УДК:519.25.004.8

МАТЕМАТИЧЕСКИЙ МЕТОД ОПИСАНИЯ НЕЙРОННЫХ СЕТЕЙ ДЛЯ КЛАССИФИКАЦИИ СВОЙСТВ ТЕКСТОВ

А.А. Косимов, Н.М. Курбонов

Таджикский технический университет имени академика М.С. Осими

В данной статье представлен метод классификации текстов с использованием нейронных сетей. Рассматриваются подходы на основе встраивания слов (word embeddings) и различных архитектур нейросетей. Описываются математические основы метода, включая преобразование текста в векторное представление, процесс классификации через полносвязный слой и вычисление вероятностей классов с помощью функции softmax. Представленный метод демонстрирует высокую адаптивность и применимость в задачах анализа тональности, тематической классификации и других направлениях обработки естественного языка.

Ключевые слова: *классификация, текст, нейрон, сеть, встраивания, softmax, функция, потеря, обработка, естественный.*

УСУЛИ МАТЕМАТИКИИ ТАВСИФИ ШАБАКАҲОИ НЕЙРОНӢ БАРОИ ТАСНИФИ ХОСИЯТҲОИ МАТНҲО

А.А. Косимов, Н.М. Курбонов

Мақолаи мазкур усули таснифи матнҳоро бо истифода аз шабакаҳои нейронӣ пешниҳод мекунад. Усулҳои воридкунии калимаҳо (word embeddings) ва архитектураи гуногуни шабакаҳои нейронӣ баррасӣ мешаванд. Асосҳои усули математикӣ, аз ҷумла табдили матн ба намояндагии векторӣ, раванди тасниф тавассути қабати пурра пайваст ва ҳисоб кардани эҳтимолияти синфҳо бо истифода аз функсияи softmax тавсиф карда шудаанд. Усули пешниҳодшуда мутобиқшавӣ ва татбиқи баландро дар вазифаҳои таҳлили эҳсосот, таснифоти мавзӯӣ ва дигар соҳаҳои коркарди забони табиӣ нишон медиҳад.

Калидвожаҳо: *тасниф, матн, нейрон, шабака, embeddings, softmax, функсия, талафот, коркард, табиӣ.*

MATHEMATICAL METHOD OF DESCRIPTION OF NEURAL NETWORKS FOR CLASSIFICATION OF TEXTS PROPERTIES

A.A. Kosimov, N.M. Kurbonov

This paper presents a method for text classification using neural networks. Word embeddings and various neural network architectures are considered. The mathematical foundations of the method are described, including the transformation of text into a vector representation, the classification process through a fully connected layer, and the calculation of class probabilities using softmax functions.. The presented method demonstrates high adaptability and applicability to sentiment analysis, topic classification, and other areas of natural language processing.

Keywords: *classification, text, neuron, network, embeddings, softmax, function, loss, processing, natural.*

Введение

В современном мире объем текстовой информации постоянно растет, что делает автоматическую обработку текстов все более актуальной задачей [1-18]. Одним из ключевых направлений в этой области является классификация текстов, позволяющая определять их принадлежность к определённым категориям, например, в задачах анализа тональности, тематического моделирования и фильтрации контента.

Традиционные методы классификации, основанные на статистическом анализе и ручном выделении признаков, часто оказываются недостаточно эффективными при работе с большими объемами данных и сложной текстовой структурой. В связи с этим все большее распространение получают методы, основанные на нейронных сетях.

В статье рассматривается метод классификации текстов, основанный на встраивании слов в векторное пространство и обработке их с помощью различных архитектур нейронных сетей. Представленный подход позволяет автоматически извлекать значимые признаки текста, обеспечивая высокую точность и адаптивность модели к различным задачам обработки естественного языка.

Для классификации текстов с помощью нейронной сети мы можем использовать подход на основе **встраиваний слов (word embeddings)** и **многослойной нейронной сети (MLP).** Предположим, что имеется текст и необходимо классифицировать его в одну из C категорий, например, "позитивный", "негативный", "нейтральный".

