83.77.23

Применение искусственных нейронных сетей для решения задач классификации при обработке научных текстов (на примере Weka)

Шмалько Ю.В.
Крымский федеральный университет имени В.И. Вернадского (КФУ)
г. Симферополь, Российская Федерация
ORCID: https://orcid.org/0000-0001-9760-5839
e-mail: uliasmalko73543@gmail.com

С появлением технологий глубокого обучения и их применения в обработке естественного языка было сделано улучшение точности этих методов в двух основных направлениях: использование нейронной сети с учителем для обучения классификатора и без учителя для оптимизации предварительной обработки данных и выбора характеристик. За последние несколько лет нейронные сети вновь появились в качестве мощных моделей машинного обучения, показали лучшие результаты в таких областях, как распознавание образов и обработки речи. Еще совсем недавно нейросетевые модели начали применяться также к различным задачам обработки естественного языка с очень хорошими результатами. Исследование предполагает рассмотрение метода обучения нейронной сети с учителем для классификации научных статей по принадлежности к тем или иным научным журналам.

Ключевые слова: искусственные нейронные сети, научный текст, машинное обучение, классификация.

Для цитаты: *Шмалько Ю.В.* Применение искусственных нейронных сетей для решения задач классификации при обработке научных текстов // Цифровая гуманитаристика и технологии в образовании (DHTE 2023): сб. статей IV Международной научно-практической конференции. 16–17 ноября 2023 г. / Под ред. В.В. Рубцова, М.Г. Сороковой, Н.П. Радчиковой. М.: Издательство ФГБОУ ВО МГППУ, 2023. 591–596 с.

Введение

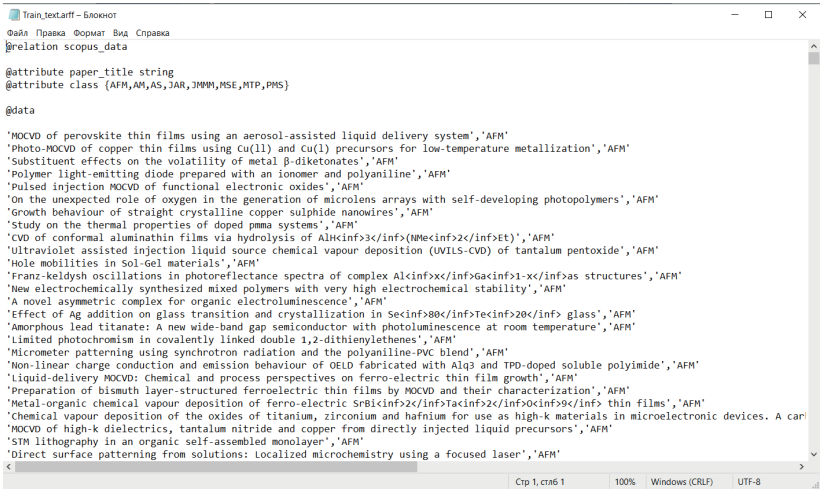
Нейронные сети относятся к направлению искусственного интеллекта (ИИ) и применяются для распознавания скрытых закономерностей в необработанных данных, группировки и классификации, а также решения задач в области ИИ, машинного и глубокого обучения. В частности, нейронные сети могут использоваться для решения задач классификации при обработке научных текстов. Наше исследование предполагает рассмотрение метода обучения

нейронной сети с учителем для классификации научных статей по принадлежности к тем или иным научным журналам.

Целью работы является изучение работы свободного программного обеспечения Weka при обработке научных статей физико-технического направления, анализ полученных результатов и выявление качественных и количественных показателей эксперимента.

Методы

Для работы с Weka (рис. 1.) был подготовлен некоторый модельный файл. Был взят набор наименований англоязычных статей из журналов, индексируемых в Scopus. Все журналы были на разные темы. Это проводилось с целью наблюдения за качеством обучаемости нейросети. Первый этап заключался в объединении статей из разных журналов в один файл (Train_text) для обучения нейронной сети. Журналы заведомо были определенной специфики. Работа проводится с *.arff файлами, которые поддерживает программа Weka. Задача состоит в том, чтобы обучить нейросеть определять по названию к какому журналу относится статья. Второй этап состоял в проверке работоспособности обученной нейронной сети, для чего был создан тестовый файл, содержащий 40 наименований статей из различных журналов. Третий этап был нацелен на анализ экспериментальных данных, полученных в ходе второго этапа.


