УДК 004.8-004.912

Е.В. Бодянский, О.В. Шубкина

СЕМАНТИЧЕСКОЕ АННОТИРОВАНИЕ ТЕКСТОВЫХ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ МОДИФИЦИРОВАННОЙ ВЕРОЯТНОСТНОЙ НЕЙРОННОЙ СЕТИ

В работе предложена модифицированная вероятностная нейронная сеть, которая используется для интеллектуального анализа текстовой информации и построения семантических аннотаций на основе полученных знаний и онтологии предметной области.

Ключевые слова: семантическое аннотирование, вероятностная нейронная сеть, многослойная архитектура

1. Введение

Ввиду того, что большая часть информации в корпоративных системах хранится в текстовом виде (в виде электронных документов, рассылок новостей), каждому экземпляру концептов онтологии, отражающей структурные знания, можно поставить в соответствие какой-либо текстовый документ или какую-то его часть в зависимости от заданных условий. Такой процесс формирования метаданных называется семантическим аннотированием [1] и состоит из трех основных компонент: онтологии, корпуса текстов и способа построения классификатора для извлечения знаний.

Существует набор стандартных решений, которые разработаны для описания метаданных и формирования семантических аннотаций, как например, стандарт Dublin Core [2]. Однако набор заданных тегов для описания текстовых документов не отражает информацию, которая может являться актуальной для текущей онтологии предметной области, а зачастую несет лишь общие сведения. Стоит отметить также, что создание семантических аннотаций вручную занимает достаточно много времени и требует больших денежных затрат. Это привело к разработке методов полуавтоматического построения семантических аннотаций, которые в свою очередь имеют ряд недостатков, например, использование шаблонов заполнения или априори задан-

© Бодянский Е.В., Шубкина О.В., 2011

 ISSN 1562-9945

ных правил. Поэтому актуальной задачей является разработка моделей и методов семантического аннотирования текстовых документов.

2. Постановка задачи

Процесс семантического аннотирования можно рассматривать как проблему классификации, при этом автоматизация может быть достигнута путем применения методов интеллектуального анализа данных [3]. Главной задачей в таком случае является построение классификатора на основе онтологии *Ont* и исходной выборки данных *ObjectSet* для получения проекции текстового объекта на онтологию предметной области. Под проекцией в данном случае следует понимать отнесение некоторого текстового объекта к классу онтологии как экземпляра концепта данного класса.

Нейросетевой подход для получения семантических аннотаций текстовых документов дает возможность построить классификатор, независимый от уровней семантического аннотирования (уровень слов, предложений, документов). В таком случае можно использовать различные способы представления исходной текстовой информации в векторном пространстве признаков, что не отражается на качестве обработки данных.

С точки зрения задачи классификации формальное представление семантического аннотирования текстового документа можно получить следующим образом. Для данной онтологии предметной области *Ont* набор концептов (классов) определяется как $\mathrm{ConceptSet} = \left(\mathrm{c}(1), \mathrm{c}(2), \ldots, \mathrm{c}(\mathrm{i}), \ldots \mathrm{c}(\mathrm{N}_1)\right)$, где *c(i)* – *i*-й концепт из *Ont*. Для рассматриваемого текстового корпуса набор текстовых объектов, полученных на этапе предобработки, можно представить как $\mathrm{ObjectSet} = \left(\mathrm{x}(1), \mathrm{x}(2), \ldots, \mathrm{x}(\mathrm{j}), \ldots, \mathrm{x}(\mathrm{N}_2)\right)$, где *x(j)* – *j*-й текстовый объект, представленный в виде некоторого набора релевантных признаков в векторной форме, N_1 и N_2 – количество концептов (классов) онтологии и мощность исходной выборки текстовых объектов соответственно.

Тогда семантическая аннотация – разметка или набор метаданных рассматриваемого текстового документа – на основе данной онтологии будет определена как $\mathrm{LabelSet} = \left\{ \mathrm{l_i} \middle| \exists \mathrm{c_j} \in \mathrm{ConceptSet} \wedge \mathrm{l_i} = \mathrm{c_j} \right\}$, в котором *LabelSet* – уникальное множество для каждого текстового документа, состоящее из концептов (классов) онтологии предметной

области, полученных путем проекции множества текстовых объектов, принадлежащих этому документу, на заданную онтологию с использованием методов на основе ИНС. Следует уточнить, что в рамках рассматриваемого подхода, задав пороговое значение ThresholdValue, которое определяется на основе разницы принадлежности объекта к каждому классу, появляется возможность отнести объект к нескольким из потенциально возможных концептов онтологии для устранения неоднозначности.

В связи с этим, введем модифицированную вероятностную нейронную сеть с многослойной архитектурой как основу для получения семантических аннотаций текстовых документов. При этом априорно заданная обучающая выборка $x(1), x(2), \ldots, x(N_2)$ состоит из текстовых объектов *ObjectSet*, полученных после предобработки данных, а значения на выходах нейронной сети формируют множество *LabelSet*.

