УДК 004.896

Исследование способов фильтрации спама в почтовых сообщениях с помощью нейронных сетей

Фам Хю Лок
Московский физико-технический институт (государственный университет)
Центр информационных технологий и систем органов исполнительной власти

Спам или почтовый спам (англ. spam) – это письма, которые приходят к вам на почту от неизвестных вам людей или компаний, которым вы не давали на это свое разрешение. Проблема фильтрации спама является актуальной, так как наносит прямой вред пользователям. В работе применяется нейронная сеть для распознавания спам-сообщений. Решение основано на одном из самых точных, эффективных, популярных в настоящее время методов классификации текстов – контент-фильтрации [1]. Данный метод осуществляет поиск ключевых слов в текстовом сообщении электронной почты и использует алгоритм, позволяющий классифицировать письмо как спам или не-спам (1 или 0). Алгоритм способен определить в каком порядке следуют ключевые слова и словосочетаниях в заранее составленном списке и где они появляются в почтовом сообщении. Таким образом, проблема фильтрации спам-сообщений может быть разбита на простые классификации и большая часть времени отводится на обучение нейронных сетей, таких как обратное распространение ошибки [2].

Решаемая задача заключается в следующем: даны множество объектов $X = \{x_1, \dots, x_N\}$ и множество допустимых ответов $Y = \{y_1, \dots, y_N\}$. Надо определить, к какому классу из $Y = \{0,1\}$ принадлежит объект из $X = \{x_1, \dots, x_N\}$, то есть объект принадлежит к спаму или нет.

Реализована классификация спам-сообщений и было проведено сравнение производительности различных сетевых конфигураций MLP (таблицы 1 и 3). Нейронная сеть состоит из входного слоя, скрытых слоёв и выходного слоя. Количество элементов в входном слое зависит от количества входных атрибутов.Количество слоёв в скрытых слоях определяет, например, в сетевой конфигурации (20-10-20-8) имеет 4 слоя: первый – 20 нейронов, второй – 10 нейронов, третий – 20 нейронов, четвертый – 8 нейронов. Также программы были оптимизированы для параметров импульса и обучения (таблицы 2 и 4). Параметры процесса обучения следующие: коэффициент скорости обучения (Learning Rate), импульс (Momentum). В таблицах ниже показаны значения данных параметров, используемые в процессе обучения и теста. В качестве алгоритма обучения, был использован алгоритм обратного распространения ошибки.

Используемые данные в работе были получены из «Enron-Spam datasets - V. Metsis, I. Androutsopoulos and G. Paliouras », анализированные учёными V. Metsis, I. Androutsopoulosand и G. Paliouras [3].

Выходные данные: Последний столбец "spambase.txt" обозначается (1), если электронное сообщение было спамом (нежелательным электронным сообщением) и обозначается (0), если сообщение не было спамом.

Входные атрибуты: Большинство атрибутов является конкретным словом или символом, часто встречающимся в электронной почте. Суммарная длина атрибутов (55-57) является длиной последовательности.

1. Результаты моделирования для набора данных «non-stemming» (рис. 1).

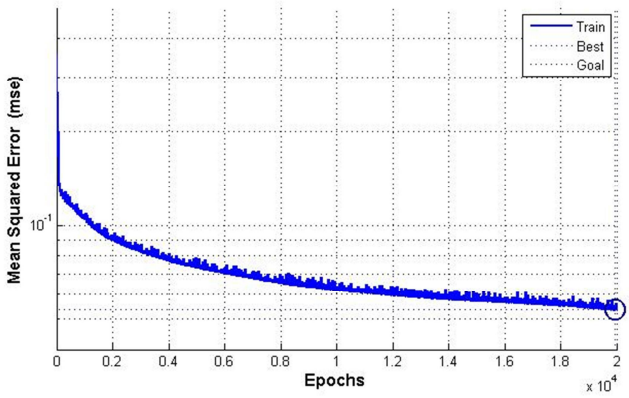

Рис. 1. График ошибки зависит от числа шагов обучения

2. Результаты моделирования для набора данных «Stemming» (рис. 2.)

