Е.В. Старкова, Я.В. Соловьева

ИССЛЕДОВАНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

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

На сегодняшний день в условиях стремительно растущего объема информации и в связи с потребностью в ней ориентироваться все более актуальной становится проблема построения универсального классификатора текстов, имеющего возможность распределения исходного набора текстов по нескольким заранее установленным рубрикам в соответствии с их смысловым содержанием. Использование такого классификатора позволит сократить трудозатраты на поиск нужной информации, представленной электронными текстами.

Различные решения данной задачи находят свое применение в таких областях, как обработка новостей, фильтрация спама, разделение сайтов по тематическим каталогам, классификация библиотечных материалов и т.д. Основными методами решения являются методы машинного обучения (метод Байеса, метод Роше, метод k-ближайших соседей и т.д.), а так же методы, основанные на знаниях (экспертные системы). Одним из перспективных направлений на сегодняшний день считается использование нейронных сетей в качестве основы подобного рода классификатора.

Основным преимуществом нейронных сетей является возможность выявления зависимостей, не поддающихся обнаружению при использовании других методов обработки информации. Нейросетевой подход к анализу текстовой информации обладает достаточным быстродействием и не зависит от языка предметной области, но при этом, в отличие от большинства алгоритмов обработки текстов, реализованных на основе статистического подхода, дает хорошие результаты [1].

В настоящее время примерами классификатора текстов являются такие системы как NNCS (Neural Network Classification & Search), TextAnalyistPro, TextCat, CBTReader, а также проект ДИАЛИНГ, разработанный специалистами факультета лингвистики РГГУ. Однако они имеют ряд недостатков: во-первых, это коммерческие проекты, стоимость которых достаточно высока, а во-вторых, эти проекты рассчитаны на профессионального пользователя, следовательно, только обучение использованию предлагаемых пакетов займет слишком много времени.

Целью данной работы является исследование возможностей нейронных сетей в решении задач классификации текстовых фрагментов в соответствии с их смысловым содержанием, проектирование и реализация нейросетевого классификатора текстов на естественном языке для экспериментальной оценки работы нейронной сети, а также сравнение полученных результатов с результата-

ми других методов классификации. Основой классификатора является нейронная сеть многослойный персептрон (рис. 1).

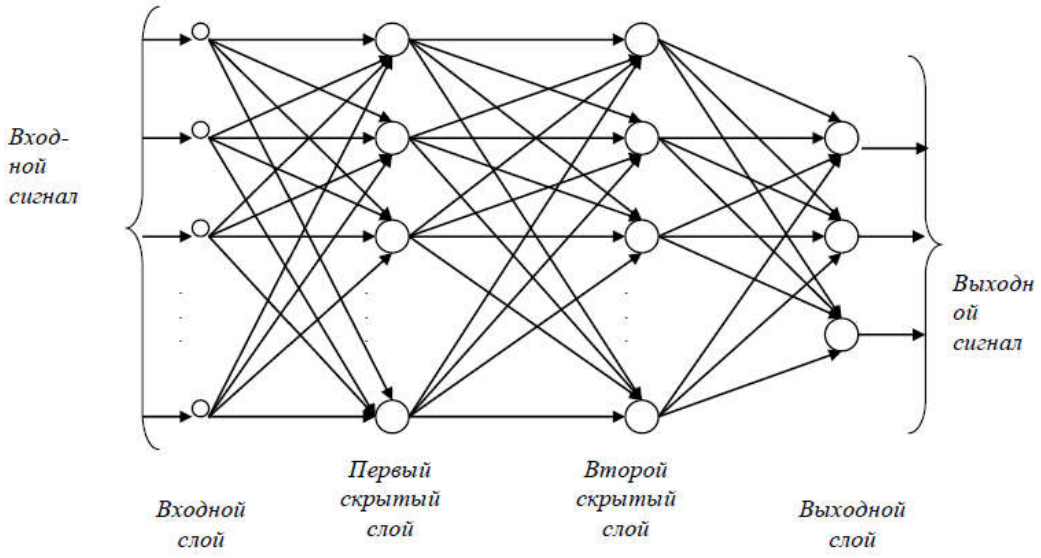

Рис. 1. Многослойный персептрон

Система классификации состоит из двух основных частей: частотный анализатор со словарем и нейросетевой классификатор (рис. 2). На вход системы поступает текст, на выходе получаем тему, которой посвящен этот текст.

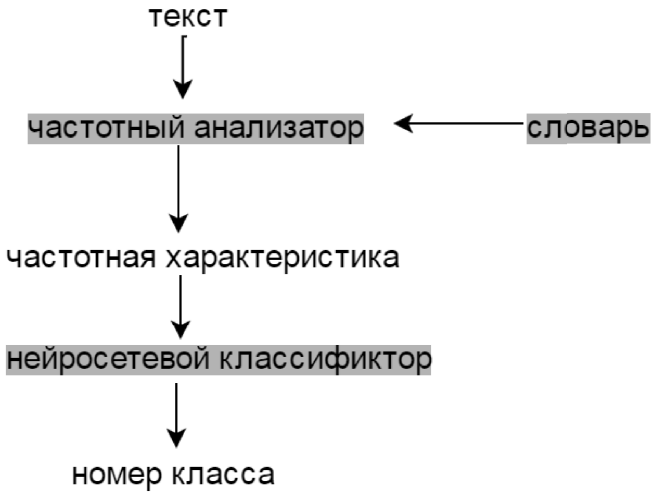

Рис. 2. Общая схема классификатора

Частотный анализатор реализует известный лингвистический метод для обработки текстов на естественных языках - частотный анализ, который показывает распределение повторов слов в тексте. Эта часть системы определяет для каждого слова v_i из словаря V его частоту вхождения f_i в данный текст t (рис. 2). Частотная характеристика - это вектор $f=(f_1,...f_n)\in F$, длина которого равна количеству слов в словаре V, каждая компонента f_i это целое неотрицательное число:

$$f_i = \sum_{i=0}^{k} e(v_i, t_j); \quad e(v_i, t_j) = \begin{cases} 0, v_i \neq t_j \\ 1, v_i = t_j \end{cases}$$

Другими словами, для каждого слова $v_i \in V$ определяется число его вхождений $f_i \geq 0$ в данный текст $t=t_1t_2t_3...t_k$.

Частотную характеристику f можно рассматривать как точку в пространстве признаков F, соответствующую тексту t. Таким образом, на входе имеем текст t и словарь V, на выходе точку в пространстве признаков F.

Вторая часть системы классифицирует вектор частотных характеристик, полученный с помощью частотного анализатора, т.е. разделяет все пространство признаков на определенное количество областей.

Рис.3. Схема частотного анализатора

На вход нейронной сети подается вектор частотной характеристики, на выходе получаем вектор $(y_0...y_m)$. Номер j , для которого выход y_j имеет максимальную активность (т.е. $y_j=\max(y_i)$; i=0...m), соответствует номеру класса входного образца [2].

Для обучения этой нейронной сети был выбран метод обратного распространения ошибки.

Обучение классификатора на множестве учебных текстов можно представить в виде четырех этапов:

1. определение количества классов;
2. составление словаря;
3. частотный анализ множества учебных текстов;
4. обучение нейросетевого классификатора.

Процесс обучения нейронной сети сводится к корректировке весовых коэффициентов ее связей в соответствии с методом обратного распространения ошибки.

Литература

1. Борисов В.В. Нейронные сети и алгоритмы. – М.:Информ, 2007. – 283 с.

2. Головко В.А. Нейронные сети: обучение, организация и применение. – М.: ИПРЖР, 2001. – 256 с.