информации о цвете / Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. – 2011. – С. 70-79

3. Kulchandani J. S., Dangarwala K. J. Moving object detection: Review of recent research trends / 2015 International Conference on Pervasive Computing (ICPC). – 2015. – pp. 1-5

4. Mitrokhin, A. Event-based moving object detection and tracking / 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). – 2018. – pp. 1-9

5. Shijila B., Tom A. J., George S. N. Simultaneous denoising and moving object detection using low rank approximation / Future Generation Computer Systems. – 2019. – pp. 198-210

Е.Г. Плешаков, Л.С. Зеленко, Д.С.Оплачко

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ «НЕЙРОСЕТЕВОЙ КЛАССИФИКАТОР ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ»

(Самарский университет)

Введение

Текстовое представление информации является наиболее часто используемым, особенно в сети интернет. Книги, статьи, новости, твиты, посты, – всё это различные формы текста. На множестве серверов хранятся терабайты, а порой и петабайты информации в текстовом виде. И с каждым днём количество текстовой информации продолжает расти. В таком большом объёме информации, которая чаще всего не структурирована, очень трудно найти текст по конкретной теме, что серьёзно обесценивает эту информацию.

Наиболее перспективным подходом при поиске текстовой информации на данный момент является машинное обучение, когда система на основе небольшой выборки размеченных данных сама создаёт правила для классификации и впоследствии на их основе присваивает категории новым текстам. Наиболее часто в качестве основы для систем классификации используется нейронная сеть [1].

Структура нейросетевого классификатора текстов на естественном языке

Система классификации состоит из двух основных частей: частотный анализатор со словарем и нейросетевой классификатор, схема представлена на рисунке 1. На вход системы поступает текст, на выходе получаем тему (номер класса), которой посвящен этот текст.

Для решения задачи приведения слов к основной форме существует несколько способов: лематизация (все слова в тексте приводятся к нормальной форме (единственное число, именительного падежа)) и стеминг (выделение

основы слов путём отбрасывания приставок и окончаний), причем второй менее качественный, но более быстрый [2].

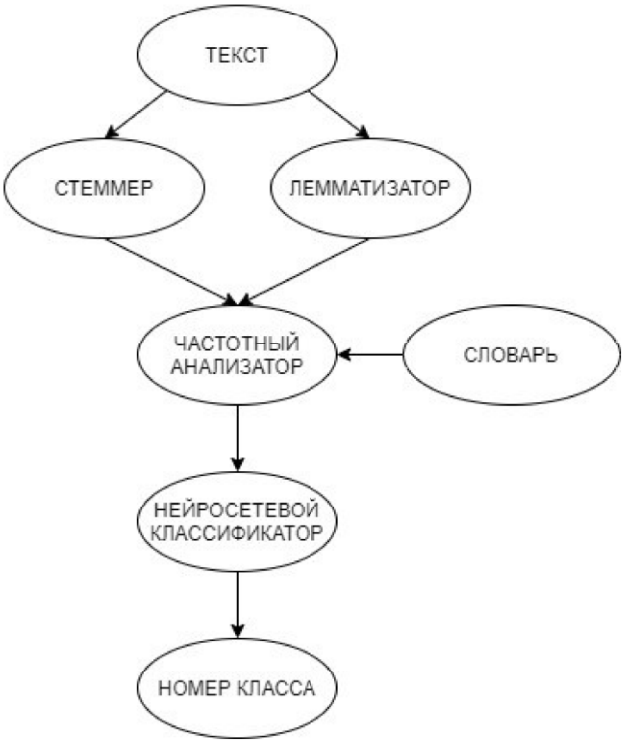

Рис. 1. Схема нейросетевого классификатора

Для уменьшения размерности вектора признаков текста применяется метод нормализованного частотного анализа или TF (Term Frequency) [2], когда значения частоты делятся на общее число слов в тексте, или метод TF-IDF (Inverse Document Frequency), данный метод не выбрасывает часто употребляемые слова из словаря, а уменьшает их вес в вектор-признаке [3]. Итоговая частотная характеристика выглядит как произведение частотной характеристики TF на коэффициент обратной частоты IDF.

Второй частью классификатора является нейронная сеть, которая классифицирует вектор частотных характеристик (точку в пространстве признаков), полученный с помощью частотного анализатора, т.е. разделяет все пространство признаков на определенное количество областей.

В качестве архитектуры нейронной сети была выбрана схема многослойного персептрона. Такая сеть состоит из множества наборов нейронов, называемых слоями. Множество входных узлов называют входным слоем сети, при этом сигнал двигается от слоя к слою в прямом направлении. На рисунке 2 представлен пример многослойного персептрона с двумя скрытыми слоями и с сигмоидами в качестве функции активации

Число входных нейронов равно размеру составленного словаря, а число выходных – количеству классов, число скрытых слоёв и нейронов в них устанавливается при создании сети. Для обучения сети использовался метод обратного распространения ошибки [4].

.

