УДК 004.421.6

Е.В. Мешкова

МЕТОДИКА ПОСТРОЕНИЯ КЛАССИФИКАТОРА ТЕКСТА НА ОСНОВЕ ГИБРИДНОЙ НЕЙРОСЕТЕВОЙ МОДЕЛИ

Представленная в данной статье модель нейронной сети является попыткой найти гибридную архитектуру, в которой можно компенсировать недостатки семантической и ассоциативной нейронной сетевых парадигм для решения задачи автоматической классификации текста [1].

Исходя из сравнительного анализа семантических и нейронных сетей [2], предложена модель, позволяющая взаимно компенсировать недостатки обоих подходов. Предложенная модель нейронной сети состоит из трех слоев, которые выстроены иерархически, обобщая первоначальные единицы текста (слова) в понятия, и, затем, в области знаний.

Первый слой содержит нейроны, которым присваиваются значения слов (в дальнейшем – *слова*). Нейроны-*слова* связываются между собой на основе словарных определений *понятий*, в которые входят, и имеют прямые связи с соответствующими нейронами-*понятиями* второго слоя. Каждому *понятию* соответствует ряд нейронов-*слов* первого слоя.

Второй слой включает в себя нейроны как *понятия*, связанные с нейронами-*словами* первого слоя, которые входят в его определение, и с нейронами *областей знаний* (третий слой). Нейроны-*понятия* принадлежат различным *областям знаний*.

Третий слой составляют нейроны как *области знаний*. За каждой *областью знаний* закреплены соответствующие им *понятия*, которые могут принадлежать одновременно нескольким *областям знаний*. Третий слой является выходным и показывает, к какой области знания относится текст.

Таким образом, каждый нейрон сети соответствует слову, понятию, области знаний. Закрепление за нейроном конкретного значения (слова, понятия, области определения) осуществляется посредством присвоения порядкового номера каждому нейрону, которому соответствует закрепленное за ним слово, понятие, область определения, находящееся под таким же номером в библиотеке. Для каждого слоя создается своя библиотека.

Библиотека первого слоя включает в себя *слова*, за которыми закреплены нейроны, а также наименования всех *понятий* и *областей знаний* (ссылающиеся на соответствующие им *понятия* и *области знаний* во втором и третьем слоях). Это обусловлено тем, что слова, содержащиеся в тексте, проверяются на наличие в библиотеке первого слоя и подаются на вход. Библиотеки второго и третьего слоя содержат, соответственно, *понятия* и *области знания*, за которыми закреплены нейроны.

Закрепление отношений между нейронами внутри слоя, и с нейронами верхнего, по отношению к взятому, слоя, реализуется с помощью матриц смежности. Содержанию слоев сети соответствуют библиотеки, нейроны связаны между собой как внутри слоя, так и с нейронами верхнего, по отношению к взятому, слоя.

Построенная сеть представляет собой гибрид семантической и ассоциативной нейронной сети. Как известно, семантическая сеть – множество понятий (слов и словосочетаний), связанных между собой. Для каждого нейрона сети сформирован набор смысловых связей и заранее установлены семантические отношения, отражен-

ные в матрицах смежности, приведенных ниже. В данной модели применены семантические сети на начальном этапе обучения ассоциативной нейронной сети [3].

Если структура и принцип построения сети являются семантическими, то способ передачи сигнала нейронами аналогичен ассоциативной нейронной сети [3].

Нейрон первого слоя построенной сети (на начальном этапе) имеет количество входов, равное количеству установленных связей. Весовой коэффициент каждого синапса отражен в матрице весовых коэффициентов. Нейрон может, как принимать сигнал, так и передавать его.

