УДК 004.91

Л.В. Серебряная, В.В. Потараев

МЕТОДЫ КЛАССИФИКАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ НА ОСНОВЕ ИСКУССТВЕННОЙ НЕЙРОННОЙ И СЕМАНТИЧЕСКОЙ СЕТЕЙ

Рассматривается применение искусственной нейронной сети в виде персептрона, сети Хопфилда и семантической сети для классификации текстовой информации. Изучаются процедуры обучения сетей, реализуются алгоритм обратного распространения ошибки в персептроне и алгоритм сходимости сети Хопфилда. Предлагается программное средство автоматической классификации текстов на основе разработанных моделей и алгоритмов. Оцениваются результаты работы программного средства.

Введение

В последние десятилетия резко возросла актуальность работы с документами, представленными в цифровом виде. Вследствие этого появились новые модели, методы и инструменты для работы с ними.

Хранение больших объемов информации практически оправдано только при условии, если ее поиск и обработка осуществляются быстро и выдается она в доступной для понимания форме. Один из видов обработки текстовой информации – это ее классификация (или рубрикация), которая находит ряд применений: фильтрация спама, разделение электронных сообщений по категориям, подбор контекстной рекламы, классификация научных статей и др.

Классификации информации предшествует этап ее анализа. Большинство существующих подходов к анализу текстов можно разбить на два типа. К первому относятся простые, быстрые, но не очень точные механизмы анализа. Чаще всего эти подходы используют формальные статистические методы, основанные на частоте появления в тексте слов различных тематик. Второй тип формируют достаточно сложные, дающие хороший результат, но сравнительно медленные подходы, основанные на лингвистических методах. Эффективным же можно считать такой подход, который сочетал бы в себе «простоту» статистических алгоритмов с достаточно высоким качеством обработки лингвистических методов.

Одним из видов лингвистического анализа является семантический анализ. С его помощью исследуются смысл слов и предложений, их связь между собой и с окружающей действительностью. Методы реализации семантического анализа и классификации текстовой информации связаны с использованием искусственных нейронных сетей (ИНС) и семантических сетей, которым посвящена данная работа.

1. Модели обработки текстовой информации на основе семантического анализа

Накопленный теоретический и практический опыт работы с ИНС позволяет выделить множество вариантов их построения, рассмотренных ниже.

Если ИНС состоит из нейронов одного типа, она называется однородной, если же в ней комбинируются слои нейронов разных типов, – гибридной.

По количеству слоев нейронов сети делятся на однослойные, в которых отсутствуют непосредственные связи выходов одних нейронов со входами других, и многослойные, где имеются указанные связи.

Сеть является однонаправленной, если в ней отсутствует передача сигналов с последующих слоев на предыдущие. Нейронная сеть с обратными связями называется рекуррентной.

По типу обучения ИНС делятся на обучающиеся с учителем (контролируемое обучение) и самообучающиеся. Для сетей первого типа задается обучающая выборка, для каждого образа которой заранее известна принадлежность к одному из заданных классов. В сетях второго типа,

популярным представителем которых считается сеть Кохонена, обучающая выборка отсутствует, а определение количества классов и отнесение к ним образов выполняются в ходе реализации алгоритма распознавания.

Многослойные сети могут привести к увеличению вычислительной мощности по сравнению с однослойной сетью лишь в том случае, если активационная функция между слоями будет нелинейной. Иначе любая многослойная линейная сеть может быть заменена эквивалентной однослойной сетью, которая весьма ограничена по своим вычислительным возможностям.

Анализ различных типов ИНС показал, что для классификации текстовой информации наиболее подходят два из них: многослойный персептрон и сеть Хопфилда. Рассмотрим их более подробно [1].

Одним из самых важных условий успешной работы ИНС является алгоритм обучения сети. Многослойный персептрон (рис. 1) обучается с помощью алгоритма обратного распространения ошибки. Целью обучения сети этим алгоритмом является такая корректировка ее коэффициентов, при которой инициализация множества входов приводит к требуемому множеству выходов. Обычно множества входов и выходов называются векторами X и Y соответственно. При обучении предполагается, что для каждого входного вектора существует парный ему целевой вектор, задающий требуемый выход. Вместе они образуют обучающую пару. Сеть обучается на многих парах.

