Санкт-Петербургский Государственный Университет

Математико-механический факультет

Кафедра системного программирования

Курсовая работа на тему:

«Оценка сайта на наличие нежелательного контента»

Кривых Алексей, 361гр.

Научный руководитель: Тарасов Александр

Рецензент: Баклановский Максим Викторович

Санкт-Петербург

2011г.

Оглавление

1. Введение

С развитием новых технологий тема искусственного интеллекта перестала быть уделом научных фантастов. За последние пару десятков лет машинное обучение шагнуло далеко вперед. Этот раздел искусственного интеллекта нашел применение в таких областях, как: распознавание речи, распознавание образов, распознавание жестов, категоризация документов, итд. Также машинное обучение широко применяется в сети Интернет, например, для определения тематики интернет ресурса или для определения спама. В данных задачах используется раздел машинного обучения – классификация.

Задача классификации: Имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется обучающей выборкой. Классовая принадлежность остальных объектов не известна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

В данной работе используется один из алгоритмов классификации – Байесовский классификатор, основанный на теореме Байеса. Данный алгоритм используется в большинстве современных спам фильтрах, и показывает хорошие результаты при обнаружении спама (отсекает 95-97% спама при достаточно большой обучающей выборке). Также этот метод имеет ряд преимуществ: прост в реализации, эффективен (порой эффективнее более сложных алгоритмов).

2. Постановка задачи

Цель данной работы:

- изучить алгоритм Байесовской классификации
- реализовать методы оценки сайта на наличие «нежелательного контента»
    - реализовать оценку теста сайта
    - реализовать оценку метаинформации сайта
    - предложить метод вычисления суммарной оценки для домена

3. Классификатор Байеса

Байесовский классификатор - это классификатор, использующий теорему Байеса для определения вероятности принадлежности наблюдения (элемента выборки) к одному из классов C при условии того, что зависимые переменные принимают заданные значения: P(C|F1 ,...,Fn). То есть, если на основе значений переменных можно однозначно определить, какому классу относится наблюдение, байесовский классификатор сообщит, что вероятность принадлежности к этому классу равна 1. В промежуточных же случаях, когда наблюдение может с разной вероятностью принадлежать к различным классам, результатом работы классификатора будет вектор, компоненты которого являются вероятностями принадлежности к тому или иному классу.

Можно видеть, что идеальный байесовский классификатор в каком-то смысле является оптимальным. Его результат не может быть улучшен, т.к. во всех случаях, когда возможен однозначный ответ, он его даст - а в тех случаях, когда ответ неоднозначен, результат количественно характеризует меру этой неоднозначности. Вместе с тем, в оптимальности кроется и основной недостаток идеального байесовского классификатора: для его построения требуется выборка, содержащая все возможные комбинации переменных - а размер такой выборки экспоненциально растет с ростом числа переменных (т.н. "проклятие размерности").

Для преодоления описанной выше проблемы на практике используют т.н. наивный байесовский классификатор - классификатор, построенный на основе предположения о независимости переменных. Тогда по теореме Байеса:

$$\mathrm{P}(\mathrm{C}|\mathrm{F}_1, \dots, \mathrm{F}_\mathrm{n}) = \frac{\mathrm{P}(\mathrm{F}_1, \dots, \mathrm{F}_\mathrm{n}|\mathrm{C})\mathrm{P}(\mathrm{C})}{\mathrm{P}(\mathrm{F}_1, \dots, \mathrm{F}_\mathrm{n})}$$

Учитывая условие независимости событий $\mathrm{F}_1, \dots, \mathrm{F}_\mathrm{n}$ получим:

$$\mathrm{P}(\mathrm{C}|\mathrm{F}_1, \dots, \mathrm{F}_\mathrm{n}) = \frac{\mathrm{P}(\mathrm{F}_1|\mathrm{C}) \times \mathrm{P}(\mathrm{F}_2|\mathrm{C}) \times \dots \times \mathrm{P}(\mathrm{F}_\mathrm{n}|\mathrm{C}) \times \mathrm{P}(\mathrm{C})}{\mathrm{P}(\mathrm{F}_1, \dots, \mathrm{F}_\mathrm{n})}$$

В случае классификации сайтов: события $\mathrm{F}_1, \dots, \mathrm{F}_\mathrm{n}$ - это слова, множество классов C состоит из двух классов: «порно» сайты и «не порно» сайты. Конечно, предположение о независимости слов в тексте для естественных языков не верно. Но данное предположение существенно упрощает задачу, а погрешность в вычислениях получается допустимой для задачи классификации.

