М.Е. Бурлаков

ОПТИМИЗАЦИЯ НАИВНОГО БАЙЕСОВСКОГО КЛАССИФИКАТОРА ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ СМС СООБЩЕНИЙ

(Самарский национальный исследовательский университет
имени академика С.П. Королёва)

Введение. 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 отсутствии 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 *DataMining*'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 информа-

ционные 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 $X_1,...,X_n$: $\vec{x} = \langle X_1,..., X_n \rangle$.

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:

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

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

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

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

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

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

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

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 $\vec{x}_1 = \lceil X_1 \rceil$, $\vec{x}_2 = \lceil X_1 \_ X_2 \rceil$ относящиеся к классу *nS* и три $\vec{x}_3 = \lceil X_3 \rceil, \vec{x}_4 = \lceil X_2 \_ X_3 \rceil, \vec{x}_5 = \lceil X_2 \_ X_3, X_2 \_ X_3 ! \rceil$ к классу *S*. Тогда таблица векторов для выборки будет иметь вид:

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

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 (Таблица 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

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:

$$p(nS \mid X_1, X_2, X_3) = p(nS)\,p(X_1 \mid nS)\,p(X_2 \mid nS)\,p(X_3 \mid nS) \quad (5)$$

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

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

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 построений ассоциаций слов.

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: $\vec{x}_1 = \lceil X_1, X_2, X_5 \rceil$, $\vec{x}_2 = \lceil X_2, X_3 \rceil, \vec{x}_3 = \lceil X_1, X_3 \rceil, \vec{x}_4 = \lceil X_1, X_2 \rceil, \vec{x}_5 = \lceil X_1, X_2, X_3 \rceil$

nS: $\vec{x}_6 = \lceil X_2, X_4 \rceil, \vec{x}_7 = \lceil X_1, X_2, X_4 \rceil, \vec{x}_8 = \lceil X_2, X_3 \rceil, \vec{x}_9 = \lceil X_1, X_2, X_3, X_5 \rceil$

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 $\lceil X_1 \rceil, \lceil X_2 \rceil, \lceil X_3 \rceil, \lceil X_5 \rceil, \lceil X_1, X_2 \rceil, \lceil X_1, X_3 \rceil, \lceil X_2, X_3 \rceil$, тогда как множество *nS* состоит из элементов $\lceil X_1 \rceil, \lceil X_2 \rceil, \lceil X_3 \rceil, \lceil X_4 \rceil, \lceil X_5 \rceil, \lceil X_1, X_2 \rceil, \lceil X_2, X_3 \rceil, \lceil X_2, X_4 \rceil$.

Составим таблицу векторов для классов *nS* и *S*. Для этого из каждой определенной последовательности для своего соответствующего класса, определим количество вхождений в соответствующее сообщение. Тогда, таблица векторов для класса *S* будет иметь следующий вид:

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 векторов для класса *nS* имеет следующий вид:

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

Далее сформируем таблицу вхождения слов в определенный класс:

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				

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:

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

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

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

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

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

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

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

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

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 Таблице 6.

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

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

Выводы. 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.

Литература

1. Portio Research оценивает мировой рынок мобильных сообщений в 2011 году [Электронный ресурс]. – 2011. – Режим доступа: http://www.mforum.ru/news/article/099200.htm

2. Worldwide A2P SMS Markets 2014-2017 [Электронный ресурс]. – 2014. –Режим доступа: http://www.strikeiron.com/wp-content/uploads/2014/12/whitepaper-sms-2014-2017-portio-research.pdf

3. Cloudmark наблюдает рост потоков SMS-спама [Электронный ресурс]. – 2013. – Режим доступа: https://securelist.ru/blog/novosti/3684/cloudmark-nablyudaet-rost-potokov-sms-spama/

4. Machine learning methods for spam e-mail classification [Электронный ресурс]. – 2011. – Режим доступа: http://airccse.org/journal/jcsit/0211ijcsit12.pdf

5. *S. Weiss, C. Apte* Maximizing text-mining performance // IEEE Intelligent Systems. - 1999. - 63–69 с.

6. Наивный байесовский классификатор [Электронный ресурс]. – 2015. – Режим доступа: https://ru.wikipedia.org/wiki/Наивный_байесовский _классификатор.