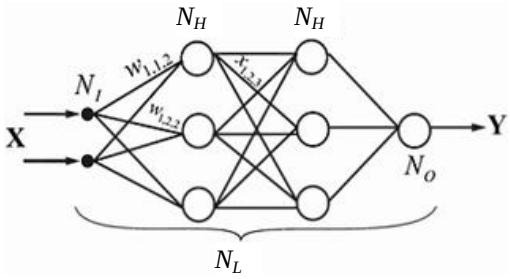

Рис. 1. Сеть в виде многослойного персептрона

Алгоритм обратного распространения ошибки предполагает два прохода по всем слоям сети: прямой и обратный. При прямом проходе входной вектор подается на входной слой нейронной сети, после чего распространяется по сети от слоя к слою. В результате генерируется набор выходных сигналов, который и является фактической реакцией сети на данный входной образ. Во время прямого прохода все весовые коэффициенты сети фиксированы. Во время обратного прохода все коэффициенты настраиваются в соответствии с правилом коррекции ошибок, а именно: фактический выход сети вычитается из желаемого, в результате чего формируется сигнал ошибки. Впоследствии этот сигнал распространяется по сети в направлении, обратном направлению связей весовых коэффициентов. Они настраиваются с целью максимального приближения выходного сигнала сети к желаемому [2].

В качестве активационной функции в многослойных персептронах, как правило, используется сигмоидальная функция, в частности бинарный сигмоид:

$$f(x) = \frac{1}{1+e^{-x}}. \quad (1)$$

Выражение

$$f'(x) = f(x) * [1 - f(x)] \quad (2)$$

является производной функцией бинарного сигмоида.

Алгоритм обратного распространения ошибки состоит из следующих шагов:

1. Инициализировать весовые коэффициенты маленькими случайными значениями.

2. Выбрать очередную обучающую пару из обучающего множества, подать входной вектор на вход сети.

3. Вычислить выход сети.

4. Вычислить разность между выходом сети и требуемым выходом (целевым вектором обучающей пары).

5. Подкорректировать коэффициенты сети для минимизации ошибки.

6. Повторять шаги 2–5 для каждого вектора обучающего множества до тех пор, пока ошибка на всем множестве не достигнет приемлемого значения.

Операции, выполняемые на шагах 2 и 3, сходны с теми, которые выполняются при функционировании уже обученной сети, т. е. подается входной вектор и вычисляется получающийся выход. Вычисления выполняются послойно. Шаги 2 и 3 образуют «проход вперед», так как сигнал распространяется по сети от входа к выходу. Шаги 4 и 5 составляют «обратный проход», когда вычисляемый сигнал ошибки распространяется обратно по сети и используется для корректировки весовых коэффициентов.

Рассмотрим подробнее процедуру корректировки коэффициентов ИНС, при которой следует выделить два возможных случая:

1. Изменение коэффициентов выходного слоя.

Введем величину δ, которая равна разности между требуемым T и реальным OUT выходами, умноженной на производную функции активации (бинарный сигмоид):

$$\delta_q = OUT_q\left(1 - OUT_q\right)\left(T_q - OUT_q\right). \tag{3}$$

Тогда коэффициенты выходного слоя после коррекции определяются по формуле

$$\omega_{p-q}(i+1) = \omega_{p-q}(i) + \eta\delta_q OUT_P, \tag{4}$$

где i – номер текущей итерации обучения;

ω_{p-q} – величина весового коэффициента, соединяющего нейрон p с нейроном q;

η – коэффициент скорости обучения, позволяющий управлять средней величиной изменения коэффициентов.

2. Корректировка весовых коэффициентов скрытого слоя.

Введем величину δ:

$$\delta_q = OUT_q\left(1 - OUT_q\right)\sum_{k=1}^{M}\delta_k\omega_{q-k}. \tag{5}$$

Коэффициенты скрытых слоев после коррекции также будут определяться с помощью выражения (4).