Пусть задан входной текст в следующим виде:

Входной текст: T={w_1, w_2,,…, w_n,}, где w_i, — это слова текста, а n — количество слов в тексте.

Встраивание слов: Каждое слово w_i представляется в виде вектора $v_i \in R^d$ где d — размерность пространства встраивания, R^d- пространство вещественных векторов.

Эти векторы обучаются таким образом, чтобы сохранять смысловую близость между словами: слова с похожими значениями располагаются ближе друг к другу в пространстве встраивания.

Другими словами, каждое слово w_i кодируется как вектор v_i, состоящий из d чисел (вещественных значений).

Популярные методы встраивания слов:

Word2Vec (CBOW, Skip-gram)

GloVe (Global Vectors for Word Representation)

FastText (учитывает морфологию слов)

BERT embeddings (контекстуальные представления слов)

1. **Агрегация векторных представлений:** Объединяем вектора слов в одно представление текста (например, используя усреднение или рекуррентные слои).

2. Классификация: На выходе сеть предсказывает вероятность принадлежности текста к каждой категории.

Математическое описание

Встраивание слов

Каждое слово $\boldsymbol{w_i}$ преобразуется в вектор с использованием таблицы встраивании:

$$\boldsymbol{v_i}=\mathbf{Embedding}(\boldsymbol{w_i})$$

где

Embedding— это матрица $E \in R^{|V|*d}$, где |V| — размер словаря, а d размерность встраивании.

2. Представление текста

Общее представление текста h вычисляется с использованием одной из следующих стратегий:

Усреднение (Bag of Words):

$$\mathrm{h} = \sum_{\mathbf{i=0}}^{\mathbf{n}} \mathbf{v_i}$$

- **1.Рекуррентная сеть (RNN)**: Каждое слово передаётся через рекуррентную ячейку:

$$\boldsymbol{h_i} = \mathbf{RNN}\ (\boldsymbol{v_i}, \boldsymbol{h_{i-1}})$$

Итоговое представление текста:

$$\mathrm{h} = h_n$$

После обработки всей последовательности (допустим, текст из **n** слов) последнее скрытое состояние $\boldsymbol{h_n}$ содержит агрегированную информацию обо всём тексте.

- **2.Сверточная сеть (CNN):**

$$\boldsymbol{h_i}=\mathbf{f}(\mathbf{W}*\boldsymbol{v_{i:i+k-1}} + \boldsymbol{b})$$

где W – фильтр, k – размер окна, f – нелинейная функция активации (например, ReLU).

- **3.Классификация**

Итоговое представление h передаётся в полносвязный слой с C выходами:

$$\mathbf{o}=\boldsymbol{W^{(0)}h} + \boldsymbol{b^{(0)}}$$

где

$\boldsymbol{W^{(0)}} \in R^{C*d}$— веса полносвязного слоя,

$\boldsymbol{b^{(0)}} \in R^{C}$— смещение.

- **4.Вероятности классов**

Для вычисления вероятностей используется **softmax**:

$$\mathbf{P(c|T)} = \frac{\mathbf{exp}(\boldsymbol{o_c})}{\sum_{\mathbf{j=1}}^{\mathbf{C}} \boldsymbol{exp}(\mathbf{o_j})}$$

где

$\boldsymbol{o_c}$— выход сети для класса c,

C — количество классов.

- **5.Функция потерь**

Для обучения модели используется **кросс-энтропийная функция потерь**:

$$\mathbf{L}= \sum_{i=1}^{N} \sum_{c=1}^{C} \boldsymbol{y_{i,c}} \boldsymbol{log} \boldsymbol{P}(\boldsymbol{c}|\boldsymbol{T_i})$$

где

$\boldsymbol{y_{i,c}}$ — истинная метка класса для текста T_i,

$\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{T_i})$ — предсказанная вероятность.