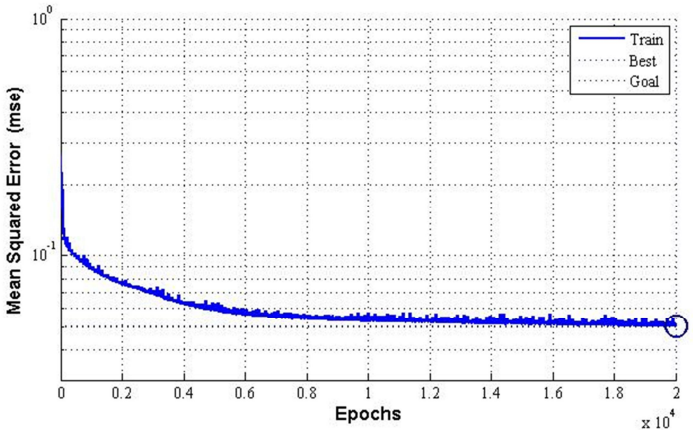

Рис. 2. График ошибки зависит от числа шагов обучения

Результаты фильтрации спама получены из примера электронной почты, в которой среднее число спам-сообщений составляет около 85% (таблица 4).

Скрытый слой	Learning Rate	Momentum	Средние проценты (набор данных обучения)		
			норм %	спам %	сумма %
20 - 10 -10 - 8	0.1	0.8	93.3	84.8	89.05
20 -10 - 10 - 5	0.1	0.8	93.2	84.1	88.65
20 - 10 -10 - 5	0.1	0.95	93.5	86.3	89.9
15 - 10 -10 - 5	0.1	0.95	93	88	90.5
15 - 10 -10 - 8	0.1	0.85	90.6	82.8	86.7
10 - 10 - 5	0.1	0.85	93.7	84.8	89.25
15 - 15 - 5	0.1	0.85	93.2	87.1	90.15

Таблица 1. Результаты моделирования для набора данных обучения «non-stemming»

Скрытый слой	Learning Rate	Momentum	Средние проценты (набор данных теста)		
			норм %	спам %	сумма %
20 - 10 -10 - 8	0.1	0.8	86	78	82
20 -10 - 10 -5	0.1	0.8	88	77	82.5
20 - 10 -10 - 5	0.1	0.95	88	73	80.5

15 - 10 -10 - 5	0.1	0.95	86	78	82
15 - 10 -10 - 8	0.1	0.85	86	77	81.5
10 - 10 - 5	0.1	0.85	82	72	77
15 - 15 - 5	0.1	0.85	89	75	82

Таблица 2. Результаты моделирования для набора данных теста «non-stemming»

Скрытый слой	Learning Rate	Momentum	Средние проценты (набор данных обучения)		
			Норм %	Спам %	Сумма %
20 - 10 -10 - 8	0.1	0.8	93.5	85.7	89.6
20 -10 - 10 -5	0.1	0.8	94.3	84.1	89.2
20 - 10 -10 - 5	0.1	0.95	94.6	85.1	89.85
15 - 10 -10 - 5	0.1	0.95	94.7	87	90.85
15 - 10 -10 - 8	0.1	0.85	95.4	88.5	91.95
10 - 10 - 5	0.1	0.85	94	83.8	88.9
15 - 15 - 5	0.1	0.85	95	88.3	91.65

Таблица 3. Результаты моделирования для набора данных обучения «Stemming»

Скрытый слой	Learning Rate	Momentum	Средние проценты (набор данных теста)		
			Норм %	Спам %	Сумма %
20 - 10 -10 - 8	0.1	0.8	88	81	84.5
20 -10 - 10 -5	0.1	0.8	87	76	81.5
20 - 10 -10 - 5	0.1	0.95	90	77	83.5
15 - 10 -10 - 5	0.1	0.95	90	75	82.5
15 - 10 -10 - 8	0.1	0.85	88	78	83
10 - 10 - 5	0.1	0.85	88	79	83.5
15 - 15 - 5	0.1	0.85	86	80	83

Таблица 4. Результаты моделирования для набора данных теста «Stemming»

Литература

1. *Pfleeger S. L., Bloom G.* Canning Spam: Proposed Solutions to Unwanted Email. IEEE Security & Privacy, 2005. С. 40-47.
2. *Wu C. T., Cheng K. T.* Using visual features for anti-spam filtering. IEEE International Conference on Image Processing, 2005. 12 с.
3. *Metsis V., Androutsopoulos I., Paliouras G.* Spam Filtering with Naive Bayes - Which Naive Bayes?. Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006), Mountain View, CA, USA, 2006. 9 с.