Титов И.В., Третьяков В.С.

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ АНАЛИЗА ТЕКСТА

IgrikWork@mail.ru, vastwork@do.ustu.ru
Российский Государственный Профессионально – Педагогический Университет
г. Екатеринбург

Наличие большого количества электронных документов, хранимых в форматах без семантической разметки текста документа, обуславливает актуальность решения задач по созданию универсальных алгоритмов, позволяющих проводить автоматизированный анализ подобных документов с целью разметки смыслового содержания отдельных фрагментов текста. Задача заключается в выделении в цельном тексте документа отдельных фрагментов и отнесению этих фрагментов к одному из заранее определенных типов данных. Исходными данными для принятия решения о типе данных каждого фрагмента является его стилистическое оформление, наличие ключевых слов или символов, его положение в тексте относительно других фрагментов и т.д. Четких алгоритмов и правил отнесения фрагмента к определенному типу данных часто не существует.

Для решения такой задачи необходим алгоритм, позволяющий принимать решения в условиях неопределенности. К таким алгоритмам в первую очередь относятся алгоритмы, выполняющие расчеты при помощи искусственных нейронных сетей, которые обладают способностью обучаться и принимать решение, опираясь на свои «знания», полученные в процессе обучения [1].

Среди существующих вариантов использования нейронных сетей для распознавания текста наилучшим образом подходит применение множества однослойных сетей, которые имеют бесконечное число нейронов с алгоритмом обучения «обратного распределения» [2]. Теория нейронных сетей допускает использование произвольного числа слоев и числа нейронов в каждом слое, однако фактически оно ограничено ресурсами компьютера.

Нейронная сеть состоит из множества нейронов, каждый из которых имеет множество входов называемыми синапсами и один выход называемый аксон. Каждый нейрон связан синапсами с входными параметрами нейронной сети, каждая такая связь имеет свой весовой коэффициент.

Процесс обучения нейронной сети сводится к поиску оптимальных значений всех весовых коэффициентов синоптических связей (некоторые из них могут быть постоянными), которые обеспечивают необходимую реакцию нейронной сети. От того, на сколько качественно будет выполнено обучение, зависит способность сети решать поставленные перед ней задачи во время эксплуатации. Качество обучения зависит от продолжительности процесса обучения и точности проводимых во время обучения расчетов и в большинстве случаев определяется из оптимального соотношения времени, потраченного на обучение, и приемлемостью получаемых результатов работы сети. Однако еще в большей степени способность нейронной сети принимать адекватные решения зависит от правильного выбора подаваемых на входы сети данных, которые будут служить основой для принятия решения.

Для использования нейронных сетей при анализе текста необходимо решить следующие задачи:

- представление входных данных для нейронной сети и обеспечение процесса их получения из анализируемого файла;
- хранение структуры и результатов обучения нейронной сети;
- выполнение действий по регистрации результатов в соответствии с реакциями на выходах нейронной сети.

Предлагается структура универсальной системы, использующей нейронные сети для анализа текста, которая включает в себя ряд программных модулей, решающих набор выше озвученных задач (см. рис. 1).

На рисунке видно, что имеется центральный модуль, который обеспечивает сопряжение всех остальных модулей. Взаимодействие между элементами ведется через строго определенный набор методов – программный интерфейс. Такая структура позволяет легко модернизировать систему, путем замены отдельного модуля системы другим в целях усовершенствования или применения системы в новых условиях.

Например, система может иметь несколько взаимозаменяемых модулей пользовательского интерфейса для работы в разных средах (Windows-приложение, WEB-интерфейс, консольное приложение, WEB-служба) независимо от остальных модулей системы. При этом каждый из этих модулей должен выполнять набор строго определенных функций, таких как выбор исходного файла для анализа, задание параметров работы алгоритма, информирование о ходе процесса расчета, запуск режима обучения и т.п.

Модуль взаимодействия с распознаваемым файлом (I/O) реализует программный интерфейс для работы с файлом. Он должен обеспечивать процесс считывания информации из исходного файла и обеспечивать доступ к файлу для запуска функций сбора информации для анализа и регистрации результатов. Например, этот модуль может быть реализован для файлов формата Microsoft Word и для файлов в формате XML.

В качестве параметров работы алгоритма посредством пользовательского интерфейса выбирается один из (или создается новый) наборов, включающих в себя структуру нейронной сети, результаты ее обучения, а также список функций для сбора исходной информации и записи результатов анализа. Элементы этого набора взаимозависимы и не имеют смысла друг без друга, т.к. структура сети зависит от количества функций,

обучение проводится при заданном наборе функций сбора информации, а функции регистрации результатов соответствуют выходам нейронной сети.