Пример с числами

Пусть задан текст: "Пример текста". Словарь V содержит слова "Пример", "текста" и ещё |V|−2 слов. Размерность встраивании d=5.

1.Встраивания слов: Каждое слово из словаря V имеет свое векторное представление $v_{слово} \in R^5$

Например:

$$\boldsymbol{V}_{\text{Пример}} = [0.1,0.2,0.3,0.4,0.5]$$

$$\boldsymbol{V}_{\text{текста}} = [0.5,0.4,0.3,0.2,0.1]$$

2.Усреднённое представление: Один из способов представить весь текст как один вектор — усреднить векторы слов:

$$\mathbf{h} = \frac{\boldsymbol{V}_{\text{Пример}} + \boldsymbol{V}_{\text{текста}}}{2}$$

Посчитаем поэлементно:

$$\mathbf{h} = \frac{[0.1,0.2,0.3,0.4,0.5]+[0.5,0.4,0.3,0.2,0.1]}{2}$$

$$\mathbf{h} = \frac{[0.6,0.6,0.6,0.6,0.6]}{2} = [0.3,0.3,0.3,0.3,0.3]$$

3.Полносвязный слой: Пусть веса и смещения:

$$\boldsymbol{W}^{(0)} = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 & 0.5 \\ 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \end{bmatrix} \qquad \boldsymbol{B}^{(i)}=[\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}],\ i=1,2$$

Выходное представление **O** считается по формуле:

$$\mathbf{O} = \boldsymbol{W}^{(0)}\mathbf{h}+\boldsymbol{b}^{(0)}$$

тогда

$$\mathbf{O} = \boldsymbol{W}^{(0)}\mathbf{h}+\boldsymbol{b}^{(0)} = \begin{bmatrix} 0.3*0.1+0.3*0.2+0.3*0.3+0.3*0.4+0.3*0.5+0.1 \\ 0.3*0.5+0.3*0.4+0.3*0.3+0.3*0.2+0.3*0.1+0.2 \end{bmatrix}$$

Посчитаем их:

1.Для первого компонента O_1:

$$\mathbf{O}_1 = 0.3*0.1+0.3*0.2+0.3*0.3+0.3*0.4+0.3*0.5+0.1 = 0.55$$

2.Для второго компоненты O_2:

$$\mathbf{O}_2 = 0.3*0.5+0.3*0.4+0.3*0.3+0.3*0.2+0.3*0.1+0.2 = 0.65$$

Таким образом, получаем:

$$\mathbf{O}=[0.55, 0.65]$$

Таким образом, шаг «Пропуск через полносвязный слой» является ключевым этапом, на котором усреднённое представление текста преобразуется в **логиты** — числовые значения, отражающие степень соответствия текста каждому из возможных классов. Это позволяет модели принимать окончательное решение о классификации текста.

4.Softmax: Предсказанные вероятности классов: После получения логитов из полносвязного слоя, применяется функция softmax, которая преобразует эти значения в вероятности. Это позволяет интерпретировать выход модели как вероятностное распределение по классам.

Пусть $\mathbf{o}_1$ и $\mathbf{o}_2$ — логиты (выходы полносвязного слоя) для классов $\boldsymbol{c}_1$ и $\boldsymbol{c}_2$ соответственно. Тогда вероятности классов вычисляются по следующим формулам:

$$\mathbf{P}(\boldsymbol{c}_1) = \frac{exp(o_1)}{exp(o_1+exp(o_2)}, \mathbf{P}(\boldsymbol{c}_2) = \frac{exp(o_2)}{exp(o_1+exp(o_2)},$$

где

$\boldsymbol{exp}(\cdot)$ — экспоненциальная функция (основание $\boldsymbol{e}$),

$\mathrm{P}(\boldsymbol{c}_1)$ и $\mathrm{P}(\boldsymbol{c}_2)$ — вероятности принадлежности к соответствующим классам.