7. SMS Spam Collection v. 1 [Электронный ресурс]. – 2011. – Режим доступа: http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/

А.А. Буцких, Е.А. Розылина, О.А. Синкина

ОБНАРУЖЕНИЕ УГРОЗ И ПЕРЕСТРОЙКА СЕТИ В СИСТЕМЕ БЕЗОПАСНОСТИ «УМНЫЙ ДОМ»

(Поволжский государственный университет телекоммуникаций и информатики)

В настоящее время всем давно уже известно про “ Умный дом”, главной задачей которого является обеспечение комфортной жизнью человека посредством использования высокотехнологических средств [1]. Принцип работы системы “Умный дом” заключается в автоматизации всего, что окружает жильца. Для обеспечения комфортной и безопасной жизни трудятся многочисленное количество датчиков, которые отправляют данные в центр управления, где сведения проходят обработку, после чего происходит корректировка устройств в доме. Из-за постоянного потока данных внутренний канал связи “забивается”, что приводит к падению скорости передачи данных. Падение скорости является критичным показателем для медиа информации, в частности видео. Главным источником видеоинформации является подсистема видеонаблюдения, которая относится к системе безопасности.

В работе предлагаются алгоритмы обработки и анализа видеоизображения, и перестройка сети в зависимости от угрозы безопасности человека и его собственности.

Видеонаблюдение используется в моменты, когда в доме отсутствуют люди, поэтому в кадре видеонаблюдения отсутствует движение. Кадр, в котором отсутствует изменение, будем считаться нулевым K0 (эталонный кадр). При видеосъемке получаем новые кадры Kn, которые сравниваются с эталонным. Сравнение происходит по следующему алгоритму. Кадры Kn и K0 преобразуются в числовые матрицы, размеры которой соответствуют разрешению видеоизображения. Каждый элемент матрицы соответствует уровню спектральной яркости пикселя. Далее идёт сравнительный анализ матриц, где сравниваются элементы матрицы с одинаковыми координатами. Далее происходит поэлементное вычитание значений, при этом результат берётся по модулю. Получается третья матрица, которая является матрицей разности спектральных уровней яросты. Все элементы матрицы суммируются. Полученное значение является параметром, который характеризует изменения в кадре S(t) (где t – является временем получения кадра). При появлении в кадре движущегося объекта происходит значительное изменение уровней яркости отдельно взятых пик-

селей, при этом параметр S(t) значительно возрастёт. Параметры S(t) и t отправляются в охранную организацию (только параметры, а не весь видеоряд), где на монитор выводится график зависимости полученных данных. Параметр S(t) не может быть равным нулю, т.к. показатели изменения уровней яркости всего изображения постоянно меняются. Это связано в основном с освещением в помещении, которое может быть естественным или искусственным и меняться с течением времени под действием. Поэтому график S(t) будет иметь скачкообразную форму. Чтобы не тревожить сотрудника охраны при каждом изменении в кадре, выставляется пороговое значение для S(t), выше которого включается предупредительный сигнал. Возможен вариант с резким изменением спектрального уровня яркости по всему кадру (например: при автоматическом включении света в помещении), что приводит к резкому повышению S(t) и ложному оповещению охраны. Для предотвращения подобной ситуации, происходит предварительный анализ изображения, при котором назначается новый эталонный кадр K0.

Если посторонний объект появился в кадре, и произошло оповещение об этом, то сотрудники охраны посылают запрос на получение видеоизображения на сервер системы безопасности. В случае недостатка ширины канала для обеспечения передачи без задержки, снижения качества, потери пакетов видео данных применяется приоритезация передачи пакетов.

Приоритезация передачи пакетов в коммутаторе состоит из следующих этапов:

1) прием трафика с любого порта;
2) маркировка пакетов трафика по приоритету передачи;
3) размещение трафика в приоритетные очереди на выходном порту.

Для реализации приоритезации передачи пакетов используется алгоритм Strict Priority Queuing (SPQ). SPQ сначала передаются пакеты из очереди, имеющей максимальный приоритет, и только когда она полностью освободится, коммутатор начнет передачу данных из следующей по приоритету. Данные системы безопасности приобретают первичный приоритет. При пакетной передаче информации пакеты систем безопасности маркируются (Q7) и проходят в первую очередь. Остальные пакеты от сторонних систем отбрасываются и отправляются буферную память, где будут ожидать своей отправки в канал связи. Количество буферов соответствует количеству очередей, которые поддерживает коммутатор (не более восьми). Возможен случай, когда пропускной способности канала достаточно для передачи видео данных в требуемом качестве и остаётся некоторый запас, используют смешанные алгоритмы. SPQ для очередей с наивысшим приоритетом (Q7) и обслуживают на основе алгоритма SPQ, а для всех остальных (Q6-Q0) применяют вариант и Weighted Round Robin (WRR), в котором используются специальные взвешенные процедуры для отправки пакетов. Каждой очереди (исключая Q7) выделяется определенный лимит для передачи: чем выше приоритет очереди, тем больше пакетов из нее передается, но в любом случае будут опрошены все очереди в порядке снижения приоритета: после истечения выделенного периода обслуживания одной очере-