации наивного байесовского классификатора. Рассмотрен механизм расчета вероятности и процесс построения обучающей таблицы для практического решения задачи классификации SMS сообщений наивным байесовским классификатором. В качестве экспериментальной выборки было взято множество SMS сообщений, которое используется специалистами в области классификации электронных сообщений (обучающее и тестовое множество). Качество классификации SMS сообщений у оптимизированного метода оказалось выше, чем у обычного наивного байесовского классификатора.

*Ключевые слова:* наивный байесовский классификатор, классификация SMS сообщений, оптимизация наивного байесовского классификатора, электронные сообщения, спам, классификация электронных сообщений.

С момента появления технологии СМС сообщений прошло более 15 лет, и как всякая технология были этапы интенсивного роста, стабильной эксплуатации и медленного угасания (наблюдается в настоящее время [1,2]). Однако, несмотря на снижение популярности технологии СМС, в настоящее время, это единственная технология, способная доставлять текстовые сообщения поверх GSM трафика в отсутствие активного Интернет соединения, становясь, таким образом, гарантированным способом доставки (в случае активного статуса абонента в сети). Благодаря данной отличительной черте, СМС остается активным каналом распространения заведомо неинтересной, а порой и зловредной, информации для конечного потребителя, классифицируемой далее как спам. Также, несмотря на снижение СМС трафика, количество спама в нем продолжает расти [3]. Таким образом, задача внедрения существующих и разработка новых методов классификации сообщений является актуальной.

В настоящее время существует большое количество решений основанных как на неадаптивных, так и адаптивных алгоритмах. Среди подобных решений можно выделить: искусственные нейронные сети, искусственные иммунные алгоритмы, деревья решений, методы ближайших соседей ( $k$ -NN алгоритмы), регрессионные модели, вероятностные подходы и т.д. Также, большое внимание уделяется разработке новых алгоритмов, с использованием последних достижений в области машинного обучения и *Data Mining*'а.

Наиболее популярным решением, установленным по умолчанию на многих программных

комплексах и аппаратах пользователей, при классификации СМС сообщений является их определение в один из двух классов: класс спам ( $S$ ) сообщений и сообщений, не относящихся к спаму ( $nS$ ) [4]. Техника классификации зачастую основана на обнаружении ранее внесенных ключевых слов в некоторый «черный список». Например для ОС *Android* применяется технология *Spam SMS blocker* или *SMS spam runner* в ОС *Symbian*. Функционирование системы классификации основано на конечном числе ключевых слов в спам словаре и зачастую не удовлетворяют своими характеристиками при процессе классификации сообщений для конечного пользователя.

Однако, несмотря на постоянные разработки и совершенствование текущих алгоритмов, наивный байесовский классификатор (НБК) является одним из наиболее популярных алгоритмов [5] по причине простоты реализации и минимальных как человеческих, так и финансовых затрат при внедрении в информационные системы (ИС) компаний-провайдеров или на устройствах конечных потребителей (пользователей).

Простота и формализация подхода в НБК, при которой все слова в СМС сообщении не зависят друг от друга, делает его менее эффективным в случае, если бы учитывался и порядок слов. Поэтому его оптимизация, учитывающая данный аспект есть актуальная задача. Для этого предлагается использование основанного на НБК алгоритма, рассматривающего СМС не только как набор слов, но и как набор некоторых объектов (группы слов). Предполагается, что данный подход позволит повысить общую точность процесса классификации за счет расчета количества не только каждого слова, но и количества групп слов входящих в одно сообщение.

### МЕХАНИЗМ РАСЧЕТА ВЕРОЯТНОСТИ

Наивный байесовский классификатор (НБК) – один из наиболее примитивных классификаторов, основанных на теореме Байеса с условием выполнения строгой независимости вероятностных компонент [6]. Данное допущение рассматривает каждое слово в сообщении отдельно и независимо от остальных. Вследствие этого, процесс обучения НБК сложен и нетривиален в силу затраты большого количества ресурсов для получения минимальной базы, пригодной для дальнейшей классификации сообщений.

Рассмотрим примитивное СМС сообщение  $\vec{x}$ , состоящее из конечного набора слов  $X_1, \dots, X_n$ :  $\vec{x} = \langle X_1, \dots, X_n \rangle$ .