Предложенная гибридная модель сети может быть представлена в виде графа, вершинами которого являются нейроны, а ребрами – установленные между ними связи, причем в виде графа представима, как вся сеть, так и отдельный слой. Следовательно, связи между нейронами, как внутри одного слоя, так и между нейронами разных слоев могут быть описаны матрицами смежности. Связи между нейронами первого слоя отражены квадратной матрицей V_1 размерности $n \times n$, где n – количество нейронов первого слоя, элемент которой v_{ij} характеризует связь i-го нейрона с j-м нейроном. Отношения между нейронами-*словами* первого слоя и нейронами-*понятиями* второго слоя представлены матрицей смежности V_{12} размерности $n \times m$, где n – количество нейронов первого слоя, а m – количество нейронов второго слоя. Элемент матрицы V_{12} v'_{ij} характеризует связь i-го нейрона-*слова* с j-м нейроном-*понятием*. Наличие связи между нейронами считается установленным, если $v'_{ij} = 1$.

Следует отметить, что нейроны-*понятия* на начальном этапе не связаны между собой, поэтому матрица V_2 не рассматривается. Отношения нейронов-*понятий* (второй слой) и нейронов-*областей знаний* (третий слой) представлены в матрице V_{23} аналогичной матрицам V_1 и V_{12}.

Матрицы смежности могут быть скорректированы на основе нечетких множеств, так как одни и те же нейроны слов и понятий принадлежат одновременно разным определениям и областям знаний.

Далее на основании матриц смежности создаются матрицы весовых коэффициентов для каждого слоя, отражающие значения синапсов нейронов. Синапсы (умножители) осуществляют связь между нейронами, умножая входной сигнал на число, характеризующее силу связи (вес синапса). На начальный момент для создаваемой модели входы и выходы равносильны, т.е. нейрон может, как принимать, так и передавать сигнал по своим связям. Каждой связи нейрона соответствует синапс, варианты установления синапсов рассчитываются на основе суммы всех весовых коэффициентов для каждого нейрона, равной одному и тому же произвольно заданному числу. Таким образом, наиболее часто встречающиеся слова получают меньшее значение, так как с большой вероятностью являются стоп-словами. Подобный подход индексирования по частоте часто используется в статистических методах распознания и классификации текста.

Значения синапсов связанных между собой нейронов представлены матрицей весовых коэффициентов $W1$ (размерности $n \times n$, где n – количество нейронов первого слоя), элемент которой $w1_{ij}$ характеризует вес i-го синапса j-го нейрона, где i – номер нейрона, j – номер входа данного нейрона. Элемент матрицы $w1_{ij}$ отражает силу связи i-го нейрона с j-м нейроном.

Как уже было отмечено, одно и то же слово может входить в определения различных понятий.

Начальное возбуждение x, подаваемое на вход сети, принимается за единицу, далее сигнал передается с помощью активационной пороговой функции [4], аналогично передаче сигнала в нейронных сетях.

На начальном периоде обучения слова-определения, связанные между собой и входящие в определение одного понятия, зациклены друг на друга и посылают возбуждения соответствующему понятию благодаря не столько подстройке коэффициентов, сколько пороговой функции и самой структуре сети. На третьем этапе, когда осуществляется генерация новых весовых коэффициентов, изменяется коэффициент, влияющий на чувствительность функции, и повышается чувствительность сети к более слабым сигналам. Возбуждение передается по установленным связям, причем на него влияет активационная функция, определяющая выходной сигнал, и весовой коэффициент связи. Сеть построена так, что на начальном этапе весовые коэффициенты минимально влияют на передачу сигнала от нейрона-*слова* к нейрону-*слову* и от нейронов-*слов* к нейрону-*понятию*.

На работу сети влияют также такие параметры, как пороговое значение, при котором нейрон передает возбуждение, количество тактов передачи возбуждения, значение перехода и стабилизации сигнала. В представленной модели данные параметры могут изменяться в процессе работы сети.

Пример работы сети. *Понятию* «изотоп» ставится в соответствие определение: «Нуклиды(9) с одинаковым(10) числом(11) протонов(12)», где каждое слово пронумеровано, и еще несколько понятий. *Понятию* «кварки» – определение: «фундаментальные(4) частицы(2) материи(5), из которых состоят(6) протоны(12) и нейтроны(13)».