3. Модифицированная вероятностная нейронная сеть

Рассматриваемая задача может быть решена на основе методов байесовской классификации с помощью вероятностных нейронных сетей (PNN), введенных Д.Ф. Шпехтом [4].

Идея байесовской классификации состоит в том, что для каждого входного образа-вектора можно принять решение на основе выбора наиболее вероятного класса из тех, которым мог бы принадлежать данный образ. Это решение требует оценки функции плотности вероятностей для каждого класса, восстанавливаемой на основе анализа данных из обучающей выборки. Для восстановления этих функций широкое распространение получили оценки Парзена (Надарая – Ватсона), использующие весовые функции (потенциальные функции, ядерные функции), имеющие центр в точках, соответствующих образам с известной классификацией из обучающей выборки.

И хотя байесовские методы классификации известны давно, их параллельная нейросетевая реализация позволила обеспечить более высокое быстродействие процессом обработки информации, связанным с распознаванием образов, классификацией, диагностикой и т.п.

Важно отметить, что решение, получаемое с помощью стандартной вероятностной сети, позволяет отнести предъявляемый образ $x(k)$ к одному единственному классу с наиболее плотным распределением в области этого образа. Вместе с тем существует достаточно ши-

 ISSN 1562-9945

рокий класс задач, где такое однозначное решение не достаточно, т.е. в процессе обработки информации необходимо определить не один наиболее вероятный класс, а вероятности принадлежности $x(k)$ к каждому из потенциально возможных классов.

Такое решение может быть получено с помощью предлагаемой модифицированной нейронной сети (MPNN), архитектура которой приведена на рис.1. Сразу же можно заметить, что MPNN представляет собой гибрид стандартной PNN и обобщенной регрессионной нейронной сети (GRNN), также введенной Шпехтом [5], и содержит четыре слоя обработки информации: первый скрытый, именуемый слоем образов, второй скрытый слой локальных сумматоров, третий скрытый слой, содержащий единственный общий сумматор, и, наконец, выходной слой делителей.