Обозначим за  $Y$  класс (в нашем случае это классы  $S$  и  $nS$ ), к которому данное сообщение может быть отнесено. Тогда для любых двух слов  $X_1$  и  $X_2$  из сообщения  $\vec{x}$  вероятность отнесения к классу  $Y$  равно:

$$\begin{aligned} p(\vec{x} | Y) &= p(X_1, X_2 | Y) = \\ &= p(X_1 | X_2, Y) p(X_2 | Y) = \\ &= p(X_1 | Y) p(X_2 | Y) \end{aligned} \quad (1)$$

Экстраполируя на конечное число слов  $X_1, \dots, X_n$ , получим соотношение:

$$p(X_1, \dots, X_n | Y) = \prod_{i=1}^n p(X_i | Y). \quad (2)$$

С другой стороны, вероятность того что  $Y$  примет  $l$ -ое возможное значение для сообщения, состоящего из  $X_1, \dots, X_n$  слов, в соответствии с теоремой Байеса равна:

$$\begin{aligned} p(Y = y_l | X_1, \dots, X_n) &= \\ &= \frac{p(Y = y_l) p(X_1, \dots, X_n | Y = y_l)}{\sum_j p(Y = y_j) p(X_1, \dots, X_n | Y = y_j)}. \end{aligned} \quad (3)$$

В силу полной независимости слов сообщения друг от друга, вероятность примет вид:

$$\begin{aligned} p(Y = y_l | X_1, \dots, X_n) &= \\ &= \frac{p(Y = y_l) \prod_i p(X_i | Y = y_l)}{\sum_j p(Y = y_j) \prod_i p(X_i | Y = y_j)}. \end{aligned} \quad (4)$$

Получив механизм расчета вероятности, рассмотрим процесс построения обучающей таблицы. Обучающая таблица позволяет систематизировать появление тех или иных слов с дальнейшим расчетом вероятности отнесения сообщений, в которых данные слова встречаются к классу  $Y$  (классам  $S$  и  $nS$ ).

### ПОСТРОЕНИЕ ОБУЧАЮЩЕЙ ТАБЛИЦЫ

Для наглядности описания процесса построения обучающей таблицы для СМС сообщений, рассмотрим конкретный пример. Пусть дано пять СМС сообщений: два сообщения  $\vec{x}_1 = \langle X_1 \rangle$ ,  $\vec{x}_2 = \langle X_1 \_ X_2 \rangle$  относящиеся к классу  $nS$  и три  $\vec{x}_3 = \langle X_3 \rangle$ ,  $\vec{x}_4 = \langle X_2 \_ X_3 \rangle$ ,  $\vec{x}_5 = \langle X_2 \_ X_3, X_2 \_ X_3! \rangle$  к классу  $S$ . Тогда таблица векторов для выборки будет иметь вид (табл. 1):

В данном случае, из СМС сообщений были извлечены слова  $\langle X_1, X_2, X_3 \rangle$ , которые являются характеристическими признаками при дальнейшей классификации СМС сообщений. Извлечение слов подразумевает отсутствие учета знаков препинания. Тогда общее кол-во слов входящих в тот или иной класс (табл. 2):

В случае появления в системе нового сообщения  $\vec{x}$ , процесс выделения признаков (слов) будет состоять из следующих этапов:

1. Из сообщения удаляются все знаки препинания и выделяются только слова;
2. Слова не встречающиеся в исходной векторной таблице исключаются;
3. Подсчитывается общее количество оставшихся слов.

В нашем представленном примере, вероятность того, что новое СМС сообщение  $\vec{x}$  будет отнесено к классу  $nS$  равно:

$$\begin{aligned} p(nS | X_1, X_2, X_3) &= \\ &= p(nS) p(X_1 | nS) p(X_2 | nS) p(X_3 | nS). \end{aligned} \quad (5)$$

А вероятность того что сообщение  $\vec{x}$  будет отнесено к классу  $S$  равно:

$$p(S | X_1, X_2, X_3) = p(S) p(X_1 | S) p(X_2 | S) p(X_3 | S). \quad (6)$$

Таким образом, рассчитав значение обеих вероятностей можно сказать, к какому классу будет отнесено сообщение  $\vec{x}$ .