Предположим, на вход пришло одно из слов-*определений* – «нуклид»(9). Нейрон **9** передает свое возбуждение (начальное возбуждение равно 1) связанным с ним нейронам **10**, **11**, **12**. На первом такте нейрон **10** передает полученное возбуждение на **11** и **12**. Полученное возбуждение умножается на вес синапса и суммируется [5], далее преобразуется пороговой активационной функцией *f(s)* и идет на выход. Аналогично с нейронами **11**, **12,** которые также получили первоначальный сигнал от нейрона **9** и передают возбуждение связанным с ними нейронам. Задается пороговая величина *D*, при значении $f(s) > D$ сигнал не передается. В данном случае принимается $D = 0{,}5$. На втором такте нейроны **10**, **11** и **12** передают возбуждение друг другу, помимо этого нейрон **12** передал возбуждение не связанным с другими нейронам **2**, **4**, **5**. Уже на 3 такте получается зацикливание связанных между собой нейронов друг на друга, в случае такого зацикливания возбуждение передается в слой 2 «понятий». Помимо этого, нейрон **12** на такте 1 передал сигнал связанным с ним нейронам **2**, **4**, **5**.

Из расчетов передачи сигналов нейронов видно, что активационная функция сильно увеличивает небольшой по значимости сигнал. Поэтому на втором этапе обучения сети ее параметры будут скорректированы, так как возникнет необходимость большего влияния со стороны весовых коэффициентов для сохранения установленных на этапе 1 связей. Если необходимо получить больше «родственных» понятий, принадлежащих к этой же области знаний, можно увеличить количество тактов.

Таким образом, на 1 этапе обучения (функционирования) сети выявляется количество в тексте специализированных терминов, порождающих зацикливание слов-определений и вызывающих обращение к слою понятий, а далее – к слою областей знаний или тематических разделов.

Нужно отметить, что модель может использоваться практически для всех областей, в которых существует четкая, сложившаяся терминология, и для смежных областей знаний. Преимуществом данной модели является также то, что она является не только классификатором, но и, в некоторой степени, системой, выделяющей ключевые темы, исходя из заложенной в текст терминологии. Эта же черта, а

также способность к выделению неявных ассоциаций [1, 3], отличает ее от классических статистических методов, частично использованных в модели. Отметим также, что предложенная модель позволяет значительно упростить сложный и трудоемкий процесс обучения нейронной сети. Разработанная методика создания гибридной нейросетевой модели позволяет классифицировать текст на основе заложенной в текст терминологии и выделения неявных ассоциаций. Разработан алгоритм построения и функционирования гибридной нейросетевой модели, структура сети и ключевые параметры сети.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Мешкова Е.В.* Построение гибридной модели на основе семантической и ассоциативной сетевых парадигм. – Шахты: ЮРГУЭС, 2004.
2. *Мешков В.Е., Мешкова Е.В.* Автоматическая классификация текстов на основе ассоциативных нейронных сетей // Материалы международ. научн.-практ. конф. «Информационные технологии и информационная безопасность в науке, технике и образовании» «ИНФОТЕХ – 2002» 30 сентября – 5 октября 2002г., Севастополь, Украина.
3. *Мешкова Е.В., Мешков В.Е.* Применение семантических сетей на начальном этапе обучения ассоциативных нейронных сетей //Материалы Международной научной конференции «Анализ и синтез как методы научного познания». – Таганрог: ТРТУ, 2004, Ч. 3. – 76 с.
4. *Круглов В.В., Борисов В.В., Харитонов Е.В.* Нейронные сети: конфигурации, обучение, применение. – Смоленск: Изд-во МЭИ, филиал в г. Смоленске, 1998.
5. *Круглов В.В., Борисов В.В.* Искусственные нейронные сети. Теория и практика / 2-е изд., стереотип. – М.: Горячая линия-Телеком, 2002. – 382 с.