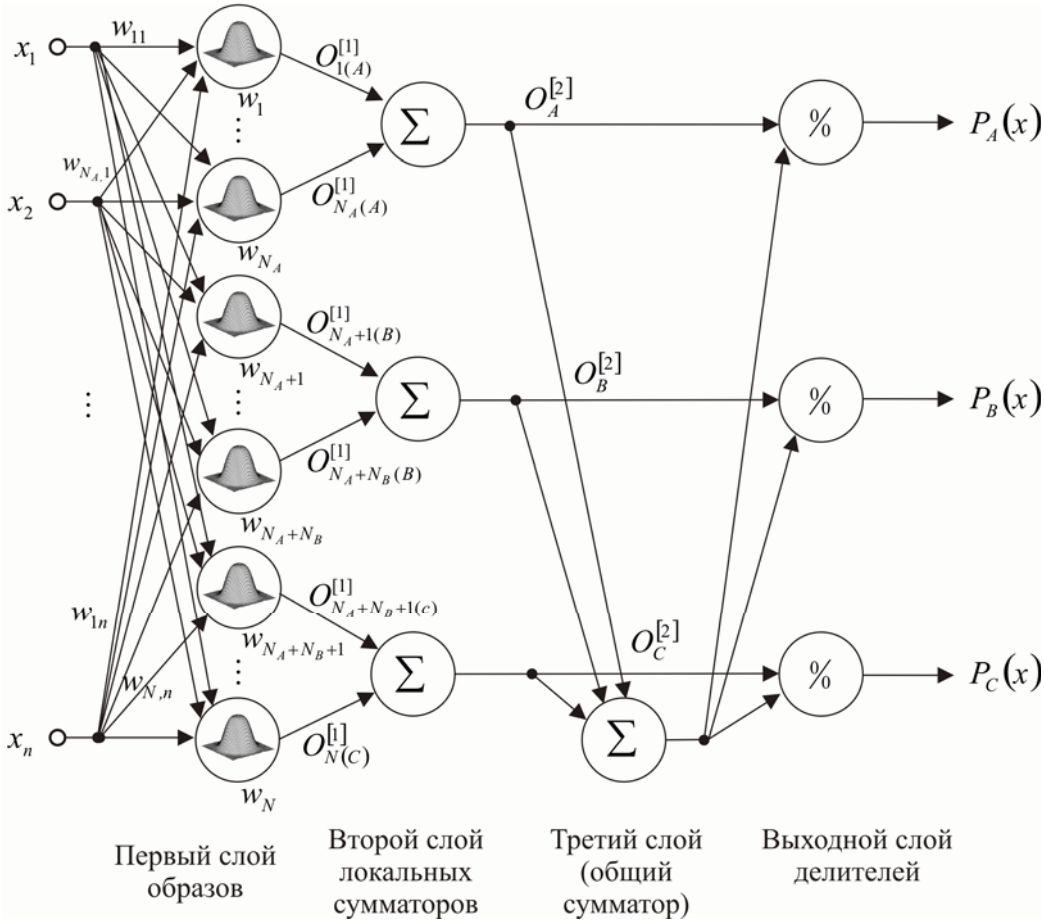

Рисунок 1 – Модифицированная вероятностная нейронная сеть

Исходной информацией для синтеза сети является обучающая выборка, образованная «пакетом» n-мерных векторов $x(1), x(2), \ldots, \mathbf{x}(\mathbf{N})$ с известной классификацией, причем место кон-

кретного образа в пакете значения не имеет. Предполагается также, что все входные векторы пронормированы так, что $\|x(j)\| = 1, j = 1,2,\ldots N$, а сами образы (без потери общности) могут принадлежать, к примеру, одному из трех классов A, B или C. Предполагается также, что N_A образов относятся к классу A, N_B – к классу B и N_C – к классу C, т.е. $N_A + N_B + N_C = N$.

Количество нейронов в слое образов принимается равным N (по одному нейрону на каждый обучающий образ), а их параметры (центры активационных функций) определяются значениями компонент входных векторов так, что

$$w_{ji} = x_i(j), j = 1,2,\ldots N; i = 1,2,\ldots n, \quad (1)$$

или в векторной форме

$$w_j = x(j), j = (x_1(j), x_2(j), \ldots, x_n(j))^T. \quad (2)$$

Таким образом, в этой сети реализуется обучение, основанное на памяти [6], по принципу «нейроны в точках данных» [7], что делает его крайне простым и практически мгновенным.

Каждый из нейронов слоя образов вычисляет взвешенную сумму компонент входных сигналов и преобразует ее с помощью нелинейной активационной функции так, что на выходе нейронов первого скрытого слоя появляется сигнал в форме

$$O_j^{[1]}(k) = \exp\left(-\frac{\|x(k) - w_j\|^2}{2\sigma^2}\right), \quad (3)$$

где $j=1_{(A)}, 2_{(A)},\ldots,N_{A(A)},\ldots,N_A+1_{(B)},\ldots,N_A+N_{B(B)},N_A+N_B+1_{(C)},\ldots,N_{(C)}$, $k=N+1, N+2\ldots$ – индекс (номер) наблюдения, не принадлежащего обучающей выборке, $_{(R)}$ – индекс класса, R принимает значение A или, B или C, σ – параметр ширины ядерной активационной функции.

Заметим, что с учетом нормировки, выражение (3) можно переписать в более удобной форме

$$O_j^{[1]}(k) = \exp\left(\frac{w_j^T x(k) - 1}{\sigma^2}\right) = \exp\left(\frac{\cos(w_j, x(k)) - 1}{\sigma^2}\right), \quad (4)$$

при этом, поскольку $-1 \le \cos(w_j, x(k)) \le 1$, аргумент (3) может изменяться в интервале

 ISSN 1562-9945

$$-\frac{2}{\sigma^2} \le \frac{\cos\left(w_j, x(k)\right) - 1}{\sigma^2} \le 0, \qquad (5)$$

а выходной сигнал каждого нейрона –

$$\exp\left(-2\sigma^{-2}\right) \le O_j^{[1]}(k) \le 1. \qquad (6)$$

Второй скрытый слой локальных сумматоров (по одному на каждый класс) вычисляет сумму выходов первого слоя в виде

$$\begin{aligned} 0 < O_A^{[2]}(k) &= \sum_{j=1_{(A)}}^{N_{A(A)}} O_j^{[1]}(k) < N_A, \\ 0 < O_B^{[2]}(k) &= \sum_{j=N_A+1_{(B)}}^{N_A+N_{B(B)}} O_j^{[1]}(k) < N_B, \\ 0 < O_C^{[2]}(k) &= \sum_{j=N_A+N_B+1_{(C)}}^{N_{(C)}} O_j^{[1]}(k) < N, \end{aligned} \qquad (7)$$

которые затем подаются на входы общего сумматора третьего слоя, вычисляющего сумму $O_A^{[2]}(k) + O_B^{[2]}(k) + O_C^{[2]}(k)$, и входы делимого делителей выходного слоя. Поскольку выходной сигнал третьего слоя подается на входы делителей выходного слоя, на выходах сети появляются значения вероятностей

$$\begin{aligned} 0 < P_A\left(x(k)\right) &= \frac{O_A^{[2]}(k)}{O_A^{[2]}(k) + O_B^{[2]}(k) + O_C^{[2]}(k)} < 1, \\ 0 < P_B\left(x(k)\right) &= \frac{O_B^{[2]}(k)}{O_A^{[2]}(k) + O_B^{[2]}(k) + O_C^{[2]}(k)} < 1, \\ 0 < P_C\left(x(k)\right) &= \frac{O_C^{[2]}(k)}{O_A^{[2]}(k) + O_B^{[2]}(k) + O_C^{[2]}(k)} < 1. \end{aligned} \qquad (8)$$