Подставим значения:

$$\mathbf{P}(\boldsymbol{c}_1) = \frac{e^{0.55}}{e^{0.55}+e^{0.65}}, \quad \mathbf{P}(\boldsymbol{c}_2) = \frac{e^{0.65}}{e^{0.55}+e^{0.65}}$$

имеем

$$e^{0.55} \approx 1.733, \qquad e^{0.65} \approx 1.915$$

$$\mathbf{P}(\boldsymbol{c}_1) = \frac{1.733}{1.733+1.915} \approx 0.475, \quad \mathbf{P}(\boldsymbol{c}_2) = \frac{1.915}{1.733+1.915} \approx 0.525$$

Итак, вероятность класса $\boldsymbol{c}_2$ чуть выше, чем у $\boldsymbol{c}_1$.

После получения предсказанных вероятностей $\mathbf{P}(\boldsymbol{c}_1)$ и $\mathbf{P}(\boldsymbol{c}_2)$, следующий шаг зависит от контекста (например, обучения или предсказания). Рассмотрим оба случая:

1. Если находимся на этапе обучения

Вычисляем функцию потерь (кросс-энтропию) и обновляем параметры модели с помощью градиентного спуска.

Вычисление функции потерь

Допустим, истинная метка для данного примера y=(1,0), то есть текст относится к классу c_1. Тогда функция потерь (кросс-энтропия) равна:

$$\mathbf{L = -(y_1 \log P(c_1) \ + y_2 \log P(c_2) \ )}$$

Подставляя значения, имеем:

$$\mathbf{L = -(1 \cdot \log(0.475) + 0 \cdot \log(0.525)) \approx -(-0.744) = 0.744}$$

Обновление параметров модели

Вычисляются **градиенты** функции потерь по параметрам модели (весам $W^{(0)}$ и смещениям $b^{(0)}$.
Применяется **алгоритм оптимизации** — градиентный спуск или его модификации (например, **Adam**, **RMSProp**).
Параметры корректируются так, чтобы минимизировать функцию потерь.
Этот процесс повторяется для всех примеров в обучающем наборе данных.

2. Если мы находимся на этапе предсказания

- Модель вычисляет вероятности классов на основе входного текста.
- Выбирается класс с наибольшей вероятностью:

$$\mathbf{P(c_1)} = 0.475, \ \mathbf{P(c_2)} = 0.525$$

Так как $P(c_2) > P(c_1)$ **,** модель относит текст к классу c_2.

В данной статье был рассмотрен подход к классификации текстов, основанный на встраивании слов в векторное пространство и их последующей обработке нейронными сетями. Преобразование слов в векторы позволяет сохранить семантическую информацию. Такой подход позволяет эффективно анализировать тексты и определять их принадлежность к определённым категориям.

Рассмотренный метод демонстрирует высокую гибкость и применимость в различных задачах обработки естественного языка, включая анализ тональности, тематическую классификацию и фильтрацию контента. Использование softmax-функции для получения вероятностей и кросс-энтропийной функции потерь для обучения модели делает процесс классификации интерпретируемым и эффективно настраиваемым.

Показанный числовой пример иллюстрирует работу модели на практике и подтверждает её способность различать классы на основе смыслового содержания текста. Таким образом, подход на основе встраиваний слов и нейронных сетей представляет собой мощный инструмент для анализа текстовых данных, способный обеспечить высокую точность и адаптивность в реальных приложениях.

Рецензент: Саидов Б.Б. – к.т.н., ТТУ имени академика М.С. Осими.

Литература

1. Усманов З.Д. Классификатор дискретных случайных величин. – ДАН РТ, 2017, т.60, № 7-8, с. 291-300.

2. Усманов З.Д. Алгоритм настройки кластеризатора дискретных случайных величин. – ДАН РТ, 2017, т.60, № 9, с. 392-397.

3. Курбонов Н.М. Об автоматическом распознавании на основе униграмм шифров авторефератов по педагогике // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. 2021. № 3 (55). С. 47-51.