Достоинством алгоритма обратного распространения ошибки является то, что он реализует вычислительно эффективный метод обучения многослойного персептрона. При этом его все-таки нельзя считать универсальным решением. Больше всего трудностей возникает с неопределенно долгим процессом обучения, в ходе которого величина шага коррекции коэффициентов может конфликтовать с качеством и временем обучения. В роли входных значений обучающей выборки выступают наборы ключевых слов текстов, а в роли выходных значений – номера классов, к которым относятся эти же тексты.

Сеть Хопфилда занимает особое место в ряду ИНС. В ней впервые удалось установить связь между нелинейными динамическими системами и нейронными сетями. Образы памяти сети соответствуют устойчивым предельным точкам (аттракторам) динамической системы. При этом появилась возможность теоретически оценить объем памяти сети Хопфилда, определить область ее параметров, в которой достигается наилучшее функционирование [3].

В общем случае модель Хопфилда может быть представлена сетью, содержащей произвольные обратные связи. По этим связям переданное возбуждение возвращается к данному нейрону, и он повторно выполняет свою функцию (рис. 2).

Обратные связи могут вызывать неустойчивости в поведении ИНС, что проявляется в блуждающей смене состояний нейронов, не приводящей к стационарным состояниям. Однако было показано, что сети Хопфилда устойчивы, и их можно определить как динамическую систему с обратной связью, у которой выход одной операции служит входом следующей операции сети. Каждая операция сети называется итерацией. Устойчивость сети подразумевает, что она

может сходиться к одной из зафиксированных (неподвижных) точек, которая зависит от исходной точки, выбранной для начальной итерации. Множество неподвижных точек сети Хопфилда – это ее память. В этом случае сеть может действовать как ассоциативная память.

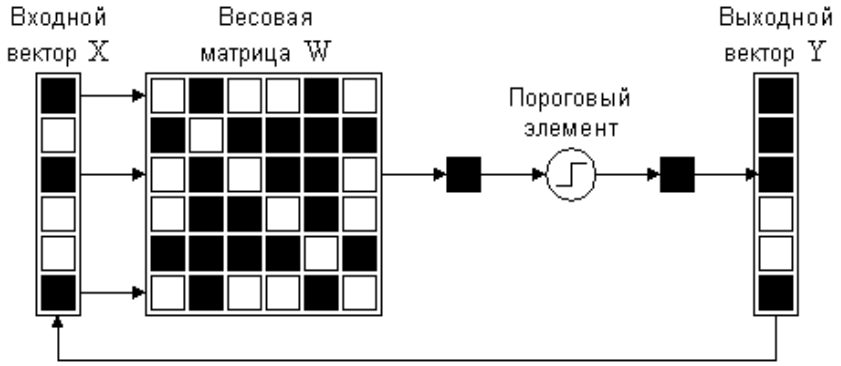

Рис. 2. Нейронная сеть Хопфилда

Рассмотрим алгоритм сходимости сети Хопфилда:

1. Вычислить компоненты выходного вектора $Y_j, j = 1,2,\dots n,$ по формуле

$$Y_j = T(\sum_{i=1}^{n} w_{ij}\, x_i), \tag{6}$$

где $T(x)$ = sign(x), т. е.

$$T(x) = \begin{cases} -1, \text{если } x < 0; \\ 1, \text{если } x > 0. \end{cases} \tag{7}$$

2. Выполнить асинхронную коррекцию.

2.1. По формуле (6) найти вектор Y_j.

2.2. Заменить $X = (x_1, x_2, x_3, \dots x_n)$ на $Y = (y_1, x_2, x_3, \dots x_n)$ и подать Y на вход X.

2.3. Повторить процесс, чтобы найти y_2, y_3 и т. д.

2.4. Повторять шаги 2.2 и 2.3 до тех пор, пока вектор не перестанет изменяться. Каждый шаг уменьшает величину энергии связей, поэтому обеспечивается сходимость к неподвижной точке (аттрактору).

Асинхронная коррекция и нули на диагонали матрицы весовых коэффициентов *W* гарантируют, что энергетическая функция будет уменьшаться с каждой итерацией. Весовая матрица отличает поведение одной сети Хопфилда от поведения другой.