### ОПТИМИЗАЦИЯ НБК

Для оптимизации НБК, основанной не только на анализе слов как независимых компонент СМС сообщения, но и учитывающей последовательность слов (группы слов) опишем необходимые этапы.

#### Построение обучающей таблицы

Построение обучающей таблицы для оптимизации несколько отличается от того построения, что описано выше. Оно учитывает не только правило независимости слов в СМС сообщении, но и правило ассоциации слов друг с другом в рамках сообщения (учет порядка следования слов). Для учета данного аспекта опишем процесс построения ассоциаций слов.

Таблица 1. Векторная таблица сообщений

Сообщение	Класс	Количество слов		
		$X_1$	$X_2$	$X_3$
$\vec{x}_1$	$nS$	1	0	0
$\vec{x}_2$	$nS$	1	1	0
$\vec{x}_3$	$S$	0	0	1
$\vec{x}_4$	$S$	0	1	1
$\vec{x}_5$	$S$	0	2	2

Таблица 2. Количество вхождений слов в определенные классы

Слова	Класс $nS$	Класс $S$
$X_1$	2	0
$X_2$	1	3
$X_3$	0	4
Итого	3	7

### Процесс построения ассоциаций

Для наглядности процесса построения ассоциаций рассмотрим конкретный пример. Пусть дано 9 СМС сообщений из классов  $nS$  и  $S$ : 5 из класса  $S$  и 4 –  $nS$  соответственно.

$$S: \vec{x}_1 = [X_1, X_2, X_5], \vec{x}_2 = [X_2, X_3],$$

$$\vec{x}_3 = [X_1, X_3], \vec{x}_4 = [X_1, X_2],$$

$$\vec{x}_5 = [X_1, X_2, X_3]$$

$$nS: \vec{x}_6 = [X_2, X_4], \vec{x}_7 = [X_1, X_2, X_4],$$

$$\vec{x}_8 = [X_2, X_3], \vec{x}_9 = [X_1, X_2, X_3, X_5]$$

Максимальная длина анализируемых ассоциаций не может быть больше минимальной длины сообщений. В нашем случае, минимальная длина равна 2 (сообщения  $\vec{x}_2, \vec{x}_3, \vec{x}_4, \vec{x}_6, \vec{x}_8$ ). Таким образом, множество слов, характеризующих сообщения, входящие в класс  $S$  состоит из элементов  $[X_1], [X_2], [X_3], [X_5], [X_1, X_2], [X_1, X_3], [X_2, X_3]$ , тогда как множество  $nS$  состоит из элементов  $[X_1], [X_2], [X_3], [X_4], [X_5], [X_1, X_2], [X_2, X_3], [X_2, X_4]$ .

Составим таблицу векторов для классов  $nS$  и  $S$ . Для этого из каждой определенной последова-

тельности для своего соответствующего класса, определим количество вхождений в соответствующее сообщение. Тогда, таблица векторов для класса  $S$  будет иметь следующий вид (табл. 3).

Таблица векторов для класса  $nS$  имеет следующий вид (табл. 4).

Далее сформируем таблицу вхождения слов в определенный класс (табл. 5).

Таким образом:

Общее число слов в словаре  $|X| = 9$ ;

Вероятность определения слова в класс  $nS$   $p(nS) = 4/9$ ;

Вероятность определения слова в класс  $S$   $p(S) = 5/9$ ;

Общее число слов класса  $nS$   $|nS| = 19$ ;

Общее число слов класса  $S$   $|S| = 17$ .