Train_text.arff – Блокнот

Файл Правка Формат Вид Справка

```
@relation scopus_data

@attribute paper_title string
@attribute class {AFM,AM,AS,JAR,JMMM,MSE,MTP,PMS}

@data

'MOCVD of perovskite thin films using an aerosol-assisted liquid delivery system','AFM'
'Photo-MOCVD of copper thin films using Cu(1l) and Cu(1) precursors for low-temperature metallization','AFM'
'Substituent effects on the volatility of metal β-diketonates','AFM'
'Polymer light-emitting diode prepared with an ionomer and polyaniline','AFM'
'Pulsed injection MOCVD of functional electronic oxides','AFM'
'On the unexpected role of oxygen in the generation of microlens arrays with self-developing photopolymers','AFM'
'Growth behaviour of straight crystalline copper sulphide nanowires','AFM'
'Study on the thermal properties of doped pmma systems','AFM'
'CVD of conformal aluminathin films via hydrolysis of Al<inf>3</inf>(NMe<inf>2</inf>Et)','AFM'
'Ultraviolet assisted injection liquid source chemical vapour deposition (UVILS-CVD) of tantalum pentoxide','AFM'
'Hole mobilities in Sol-Gel materials','AFM'
'Franz-keldysh oscillations in photoreflectance spectra of complex Al<inf>x</inf>Ga<inf>1-x</inf>as structures','AFM'
'New electrochemically synthesized mixed polymers with very high electrochemical stability','AFM'
'A novel asymmetric complex for organic electroluminescence','AFM'
'Effect of Ag addition on glass transition and crystallization in Se<inf>80</inf>Te<inf>20</inf> glass','AFM'
'Amorphous lead titanate: A new wide-band gap semiconductor with photoluminescence at room temperature','AFM'
'Limited photochromism in covalently linked double 1,2-dithienylethenes','AFM'
'Micrometer patterning using synchrotron radiation and the polyaniline-PVC blend','AFM'
'Non-linear charge conduction and emission behaviour of OELD fabricated with Alq3 and TPD-doped soluble polyimide','AFM'
'Liquid-delivery MOCVD: Chemical and process perspectives on ferro-electric thin film growth','AFM'
'Preparation of bismuth layer-structured ferroelectric thin films by MOCVD and their characterization','AFM'
'Metal-organic chemical vapour deposition of ferro-electric SrBi<inf>2</inf>Ta<inf>2</inf>O<inf>9</inf> thin films','AFM'
'Chemical vapour deposition of the oxides of titanium, zirconium and hafnium for use as high-k materials in microelectronic devices. A car
'MOCVD of high-k dielectrics, tantalum nitride and copper from directly injected liquid precursors','AFM'
'STM lithography in an organic self-assembled monolayer','AFM'
'Direct surface patterning from solutions: Localized microchemistry using a focused laser','AFM'
```

Стр 1, стлб 1 | 100% | Windows (CRLF) | UTF-8

Рис. 1. Train_test файл для обучения нейронной сети

Начало *.arff файла, открытого при помощи Notepad, приведено на рисунке. Фактически у нас есть размеченный файл для обучения ИНС.

Тестовый файл содержит названия более 16000 статей физико-технической направленности из восьми известных научных журналов. Каждая строка тренировочного *.arff файла содержит название статьи и, через запятую, класс, к которому она относится, то есть название журнала. Для удобства, вместо полного названия журналов использовались аббревиатуры. Далее, подготовленный тренировочный файл необходимо загрузить в Weka Explorer

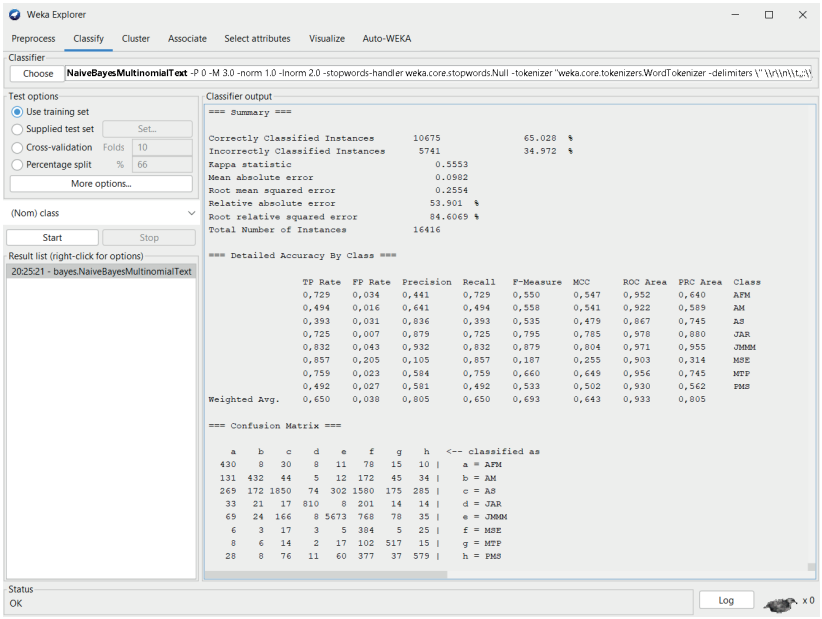


Рис. 2. Отчет об обучении искусственной нейронной сети для решения задачи классификации научных текстов с использованием тренировочного файла