К недостаткам сети Хопфилда можно отнести ее сравнительно небольшой объем памяти, вследствие чего попытка записи большего числа образов приводит к тому, что нейронная сеть перестает их распознавать. Кроме того, достижение устойчивого состояния не гарантирует правильный ответ сети. Это происходит из-за того, что ИНС может сойтись к так называемым ложным аттракторам.

Для обучения сети Хопфилда составляется обучающая выборка. Это наборы ключевых слов знакомых текстов. Наличие слова в тексте означает единичный сигнал на входе нейронной сети. При обучении можно рассчитать коэффициенты матрицы связей на основании слов обучающих текстов. При классификации текстов входной сигнал после выполнения некоторого количества итераций будет приводить нейронную сеть в стационарное состояние, соответствующее образу, на котором проводилось обучение. Каждый из обучающих образов соответствует одному из классов, к которым нужно относить незнакомые тексты.

Еще одним способом семантической обработки текстов является подход на основе семантических сетей. Семантические сети позволяют выделять смысл текста в виде понятий и связей между ними, образующих граф. Понятия семантической сети записываются в вершинах графа, а отношения между понятиями – это дуги графа. Количество типов отношений в семан-

тической сети определяется ее разработчиком исходя из конкретных целей. Часто используются иерархические семантические сети, в которых отношения образуют древовидную структуру. Отношения в сетях могут быть разных типов: функциональными, количественными, пространственными, временными, логическими и др. [4].

К достоинствам семантических сетей можно отнести:

– универсальность, достигаемую за счет выбора соответствующего набора отношений;

– наглядность системы знаний, представленную графически;

– близость структуры сети, представляющей систему знаний, к семантической структуре фраз на естественном языке;

– соответствие современным представлениям об организации долговременной памяти человека.

Недостатки семантических сетей:

– сетевая модель не всегда дает ясное представление о структуре предметной области, поэтому формирование и модификация такой модели могут быть затруднительными;

– сетевые модели представляют собой структуры, для обработки которых необходим специальный аппарат формального вывода;

– проблема поиска решения в семантической сети сводится к задаче поиска ее фрагмента, отражающего поставленный запрос. Это может обусловливать сложность поиска решения в семантических сетях;

– представление, использование и модификация знаний при описании систем реального уровня сложности оказываются трудоемкими процедурами, особенно при наличии множественных отношений между их понятиями.

При решении задачи классификации текстовой информации с помощью семантических сетей можно в качестве узлов сети принять некоторые концепты, о которых идет речь в тексте, а в качестве дуг – связи между концептами. В таком случае обучение классификатора будет представлять собой создание сетей по некоторым обучающим текстам – одна сеть на одну рубрику. При классификации текстов для каждого из них нужно построить собственную семантическую сеть, а затем на основании сходства сети текста и сетей рубрик относить текст к одной из рубрик.

2. Алгоритмы классификации текстов с помощью искусственных нейронных и семантических сетей

Рассмотрим алгоритмы классификации текстов на основе ИНС двух типов: персептрона и сети Хопфилда. В обоих случаях выполняется контролируемое обучение сетей, для чего используются наборы ключевых слов обучающих текстов, построенные по законам Зипфа.

Для выделения понятий текста, представляющих слова и словосочетания, может быть применен статистический алгоритм, основанный на анализе частоты встречаемости слов, цепочек слов и их вхождения друг в друга. Во всех созданных человеком текстах можно выделить статистические закономерности, которые никому не удается обойти. Независимо от текста и языка его написания внутренняя структура текста остается неизменной.

Если измерить количество вхождений каждого слова в текст и взять только одно значение из каждой группы, расположить частоты по мере их убывания и пронумеровать (порядковый номер частоты называется рангом частоты), то наиболее часто встречающиеся слова будут иметь ранг 1, следующие за ними – 2 и т. д. Кривая Зипфа, имеющая вид равносторонней гиперболы, отражает зависимость между частотой вхождения слов в текст и рангом слов. Исследования показали, что наиболее значимые слова лежат в средней части кривой и имеют ранг от 4 до 17. Это объясняется тем, что слова, которые встречаются очень часто, в основном оказываются предлогами и местоимениями. Редко встречающиеся слова в большинстве случаев тоже не имеют решающего смыслового значения [5].