После того, как сеть построена, необходимо задать значение параметра ширины σ, который для нормированных входов выбирается достаточно произвольно в интервале от нуля до единицы [8]. Вместе с тем, следует отметить, что простого формального решения, позволяющего получить значение этого параметра, на сегодня не существует. И, наконец, можно приступать к решению задачи собственно классификации, предъявляя MPNN образы $x(k), k > N$ с неизвестной принадлежностью.

4. Результаты экспериментальных исследований

Предложенный метод семантического аннотирования текстовых документов, основой которого является MPNN, тестировался на

выборке текстов, принадлежащих к разным онтологическим классам (50 признаков, 100 объектов). В качестве исходных данных рассматривался как корпус текстов “20 Newsgroups DataSet” (comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware), так и обработанная текстовая информация узкоспециализированной предметной области. Преобразование корпуса текстов в векторное пространство осуществлялось на основе статистической оценки TFIDF.

В ходе эксперимента рассматривалось, прежде всего, качество работы MPNN. В табл. 1 представлены результаты работы со значением $\sigma = 0.3$. Показано, что в результате работы формируется набор значений вероятностей принадлежности входного текстового объекта к нескольким классам, которые рассматриваются как концепты онтологии предметной области. Также проведено сравнение работы предложенного метода на основе MPNN для различных значений ширины параметра σ. Для тестирования при проведении экспериментов использовалось 60% исходных данных из обучающей выборки.

Таблица 1

Пример работы программы

№ текстового объекта	Вероятность принадлежности к 1-му классу	Вероятность принадлежности ко 2-му классу	Вероятность принадлежности к 3-му классу
1	0,99964	0,00010032	0,00025828
2	0,67091	0,094663	0,23443
3	0,02427	0,95577	0,019964
4	0,092507	0,88085	0,02664
5	0,32385	0,62807	0,048079

В ходе экспериментальных исследований было установлено, что предложенный метод обладает высокими показатели точности и быстротой работы, это дает возможность повысить качество извлечения знаний из текстовых источников при ограниченной выборке.

5. Выводы

В данной работе предложен метод семантического аннотирования, в основе которого лежит модифицированная вероятностная нейронная сеть, которая представляет собой гибрид стандартной PNN и GRNN. Благодаря этому, возможно определить вероятности принадлежности входящего текстового объекта к каждому из потенциально возможных классов онтологии предметной области для формирования семантических аннотаций. Данный метод предусматривает возмож-

 ISSN 1562-9945

ность обработки информации по мере ее поступления в последовательном режиме, характеризуется простотой реализации и высокой скоростью обработки информации.

ЛИТЕРАТУРА

1. Uren V. Semantic annotation for knowledge management: Requirements and a survey of the state of the art [Текст] / V. Uren, Ph. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, F. Ciravegna // Web Semantics: Science, Services and Agents on the World Wide Web. – 2006. – Vol. 4, No. 1. – P. 14–28.

Dublin Core: Metadata Initiative [Электронный ресурс]. – Режим доступа: www.dublincore.org – 01.03.2011 г. – Загл. с экрана.

2. Бодянский Е.В. Семантическое аннотирование текстовых документов на основе иерархической радиально-базисной нейронной сети [Текст] / Е.В. Бодянский, О.В. Шубкина // Восточно-Европейский журнал передовых технологий. – 2010. – Вып. 6/3 (48). – С. 72–77.
3. Specht D.F. Probabilistic neural networks [Текст] / D.F. Specht // Neural Networks. – 1990. – 3. – P. 109–118.
4. Specht D.F. A general regression neural network [Текст] / D.F. Specht // IEEE Trans. on Neural Networks. – 1991. – 2. – P. 568–576.
5. Nelles O. Nonlinear System Identification [Текст] / O. Nelles. – Berlin: Springer, 2001. – 785 p.
6. Zahirniak D.R. Pattern recognition using radial basis function network [Текст] / D.R. Zahirniak, R. Chapman, S.K. Rogers, B.W. Suter, M. Kabriski, V. Pyatti // Proc: 6-th Ann. Aerospace Application of AI Conf. – Dayton, OH, 1990. – P. 249–260.
7. Tsoukalas L.H. Fuzzy and Neural Approaches in Engineering [Текст] / L.H. Tsoukalas, R.E. Uhrig. – N.Y.: John Willey and Sons Inc., 1997. – 587 p.