Рассмотрим конкретный пример, возьмем тестовое СМС сообщение вида  $\vec{x} = \langle X_1, X_1, X_2, X_2, X_3 \rangle$ . Тогда, в рамках построенной модели вероятность того, что сообщение будет отнесено к классу  $nS$  равно:

$$p(nS, X_1, X_1, X_2, X_2, X_3) = p(nS) \times p(X_1 | nS)^2 \times p(X_2 | nS)^2 \times p(X_3 | nS) \times p(X_1, X_2 | nS) \times p(X_1, X_3 | nS) \times p(X_2, X_3 | nS)$$

Таблица 3. Таблица векторов класса  $S$ 

Сообщение	Признаки (слова и их последовательности)						
	$X_1$	$X_2$	$X_3$	$X_5$	$X_1, X_2$	$X_1, X_3$	$X_2, X_3$
$\vec{x}_1$	1	1	0	1	1	0	0
$\vec{x}_2$	0	1	1	0	0	0	1
$\vec{x}_3$	1	0	1	0	0	1	0
$\vec{x}_4$	1	1	0	0	0	1	0
$\vec{x}_5$	1	1	1	0	1	1	1

**Таблица 4.** Таблица векторов класса  $nS$

Сообщение	Признаки (слова и их последовательности)							
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_1, X_2$	$X_2, X_3$	$X_2, X_4$
$\vec{x}_6$	0	1	0	1	0	0	0	1
$\vec{x}_7$	1	1	0	1	0	1	0	1
$\vec{x}_8$	0	1	1	0	0	0	1	0
$\vec{x}_9$	1	1	1	0	1	1	1	0

**Таблица 5.** Вхождение слов в определенный класс

Признаки	Класс $nS$	Класс $S$	Признаки	Класс $nS$	Класс $S$
$X_1$	4	2	$X_1, X_2$	2	2
$X_2$	3	4	$X_1, X_3$	3	0
$X_3$	4	2	$X_2, X_3$	2	2
$X_4$	0	2	$X_2, X_4$	0	2
Итого:	Класс $nS$ :19, Класс $S$ : 17				

С другой стороны, вероятность того, что сообщение будет отнесено к классу  $S$  равно:

$$p(S, X_1, X_1, X_2, X_2, X_3) = p(S) \times p(X_1 | S)^2 \times p(X_2 | S)^2 \times p(X_3 | S) \times p(X_1, X_2 | S) \times p(X_1, X_3 | S) \times p(X_2, X_3 | S)$$

Исходя из вышесказанного, проведем сравнение с индивидуальным коэффициентом вероятности каждого слова и получим следующий результат:

$$p(X_1 | nS) = \frac{4+1}{19+|X|} = \frac{5}{28}, p(X_2 | nS) = \frac{3+1}{19+|X|} = \frac{4}{28}, p(X_3 | nS) = \frac{4+1}{19+|X|} = \frac{5}{28},$$

$$p(X_1, X_2 | nS) = \frac{2+1}{19+|X|} = \frac{3}{28}, p(X_1, X_3 | nS) = \frac{3+1}{19+|X|} = \frac{4}{28}, p(X_2, X_3 | nS) = \frac{2+1}{19+|X|} = \frac{3}{28},$$

$$p(X_1 | S) = \frac{2+1}{17+|X|} = \frac{3}{26}, p(X_2 | S) = \frac{4+1}{17+|X|} = \frac{5}{26}, p(X_3 | S) = \frac{2+1}{17+|X|} = \frac{3}{26},$$

$$p(X_1, X_2 | S) = \frac{2+1}{17+|X|} = \frac{3}{26}, p(X_1, X_3 | nS) = \frac{0+1}{17+|X|} = \frac{1}{26}, p(X_2, X_3 | nS) = \frac{2+1}{17+|X|} = \frac{3}{26},$$

Исходя из этого значение

$$p(nS, X_1, X_1, X_2, X_2, X_3) = \frac{4}{9} \times \left(\frac{5}{28}\right)^2 \times \left(\frac{4}{28}\right)^2 \times \left(\frac{5}{28}\right) \times \left(\frac{3}{28}\right)^2 \times \left(\frac{4}{28}\right) \times \left(\frac{3}{28}\right) = 9 \times 10^{-9}$$

$$p(S, X_1, X_1, X_2, X_2, X_3) = \frac{5}{9} \times \left(\frac{3}{26}\right)^2 \times \left(\frac{5}{26}\right)^2 \times \left(\frac{3}{26}\right) \times \left(\frac{3}{26}\right)^2 \times \left(\frac{1}{26}\right) \times \left(\frac{3}{26}\right) = 1.8 \times 10^{-9}$$

Исходя из проведенных вычислений, можно сказать, что тестовое сообщение  $\vec{x} = \langle X_1, X_1, X_2, X_2, X_3 \rangle$  с большей вероятностью будет отнесено к классу  $nS$ , нежели  $S$ .