Количество входных нейронов сети равно суммарному количеству выбранных ключевых слов во всех текстах:

$$N = M \times K, \tag{8}$$

где N – количество входов нейронной сети; M – количество слов, отобранных из каждого текста; K – количество обучающих текстов.

Выходам нейронной сети можно сопоставить классы, к которым относятся обучающие тексты.

Рассмотрим процедуру классификации текстов при помощи алгоритма обратного распространения ошибки.

В качестве функции активации нейронов используется бинарный сигмоид (1) и его производная (2). Начальным значениям весовых коэффициентов связей между нейронами присваиваются случайные числа от 0 до 1. Скорость обучения подбирается таким образом, чтобы оно завершалось в течение нескольких итераций обратного распространения ошибки. Еще одним параметром является максимально допустимое количество итераций обучения.

По завершении обучения выполняется проверка работы сети на обучающей выборке. Каждое из M ключевых слов каждого из K текстов подается на входы сети, и она должна корректно определить номер класса. Только в этом случае обучение считается успешным.

Для проверки классификации используется набор тестовых текстов, для каждого из которых сеть должна определить выход (класс принадлежности). Если такие выходы не найдены или их найдено несколько для одного текста, то решение о классификации не принимается. В то время как процесс обучения является итерационным, классификация представляет собой расчет функции активации для каждого из узлов нейронной сети.

В случае выполнения алгоритма классификации текстов на основе нейронной сети Хопфилда обычно используется функция активации, представленная выражением (7). Начальные весовые коэффициенты связей нейронов задаются случайными небольшими значениями, а процесс обучения заключается в расчете матрицы связей W.

После расчета матрицы связей нейронная сеть готова к выполнению классификации. Она представляет собой итерационный процесс, при котором сеть последовательно меняет свои состояния. Это происходит до тех пор, пока ИНС не попадет в некоторое стационарное состояние, соответствующее одному из обучающих сигналов. Поскольку условие прекращения итераций может никогда не выполниться, классификацию обычно ограничивают максимальным числом итераций. Если по окончании итераций состояние сети не соответствует ни одному из обучающих текстов, то классифицируемый текст относят к классу, наиболее близкому к одному из обучающих сигналов.

Для классификации текстов на основе семантической сети представим текст как множество предложений, каждое из которых состоит из подлежащего, сказуемого и дополнений. Узлы сети – это некоторые понятия, выражаемые подлежащими и дополнениями, а связи между узлами задаются сказуемыми. С целью упрощения решаемой задачи остальные члены предложений исключены из рассмотрения. Для определения части речи слов, а также их начальных форм используются словари начальных форм существительных и глаголов русского языка. Особенностью семантической сети является определение схожести текстов по их смыслу, даже если в них используются разные слова. Для обеспечения такой возможности узлом сети принято считать не только само слово, но и все его синонимы. Аналогично связью можно считать не глагол, а множество синонимичных глаголов, означающих примерно одинаковые отношения.

Рассмотрим алгоритм классификации текстов, использующий сеть, построенную предложенным способом. На этапе обучения происходит добавление каждого предложения текста к создаваемой семантической сети. Если подлежащее (либо дополнение) уже есть в сети, то к нему добавляется дуга, соответствующая сказуемому. Если в предложении нет сказуемого, то оно со всеми своими синонимами добавляется в сеть как не связанный ни с чем узел. На рис. 3 показан пример семантической сети, построенной для двух предложений.

Назовем участком сети любой узел либо дугу и пару связанных с ней узлов. На этапе классификации при поиске участка сети, наиболее близкого к предложению классифицируемого текста, близость предложения и участка сети можно вычислять по-разному. Например, если совпадают подлежащие в двух дугах, то близость равна X, а если еще совпадают сказуемые, то близость равна $X + Y$. Будем считать величины X и Y весами близости подлежащего и сказуемого соответственно.

Необходимо отметить, что количество итераций данного алгоритма зависит только от числа предложений в обрабатываемых текстах. Близость текстов можно оценивать на основе оценки близости дуг соответствующих им сетей.