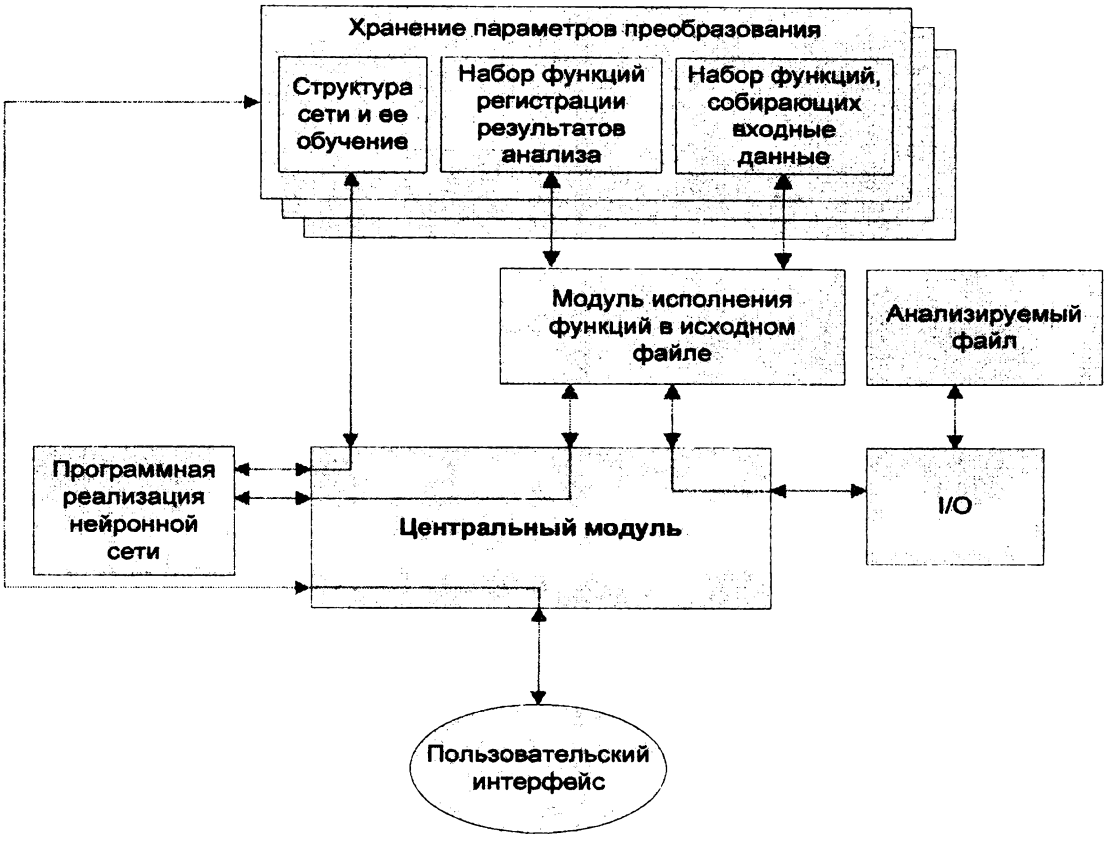

Рис.1. Структура универсальной системы анализа текста

Модуль исполнения функций отвечает за выполнение функций сбора информации и регистрации результатов в среде, соответствующей формату анализируемого файла. Например, для файлов Microsoft Word он будет осуществлять запуск функций на VBA и передачу полученных результатов в центральный модуль для дальнейшей их передачи на входы нейронной сети.

Программная реализация искусственной нейронной сети содержит все необходимые функции по инициализации нейронной сети и выполнению на ее основе вычислений. Этот модуль должен содержать следующие функции и свойства:

- Конструктор нейронной сети с указанием количества входов сети и массива имен выходов;
- Блок ввода/вывода входных и выходных данных. Отвечает за прием массива данных, подаваемых на входы сети, и вывод реакции сети в виде массива данных с выхода сети;
- Блок вывода решения, принятого нейронной сетью, по принципу «победитель получает все»;
- Функция ввода примера, который имеет следующие параметры: номер ячейки, в котором хранится пример (он необходим, в случае если понадобится перезаписать какой либо пример); массив входных данных для данного примера в виде дробных значений; имя выхода, соответствующего правильному решению для данного примера.

Модуль описывает однослойную нейронную сеть, обеспечивает динамическое увеличение или уменьшение этой сети, без потери весовых коэффициентов (знания сети). Кроме того, путем соединения входов и выходов нескольких экземпляров нейронных сетей может быть получена многослойная нейронная сеть.

Таким образом, центральный модуль получает в виде модуля программной реализации нейронной сети законченное средство для создания, динамического управления и использования сложных нейронных сетей. Причем за счет взаимозаменяемости остальных модулей системы существенно расширяется область применения уже созданных нейронных сетей без необходимости их повторного обучения, а вынесение набора функций для сбора информации и регистрации результата в хранилище параметров настройки алгоритма позволяет создавать библиотеку настроек, оптимизированных для разных вариантов анализируемых файлов.