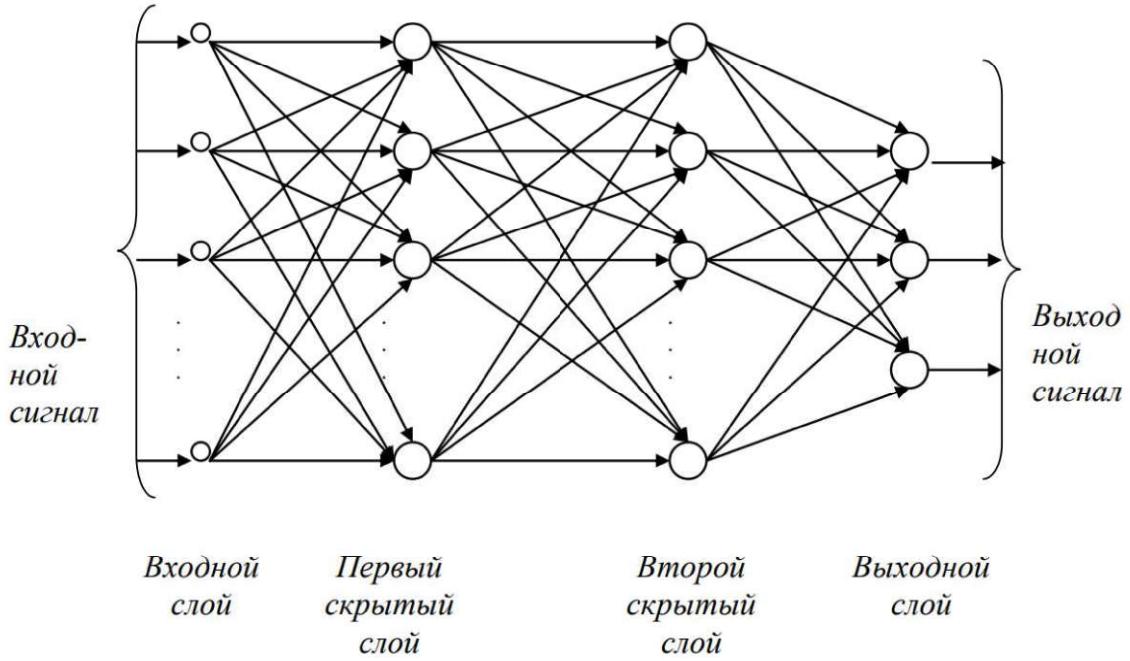

Рис. 2. Схема однослойного персептрона с двумя скрытыми слоями

Проектирование автоматизированной системы для нейросетевой классификации текстов

Для автоматизации решения задачи нейросетевой классификации текстов была спроектирована система, структурная схема которой приведена на рисунке 3.

В состав системы входят следующие подсистемы:

1 Подсистема подготовки текста, которая отвечает за преобразование текста к виду, который может обрабатывать нейронная сеть. Она включает в себя:

– Подсистему чтения текста, которая отвечает за чтение текста из файлов.

– Подсистему предобработки текста, которая нормализует текст при помощи стеммера или лемматизатора.

– Подсистему частотного анализа, которая с помощью частотного словаря и частотного анализатора преобразует текст в вектор частот.

2 Подсистема работы с нейросетью, которая включает:

– Подсистему обучения, которая отвечает за обучение нейросети.

– Подсистему тестирования, которая отвечает за тестирование сети.

3 Подсистема интерпретации, которая преобразует выходной вектор значений нейросети к однозначному ответу.

4 Файловая подсистема, которая отвечает за работу с файлами;

5 Справочная подсистема, которая отвечает за выдачу справочной информации.

Рис. 3. Структурная схема системы

Заключение

Разработанная система была апробирована на обучающей выборке из 1000 текстов, относящимся к 10 различным классам. На основании этого были определены оптимальные значения параметров нейронной сети, при которых процент распознавания находится в диапазоне 75-90%. Для получения таких результатов количество нейронов в первом скрытом слое должно составлять 100, во втором 25, коэффициент обучения должен быть равен 0,01.

Литература

1 Нейроинформатика [Электронный ресурс]: учеб. пособие. Системные требования: Google Chrome – URL: repo.ssau.ru/bitstream/Uchebnye-posobiya/Neiroinformatika-Elektronnyi-resurs-ucheb-posobie-55106/1/Солдатова%20О.П.%20Нейроинформатика.pdf (дата обращения: 02.04.2020).

2 Нейронные сети, генетические алгоритмы и нечеткие системы [Электронный ресурс] / М. Пилиньский, Л. Рутковский, пер.: И.Д. Рудинский, Д. Рутковская . 2-е изд., стер. М.: Горячая линия-Телеком, 2015. 85 с.

3 Википедия TFIDF [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/ TF-IDF (дата обращения: 23.10.2019).

4 Осовский С. Нейронные сети для обработки информации. М.: Горячая линия–Телеком, 2017. 448 с.