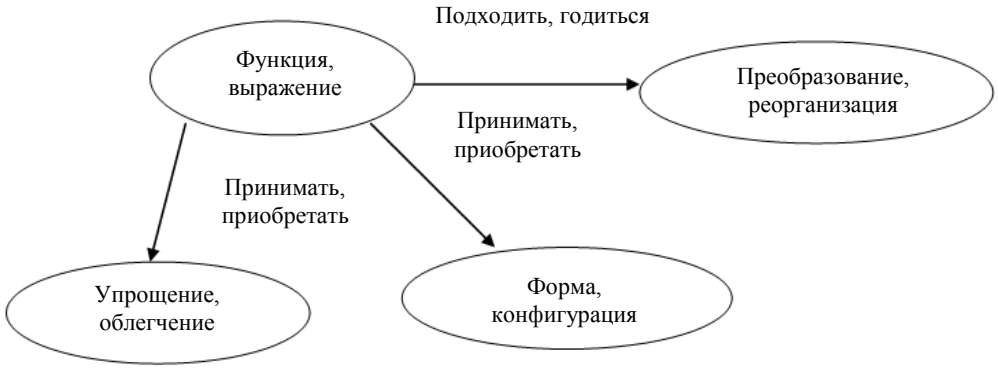

Рис. 3. Фрагмент семантической сети

Для работы алгоритма нужны словарь начальных форм существительных, словарь глаголов и словарь синонимов русского языка. На начальном этапе для ускорения алгоритма можно заменить каждый текст подмножеством его предложений (например, 100 первых предложений текста или 100 самых длинных предложений). Данную сеть можно дополнить и другими правилами, которые позволяют генерировать новые высказывания, увеличивая порождающую способность семантической сети. Тогда для повышения качества работы сети вводится вероятность каждой связи.

3. Сравнение методов классификации текстов и анализ полученных результатов

На основе ранее рассмотренных моделей и алгоритмов было разработано программное средство, предназначенное для автоматической классификации текстовых документов.

На персептроне и сети Хопфилда проводилась процедура контролируемого обучения. Для этого на входы сетей подавались векторы ключевых слов текстов, составляющих обучающую выборку. Количество классов, размер обучающей выборки, принадлежность текстов к рубрикам задает пользователь, при этом все данные могут изменяться. Результат обучения – определение номера класса для каждого обучающего образа. В качестве начальных значений весовых коэффициентов связей между нейронами использовались случайные числа от 0 до 1. Скорость обучения подбиралась таким образом, чтобы для небольшой обучающей выборки процесс обучения завершался достаточно быстро. Дополнительно можно задать максимальное количество итераций обучения.

По завершении процедуры обучения можно оценить характеристики классификации каждой из ИНС. Сначала задавались три класса, десять обучающих текстов, каждый из которых был представлен вектором из десяти ключевых слов. Размер текстов – 100–200 Кб. Качество классификации текстов оказалось у сетей приблизительно одинаковым, но сеть Хопфилда работала быстрее. С увеличением обучающей выборки и вектора ключевых слов увеличивалось время обучения, а качество классификации становилось выше.

Перейдем к реализации семантической сети. В простейшем случае в разработанной сети каждому предложению текста ставится в соответствие некоторая дуга. Близость текстов фактически означает совпадение дуги либо узла сети одного текста с дугой либо узлом другого текста. В качестве параметров сети выбирались следующие: минимальная близость для принятия решения о классификации; веса близости подлежащих, сказуемых, дополнений. Для заданных трех классов семантическая сеть выдала более 80 % корректных ответов, что явилось лучшим результатом, чем результаты обеих ИНС. Время работы семантической сети оказалось больше, чем у нейронных сетей.

Рассмотрим пример разделения текстов на пять классов с помощью двух ИНС и семантической сети (таблица). С увеличением числа классов качество классификации персептроном немного ухудшилось, сетью Хопфилда – стало существенно хуже, а семантическая сеть работала практически без изменений. Это можно объяснить следующим образом: персептрону не хватило итераций для того, чтобы закончить процесс обучения, а сеть Хопфилда начала сходиться к ложной стационарной точке. Кроме того, с увеличением обучающей выборки и длины векторов ключевых слов ИНС для обучения требуется большее количество итераций. В результате они теряют преимущество по времени перед семантической сетью.

Результаты классификации текстов на пять рубрик

Количество текстов	Рубрика текста	Количество текстов, использованных для обучения	Количество верных ответов		
			Классификация персептроном	Классификация сетью Хопфилда	Классификация семантической сетью
7	Теория вероятностей	2	6	6	5
12	Философия	3	9	7	6
15	Информатика	3	12	8	10
20	Геометрия	4	15	10	17
22	Физика	3	15	10	20

Созданное программное средство позволяет подбирать оптимальные параметры сетей в зависимости от требований пользователя. Так, например, при небольшом количестве классов и нежестких требованиях к качеству классификации можно выбирать сеть Хопфилда. Алгоритм на основе нейронной сети с обратным распространением ошибки показывает достаточно стабильные результаты при различных количествах классов, допуская много ошибок. Для достижения баланса между скоростью и качеством классификации можно использовать две сети: нейронную сеть с обратным распространением ошибок и предложенную семантическую сеть в случае, если нейронная сеть не сумела определить класс.

Заключение

В работе изучены модели и алгоритмы классификации текстовой информации. Предложено применение методов семантического анализа текстовой информации для решения задачи классификации. Семантический анализ связан с выделением информационно-логической основы текста, что и было выполнено в работе.

Предложены алгоритмы классификации текстовой информации на основе искусственных нейронных и семантической сетей. Для всех моделей применялся метод контролируемого обучения, обеспечивающий более точное решение поставленной задачи.

Создано программное средство, реализующее алгоритм обратного распространения ошибки для персептрона, алгоритм сходимости сети Хопфилда и алгоритм классификации семантической сети. Работу программного средства можно настраивать с помощью различных параметров в зависимости от решаемой прикладной задачи. Результаты работы программы представляются в удобном аналитическом и графическом виде. Сравнительный анализ полученных результатов показал, что семантическая сеть дает более точные результаты, хотя и отстает от ИНС по скорости работы. При увеличении количества классов разница в скорости работы сетей уменьшается. Для улучшения результатов по качеству и времени работы предлагается комбинировать различные модели и алгоритмы сетей.

Список литературы

1. Искусственная нейронная сеть [Электронный ресурс]. – Режим доступа : https://ru.wikipedia.org/wiki/Искусственная_нейронная_сеть. – Дата доступа : 06.06.2016.

2. Алгоритм обратного распространения ошибки [Электронный ресурс]. – Режим доступа : http://www.aiportal.ru/articles/neural–networks/back–propagation.html. – Дата доступа : 06.06.2016.

3. Нейронная сеть Хопфилда и ее применение [Электронный ресурс]. – Режим доступа : http://iasa.org.ua /lections/tpr/neuro/hopfield.htm. – Дата доступа : 06.06.2016.

4. Семантические сети или сетевые модели знаний [Электронный ресурс]. – Режим доступа : http://www.aiportal.ru/articles/knowledge-models/semantic-network.html. – Дата доступа : 06.06.2016.

5. Серебряная, Л.В. Информационное обеспечение финансовых структур / Л.В. Серебряная // Методическое пособие к лабораторным работам для студентов специальности «Программное обеспечение информационных технологий» всех форм обучения. – Минск : БГУИР, 2011. – 43 с.

Поступила 13.07.2016

Белорусский государственный университет
информатики и радиоэлектроники,
Минск, ул. П. Бровки, 6
e-mail: l_silver@mail.ru,
vic229@rambler.ru

L.V. Serebryanaya, V.V. Potaraev

METHODS OF TEXT INFORMATION CLASSIFICATION ON THE BASIS OF ARTIFICIAL NEURAL AND SEMANTIC NETWORKS

The article covers the use of perseptron, Hopfild artificial neural network and semantic network for classification of text information. Network training algorithms are studied. An algorithm of inverse mistake spreading for perceptron network and convergence algorithm for Hopfild network are implemented. On the basis of the offered models and algorithms automatic text classification software is developed and its operation results are evaluated.