3.1 Усовершенствования алгоритма

«Нейтральные» слова

Некоторые слова, например, "the", "a", "some", или "is" (в английском языке), или их эквиваленты на других языках, могут быть проигнорированы, так как невозможно определить их принадлежность к тому или иному классу в задаче классификации.

Простое решение состоит в том, чтобы игнорировать слова, для которых $P(W|C) \approx 0.5$.

Проблема редких слов

Данная проблема возникает, если слово ни разу не встречалось на этапе обучения, или встречалось недостаточное количество раз, чтобы делать выводы о свойствах такого слова.

Здесь, как и выше предлагается игнорировать такие слова.

4. Описание работы фильтра

Условно работу фильтра можно разделить на 5 частей: обучение, оценка текста, оценка метаинформации, итоговая оценка сайта, дообучение.

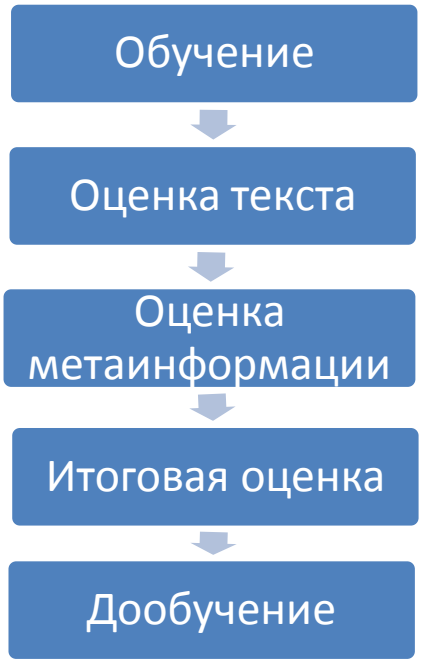

Рисунок 1 Этапы работы фильтра

4.1 Обучение

Обучение играет очень важную роль в работе фильтра, обучающая выборка должна быть достаточно полной для корректной работы фильтра.

Для обучения использовалась выборка из 40 сайтов содержащих нежелательный контент и 40 сайтов не содержащих нежелательного контента.

В результате обучения были получены 2 словаря с весовыми оценками, каждый из словарей содержит несколько десятков тысяч слов.

4.2 Оценка текста и оценка метаинформации

Оценка текста и метаинформации производится по формулам, представленным в разделе «Байесовский классификатор». В итоге получаем две условные вероятности:
p(P|T) – вероятность, что сайт – порно сайт, при условии, что в нем содержатся слова T

p(P|M) – вероятность, что сайт – порно сайт, при условии, что в мета тегах содержатся слова M. При отсутствии мета тегов считаем p(P|M) = 0.5.

4.3 Итоговая оценка

На основании двух вероятность из предыдущего пункта считаем итоговую оценку по формуле Байеса: $p = \frac{p(P|T)p(P|M)}{P(T,M)}$. Каждый фильтр имеет вес, определяющий его влияние на итоговую оценку. Вес фильтра это коэффициент, с которым вероятность, возвращенная этим фильтром, входит в итоговую формулу.

При принятии решения о сайте, используем пороговое значение 0.7, то есть, если итоговая оценка выше или равна этому значению, считаем, что сайт – порно.

4.4 Дообучения

На основании решения фильтра о свойствах сайта относим сайт к тому или иному классу и проводим дообучение. То есть добавляем адрес сайта в базу, и обновляем словари с учетом новых данных.

5. Тестирование

Для тестирования была составлена выборка из 41 хороших сайтов и 58 плохих.

Результаты тестирования:

- На хороших сайтах: 1 ошибка из 41
- На плохих сайтах: 2 ошибки из 58

6. Результат

В процессе данной работы были реализованы фильтры оценки сайта на наличие «нежелательного контента». Так же был изучен и реализован один из самых популярных методов классификации – Байесовский классификатор. По окончанию работы было проведено тестирование фильтра.

7. Список литературы

[1] Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. http://www.machinelearning.ru

[2]Лекция Сергея Николенко (Академический Университет) «Байесовские классификаторы» http://logic.pdmi.ras.ru/~sergey/teaching/mlaptu11/03-classifiers.pdf

[3] Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения)