МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего профессионального образования

«Нижегородский государственный университет им. Н.И. Лобачевского»

Институт Информационных технологий, математики и механики

**ПРОЕКТНАЯ РАБОТА**

Тема:

«Классификация RSS лент»

**Выполнили:**

Костин С.С.

Ганин А.

Нижний Новгород  
2015

**Содержание**

[Введение 3](#_Toc436565767)

[1.Описание данных 4](#_Toc436565768)

[1.1. Нехватка данных 4](#_Toc436565774)

[2.Обработка исходных данных (обучение) 5](#_Toc436565775)

[3.1 Байесов классификатор 6](#_Toc436565780)

[3.1.1 Общие принципы 6](#_Toc436565781)

[3.1.2 Анализ “критериев качества” 7](#_Toc436565782)

[3.2 Тестирование (MLP) 8](#_Toc436565783)

[4.1. Общие принципы 8](#_Toc436565784)

[4.2. Анализ “критериев качества” 9](#_Toc436565785)

[5. Сравнение алгоритмов 10](#_Toc436565786)

[6. Литература 12](#_Toc436565787)

[7. Приложение 13](#_Toc436565788)

# Введение

Каждый из нас когда-нибудь в своей жизни смотрел или читал новости в газетах, по телевизору, или, как популярно сейчас, в Интернете. Следить за новостями последним способом наиболее удобно, так как можно просто “отфильтровать” новости по интересной для вас тематике. Они могут быть абсолютно различными по содержанию: о медицине, о спорте, о красоте и уходе, о политике, экономике и т.д.

При выполнении данной проектной работы использовался сайт [http://subscribe.ru](https://vk.com/away.php?to=http%3A%2F%2Fsubscribe.ru), на котором можно наблюдать как старые, так и самые свежие новости представленные RSS-лентами.

RSS (англ. Rich Site Summary - обогащённая сводка сайта) - семейство XML-форматов, предназначенных для описания лент новостей, анонсов статей, изменений в блогах и т.п. Информация из различных источников, представленная в формате RSS, может быть собрана, обработана и представлена пользователю в удобном для него виде специальными программами-агрегаторами или онлайн-сервисами, такими как Google Reader (закрыт), Яндекс.Лента (в дальнейшем переименован в «Подписки» и включен в сервис Яндекс.Почта), BlinkFeed и другими.

Обычно с помощью RSS 2.0 даётся краткое описание новой информации, появившейся на сайте, и ссылка на её полную версию. Интернет-ресурс в формате RSS называется RSS-каналом, RSS-лентой или RSS-фидом.

В каждой RSS-ленте каждая новость принадлежит к одной из категорий, приведенных выше.

# 1.**Описание данных**

В работе в качестве входных данных являются RSS-ленты, “скачанные” с сайта [http://subscribe.ru](https://vk.com/away.php?to=http%3A%2F%2Fsubscribe.ru), а так же категории, к которым принадлежат новости из этих лент. Всего было выгружено в базу данных 2051 RSS-лент. В каждой ленте от 10 до 50 новостей.

## 1.1. Нехватка данных

Входные данные, используемые для реализации системы классификации RSS-лент, не могут иметь пропусков (недостатка) данных. Поскольку у каждой ленты имеется одна и только одна категория. Поэтому проблем с нехваткой данных у нас нет.

# 2.Обработка исходных данных (обучение)

В начале, на RSS-лентах (от 10 до 95 процентов, загруженных в базу данных) будем производить обучение, затем на оставшихся мы будем производить тестирование. Обучение происходит следующим образом. Для каждой RSS-ленты составляется вектор уникальных слов, встречающихся в новостях из данной ленты. Также составляется массив содержащий набор категорий, которым принадлежит та или иная лента. В итоге получаем ассоциативную связь вида: “category”=>”vector\_of\_unique\_words”. Иллюстрацию можно видеть на следующем рисунке:

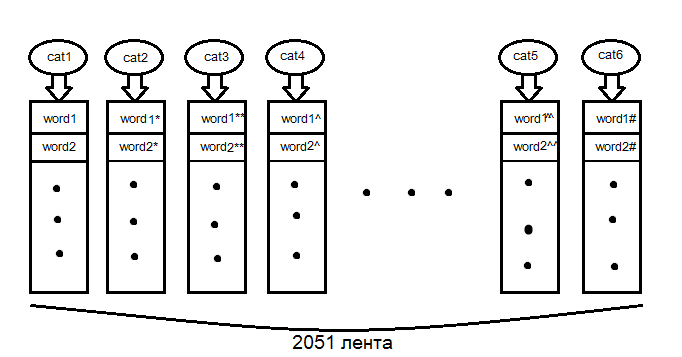


Рис1. Обучение классификатора.

Интуитивно понятно, что, чем больше RSS-лент мы возьмем в качестве обучающей выборки, тем выше будет показатель качества классификатора на оставшейся (тестовой) выборке. Тестирование будем производить с помощью двух алгоритмов, описанных в библиотеке, находящейся в файле machine\_learning.min.js : Байесов классификатор и MLP-алгоритм.

В качестве “критерия качества” при тестировании у нас будет являться соотношение: .

# 3.1 Тестирование (Байесов классификатор)

## 3.1.1 Общие принципы

Рассматривается задача классификации, множество классов обозначим . Для решения поставленной задачи необходимо минимизировать средний риск:

Функция потерь является симметричным штрафом (то есть, может принимать два значения: 0 и 1) и определяется формулой:

Подынтегральная функция в является вероятностью ошибки при заданном , значит можно представить в виде:

Отсюда следует, что регрессионная функция

Функция называется байесовым классификатором. При этом средний риск байесова классификатора называется байесовой ошибкой. Байесов классификатор играет в задаче классификации роль, аналогичную той, которую играет регрессионная функция в задаче восстановления регрессии.

— априорная вероятность появления объекта из класса y.

— апостериорная вероятность появления объекта из класса y.

Чтобы построить байесов классификатор, мы должны знать или оценить . Если классы равновероятны, то есть , то

где плотность называется правдоподобием.

## 3.1.2 Анализ “критериев качества”

В следующих таблицах можно наблюдать показатели качества при разном соотношении: обучающая/тестируемая выборка.

|  |  |  |  |
| --- | --- | --- | --- |
| Соотношение  обуч./тест. | 10%/90% | 60%/40% | 70%/30% |
| Результаты экспериментов. |  |  |  |

Таблица 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Соотношение  обуч./тест. | 80%/20% | 90%/10% | 95%/5% |
| Результаты экспериментов. |  |  |  |

Таблица 2.

Из результатов тестирования можно видеть, что при увеличении объема тестируемой выборки, количество верно “предсказанных” лент увеличивается, хоть и скачками. Это может происходить потому, что объём тестируемой выборки уменьшается. Если сохранять объем тестируемой выборки постоянным, будет происходить строгое возрастание.

# 3.2 Тестирование (MLP)

## 3.2.1 Общие принципы

**Многослойный** персептрон (MLP) – нейронная сеть прямого распространения сигнала (без обратных связей), в которой входной сигнал преобразуется в выходной, проходя последовательно через несколько *слоев*.

Первый из таких *слоев* называют входным, последний - выходным. Эти *слои* содержат так называемые вырожденные нейроны и иногда в количестве *слоев* не учитываются. Кроме входного и выходного *слоев*, в многослойном *персептроне* есть один или несколько промежуточных *слоев*, которые называют скрытыми.

В этой модели *персептрона* должен быть хотя бы один скрытый *слой*. Присутствие нескольких таких *слоев* оправдано лишь в случае использования нелинейных *функций активации*.

Пример двухслойного *персептрона* представлен на рисунке 2.

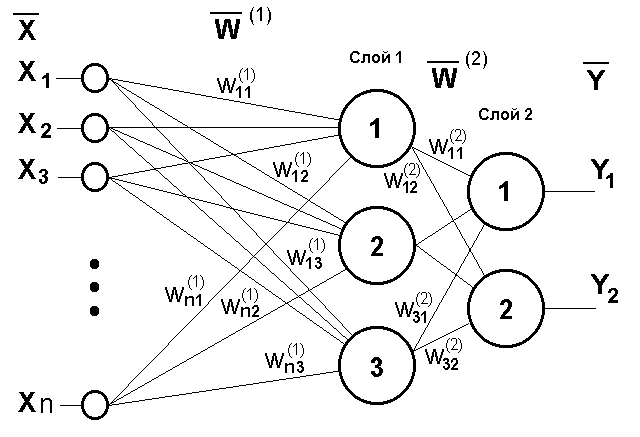


Рисунок 2.

Сеть, изображенная на рисунке, имеет n входов. На них поступают сигналы, идущие далее по *синапсам* на 3 нейрона, которые образуют первый *слой*. Выходные сигналы первого *слоя* передаются двум нейронам второго *слоя*. Последние, в свою очередь, выдают два выходных сигнала.

После обучения многослойного персептрона мы подаем ему на вход вектора тестируемой выборки и получаем на выходе категорию, к которой “возможно” принадлежит этот вектор.

## 3.1.2 Анализ “критериев качества”

В следующих таблицах можно наблюдать показатели качества при разном соотношении: обучающая/тестируемая выборка.

|  |  |  |  |
| --- | --- | --- | --- |
| Соотношение  обуч./тест. | 10%/90% | 20%/80% | 30%/70% |
|  |  |  |  |

Таблица 3.

|  |  |  |  |
| --- | --- | --- | --- |
| Соотношение  обуч./тест. | 40%/60% | 50%/50% | 70%/30% |
|  |  |  |  |

Таблица 4.

В качестве параметров для MLP была использована нейронная сеть, состоящая из трёх скрытых слоёв, в каждом из которых по 4 нейрона (4x4x4).Из результатов тестирования можно видеть, при разном соотношении обучающая/тестируемая выборка точность разбрасывается в относительно широком диапазоне.

# 4. Сравнение алгоритмов

На следующих графиках можно видеть кривые “качества” для обоих методов:

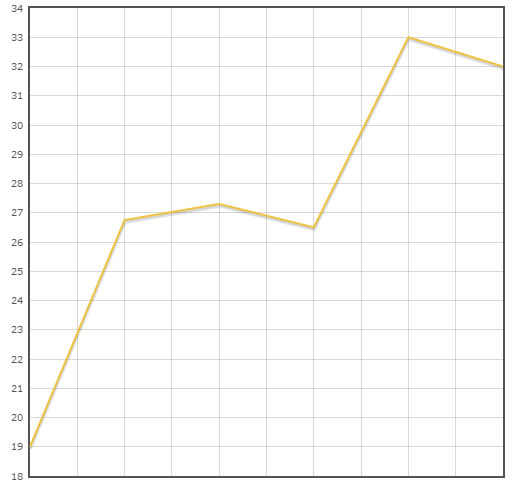


Рисунок 3.

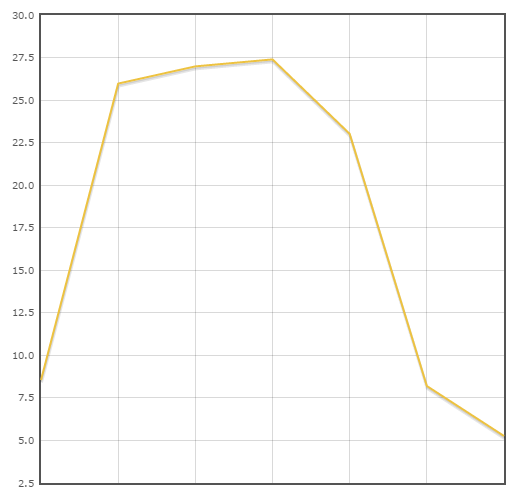


Рисунок 4.

Из графиков видно, что при работе байесова классификатора точность возрастает (не строго), с увеличением объема обучаемой выборки. При работе же MLP-алгоритма “пик” точности наблюдается при небольшом объеме обучаемой выборки. Можно сделать вывод о том, что при больших затратах времени на обучение, лучше использовать MLP-алгоритм, так как он выдает при малом размере обучающей выборки выдаёт наилучший критерий качества. Когда время обучения игнорируется или просто не имеет значения, целесообразнее использовать байесов классификатор, так как он лучше “покажет себя” при большем объеме обучающей выборки (лучше обучить на большом объеме данных).

**5. Литература**

1. Золотых Н.Ю. «Курс лекций по машинному обучению» 2015 г.
2. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001.
3. Berry, Michael J. A. “Data mining techniques: for marketing, sales, and customer relationship management “ Michael J.A. Berry, Gordon Linoff. – 2nd ed.
4. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999
5. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. М.: Наука, 1974
6. Murthy S. (1998). Automatic construction of decision trees from data: A multidisciplinary survey. Data Mining and Knowledge Discovery
7. Mitchell T. Machine learning. McGraw Hill,1997

## 6. Приложение