4. Усманов З.Д. N-граммы в распознавании однородных текстов. Материалы 20 научно-практического семинара "Новые информационные технологии в автоматизированных системах", Москва 2017, № 20. С. 52-54.

5. Косимов А.А. Оценка эффективности использования униграмм при идентификации текста. Доклады Академии наук Республики Таджикистан. - 2017. - Том 60. - № 3-4. - С. 132-137.

6. Goodfellow, I., Bengio, Y., & Courville, A. *Deep Learning*. MIT Press.. (2016).

7. Косимов А.А. Оценка эффективности использования триграмм при идентификации текста Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. - 2017. - № 1 (166). - С. 51-57.

8. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.

9. Косимов А.А. О минимальном числе высокоточных n-грамм, необходимых для распознавания автора текста. Российско-китайский научный журнал «Содружество», Ежемесячный научный журнал, научно-практической конференции. - 2017. - Часть 1. - № 17. - С. 58-59.

10. Усманов З.Д., Косимов А.А. О метризации произведений художественной литературы. Материалы 21 научно-практического семинара "Новые информационные технологии в автоматизированных системах", Москва 2018, № 21, С.183-186.

11. Назаров А.Ш., Ли И.Т., Курбонов Н.М. Моделирование системы защиты информации от угроз // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. 2019. № 1 (45). С. 10-12.

12. Усманов З.Д., Косимов А.А. О влиянии цифрового портрета текста на распознавание автора произведения. Известия АН РТ, Отделение физ.-мат., хим., геол. и техн. наук. – 2020. - № 3 (180) – С.36-42.

13. Усманов З.Д. Оценка эффективности применения -классификатора для атрибуции печатного текста // ДАН РТ - 2020.- Т.63, № 3-4 – С.172-179.

14. Усманов З.Д., Косимов А.А. Об автоматическом распознавании языка произведений // Доклады Академии наук Республики Таджикистан, 2020, т.63, № 7-8, с. 461-466.

15. Косимов А.А., Курбонов Н.М., Муродов Х.М., Зулфов Е.О. Построение структуры однородностей поэм произведения а. фирдоуси “шахнаме” на основе биграмм // Вестник ПИТТУ имени академика М.С. Осими. 2022. № 3 (24). С. 22-28.

16. Косимов А.А., Курбонов Н.М. Структура однородностей поэм произведения а.фирдоуси “шахнаме” // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. 2021. № 2 (54). С. 35-38.

17. Усманов З.Д., Косимов А.А., Каюмов М.М. База данных αβ-кодов словоформ для определения автора незнакомого текста // Свидетельство о государственной регистрации информационного ресурса, Республика Таджикистан, 07.06.2021, №1202100478.

18. Прогнозирование параметров решетки перовскитных материалов с использованием методов машинного обучения / М. М. Каюмов, А. С. Бурхонзода, Д. Д. Нематов [и др.] // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2024. – № 3(67). – С. 49-52. – EDN NUVAJY.

СВЕДЕНИЯ ОБ АВТОРАХ – МАЪЛУМОТ ДАР БОРАИ МУАЛЛИФОН – INFORMATION ABOUT THE AUTHORS

RU	TJ	EN
Косимов Абдунаби Абдурауфович	Қосимов Абдунаби Абдурауфович	Kosimov Abdunabi Abduraufovich
Доктор технических наук, доцент	Доктори илмҳои техникӣ, дотсент	Doctor of Technical Sciences, associate professor
Таджикский технический университет имени академика М.С. Осими	Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ	Tajik technical university named after academician M.S. Osimi
E-mail: abdunabi_kbtut@mail.ru		

RU	TJ	EN
Курбонов Нурулло Мирзомахмудович	Қурбонов Нурулло Мирзомаҳмудович	Kurbonov Nurullo Mirzomahmudovich
докторант Ph.D	докторанти Ph.D	Ph.D. student
Таджикский технический университет имени академика М.С. Осими	Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ	Tajik technical university named after academician M.S. Osimi
E-mail: nurullo94@gmail.com		