### СРАВНЕНИЕ НБК И ОПТИМИЗИРОВАННОГО АЛГОРИТМА НА ПРАКТИКЕ

Для практической реализации предложенной идеи бралась база данных *SMS Spam Collection v.1* [7], состоящая из 5574 СМС из  $nS$  и  $S$  классов.

Обучение было построено следующим образом, брались первые 600 строк из представленной БД и на базе них строилась векторная таблица, далее бралось 100 сообщений и проводилось тестирование полученного алгоритма. Далее брались следующие 600 строк, и к ним следующие 100 для тренировки. Процесс был повторен 5 раз. Результаты работы представлены в табл. 6.

Таким образом, для выбранной коллекции СМС сообщений эффективность оптимизированного алгоритма на 1.68% выше, нежели чем у НБК.

Таблица 6. Сводная таблица результатов работы алгоритма

№ тестовой выборки	Оптимизированный алгоритм	НБК(%)	Разница(%)
1	97.5	96.0	+1.5
2	98.2	96.3	+1.9
3	96.1	95.5	+0.6
4	99.0	97.2	+2.8
5	98.4	96.8	+1.6
Итого (среднее):	97.84	96.36	+1.68

### ЗАКЛЮЧЕНИЕ

При рассмотрении СМС сообщений не только с точки зрения отдельных слов но и их ассоциаций, точность предложенного оптимизированного алгоритма выше, нежели чем у НБК. Таким образом, более полный анализ структуры СМС сообщений позволяет не только повысить качество классификации, но и открывает большие перспективы с точки зрения разработки новых алгоритмов и оптимизации существующих.

### СПИСОК ЛИТЕРАТУРЫ

1. Portio Research оценивает мировой рынок мобильных сообщений в 2011 году [Электронный ресурс]. 2011. URL: <http://www.mforum.ru/news/article/099200.htm> (дата обращения 15.09.2016).
2. Worldwide A2P SMS Markets 2014-2017 [Электронный ресурс]. 2014. URL: <http://www.strikeiron.com/wp-content/uploads/2014/12/whitepaper-sms-2014-2017-portio-research.pdf> (дата обращения 15.09.2016).
3. Cloudmark наблюдает рост потоков SMS-спама [Электронный ресурс]. 2013. URL: <https://securelist.ru/blog/novosti/3684/cloudmark-nablyudaet-rost-potokov-sms-spama/> (дата обращения 15.09.2016).
4. Machine learning methods for spam e-mail classification [Электронный ресурс]. 2011. Режим доступа: <http://airccse.org/journal/ijcsit/0211ijcsit12.pdf> (дата обращения 15.09.2016).
5. Weiss S., Apte C. Maximizing text-mining performance // IEEE Intelligent Systems. 1999. 63–69 с.
6. Наивный байесовский классификатор [Электронный ресурс]. 2015. URL: [https://ru.wikipedia.org/wiki/Наивный\\_байесовский\\_классификатор](https://ru.wikipedia.org/wiki/Наивный_байесовский_классификатор) (дата обращения 15.09.2016).
7. SMS Spam Collection v. 1 [Электронный ресурс]. 2011. URL: <http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/> (дата обращения 15.09.2016).

### USING OPTIMIZE NAÏVE BAYES CLASSIFIER IN PROBLEM OF SMS CLASSIFICATION

© 2016 M.E. Burlakov

Samara National Research University named after Academician S.P. Korolyov

In article the process of optimizing a naive Bayesian classifier is viewed. The mechanism of calculating the probability and the process of constructing a training table are analyzed. It's needed for a practical solution to problem of SMS messages classification by Naive Bayes classifier. There is experimental set of SMS messages. It used by experts for SMS messages classification process. It has two sets – train set and test set. The quality of SMS messages classification is higher for optimized method and lower of non-optimized Naive Bayesian classifier.

*Keywords:* Naive Bayes classifier, the classification of SMS messages, the optimization of the Naive Bayesian classifier, e-mails, spam, classification of e-mails.