Существует реализация описанной системы для решения задачи распознавания структуры заданий, вопросов и ответов для системы автоматизированной проверки знаний. Распознаваемые файлы представлены в формате Microsoft Word, текст включает в себя формулировку заданий, перечень вопросов, варианты ответов к этим вопросам, правильные ответы и количество баллов за правильный ответ. Ставится задача: 1) выделить заданными стилями фрагменты текста в соответствие с типом хранимой в них информации (вопрос, количество баллов, вариант ответа, правильный ответ); 2) определить тип вопроса: выбор одного из нескольких вариантов,

выбор нескольких вариантов, ввод ответа открытого типа, ввод числа, ввод набора чисел, указание правильной последовательности, указание соответствия (граф связей). Функции для сбора исходных данных написаны на VBA, регистрация результата производится путем применения стиля к текущему блоку текста и вставкой обозначения типа вопроса.

Сервис, основанный на описанной системе распознавания структуры вопросов теста, с подключенным модулем формирования автономных модулей для автоматизированной проверки знаний доступен по адресу: http://dist.ustu.ru/serverProcess/test.aspx.

Литература:

1. "Программирование искусственного интеллекта в приложениях", М. Тим Джонс, 2004, М., ДМК Пресс.
2. Введение в теорию нейронных сетей (http://www.orc.ru/~stasson/neurox.html)

Фомина Е.А.

ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ДЛЯ ПОВЫШЕНИЯ КВАЛИФИКАЦИИ СПЕЦИАЛИСТОВ НАЛОГОВЫХ ОРГАНОВ

kaffba@list.ru

Башкирская академия государственной службы и управления при Президенте Республики Башкортостан г. Уфа

Становление России как налогового государства, совершенствование налогового законодательства актуализируют проблему повышения эффективности деятельности налоговых органов. Отказ от сплошного метода при назначении проверок вызывает необходимость более точной идентификации предприятий – объектов налогового контроля. То есть оптимальная модель деятельности налоговых органов предполагает как можно более четкую идентификацию налогоплательщиков-нарушителей и повышение на этой основе эффективности налоговых проверок.

В практике экономических исследований проблема построения моделей зависимостей возникает в следующем контексте. Имеется некоторая группа экономических объектов. Относительно каждого объекта группы имеется определенная информация, характеризующая в различных аспектах его назначение, состояние, структуру или функции. В частности, имеется и информация, позволяющая считать некоторые из характеристик объекта зависимыми. Ставится задача: на основе всей имеющейся информации оценить (построить) точно или приближенно эту зависимость. Генеральная совокупность Ω рассматривается как множество, элементами которого являются объекты ю, однородные в каком то смысле. Объекты могут обладать различными свойствами. Некоторым свойствам даны такие определения, в соответствии с которыми свойство может быть охарактеризовано числом или функцией. Подобные свойства называются характеристиками.

Характеристика объектов из Ω, выражаемая с помощью элементов некоторого множества V, тем самым является отображением $\Omega \rightarrow V$ или функцией, заданной на множестве Ω и принимающей значения из множества V. Пространство всех подобных характеристик (измеримых отображений $\Omega \rightarrow V$ будем обозначать через $M(V)$. Если $V=\boldsymbol{R}$ - числовая прямая, то такие (числовые) характеристики объектов называются *показателями.* В дальнейшем будем рассматривать только характеристики, являющиеся показателями. Следовательно, V можно рассматривать как топологическое пространство и ввести на множестве характеристик V метрику $\rho(X, Y)$ - расстояние между точками X,Y из V.

В основе формирования информационной базы статистических зависимостей лежат понятия наблюдение, измерение, оценка. В различных работах даются разные определения этих понятий. Согласно определению Г.Б. Клейнер под наблюдением необходимо понимать реально выполнимую процедуру сбора и фиксации информации об объекте. Измерение (характеристики)- это также реально выполнимая процедура, позволяющая однозначно и точно определить значение этой характеристики в заданных условиях. Характеристика, допускающая измерение, называется измеримой. Оценкой характеристики объекта называется некоторое значение измеримой характеристики этого объекта. При этом оценка может принадлежать тому же множеству V,что и оцениваемая характеристика, но может и не принадлежать ему.

Наблюдения характеристики производятся с целью ее оценки, и обратно - любая оценка является результатом какого-то наблюдения. Однако произвольная оценка может нести мало информации о нужной характеристике или вообще не нести никакой информации о ней. В этой связи можно говорить об ошибках наблюдения, понимая под ними несоответствия между наблюдаемыми значениями характеристик и их истинными значениями. Под данную трактовку подпадают и ошибки регистрации, возникающие при искажении (умышленном или неумышленном) результата наблюдения при его фиксации.

Наблюдения могут характеризоваться неопределенностью, под которой понимается неполнота или неточность информации о значениях наблюдаемой характеристики. При этом неполнота определяется по отношению к тому (порой неизвестному) объему информации, который позволяет полностью определить истинное значение характеристики, а неточность трактуется как расхождение между истинными и полученными в ходе наблюдения данными. Неопределенность присуща практически любым наблюдениям, но мера, степень и конкретные формы ее проявления могут быть различными. В то время как неопределенность характеристики незначительна, можно считать ее детерминированной.