По завершении процесса обучения, программа Weka выдаёт отчёт о работе с тренировочным набором данных (рис. 2.). Отчет содержит общую информацию об успешности классификации, детализированную информацию о весовых коэффициентах и матрицу путаницы (Confusion Matrix). Правильно классифицированными по названию оказались 65 % научных статей, остальные 35 % были отнесены к неверным журналам.

Для проверки работоспособности обученной нейронной сети был подготовлен тестовый файл (test_text), содержащий 40 наименований статей из различных журналов, по пять на каждый участвовавший в обучении нейронной сети журнал (рис. 3.).

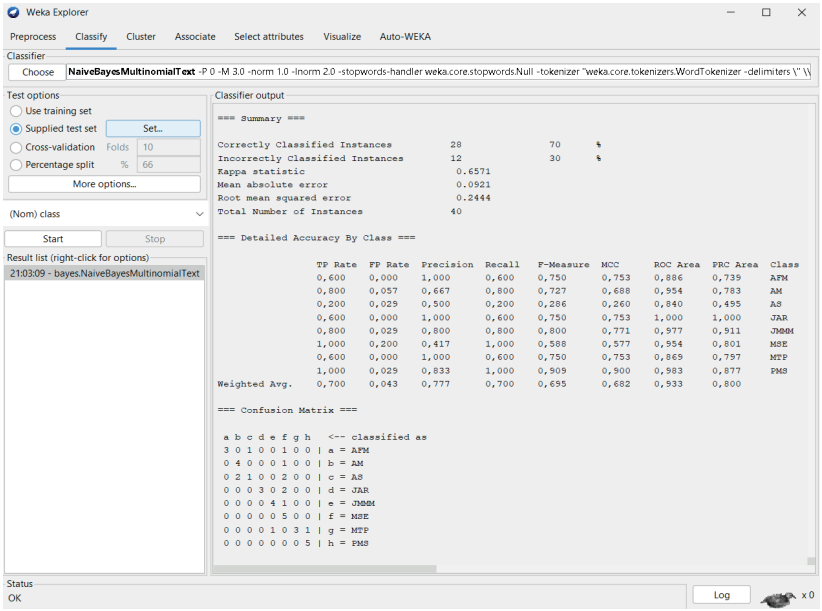


Рис. 3. Отчет о тестировании искусственной нейронной сети

Из рисунка 3 видно, что модель искусственной нейронной сети при обработке тестового файла определила принадлежность 70 % статей к правильным журналам, и лишь остальные 30 % статей были классифицированы неверно.

В целом, результат экспериментального исследования с моделью искусственной нейронной сети доказывает, что задача классификации научных текстов может быть успешно выполнена, и, хоть вероятность корректной классификации не достигает 100 %, но модель показывает достойный результат. При более детальной настройке и более широком наборе тренировочных данных, вероятность успешной классификации может быть значительно увеличена.

Обсуждение

В ходе данной работы было проведено исследование возможности использования свободного программного обеспечения Weka

для решения задач классификации научных текстов физико-технического направления. Для обучения искусственной нейронной сети был подготовлен тренировочный файл (Train_text.arff), содержащий более 16 000 наименований статей. Для каждой статьи тренировочного файла был определен атрибут – журнал, к которому относится эта статья. Вероятность успешной классификации на этапе обучения составила 65 %. Далее для проверки работоспособности обученной искусственной нейронной сети был подготовлен тестовый файл (test_text.arff), содержащий 40 новых научных статей из журналов, которые использовались при обучении. После применения обученной нейронной сети к тестовому набору данных были получены следующие результаты – процент успешно классифицированных научных статей составил 70 %, а остальные 30 % были отнесены к неверным журналам.

Такой показатель говорит о достаточно высокой эффективности применения искусственных нейронных сетей для решения задачи классификации научных статей по принадлежности к журналам. Возможность применения данного метода значительно упрощает процесс обработки научных текстов.

Информация об авторах

Шмалько Юлия Витальевна, студентка 2 курса магистратуры кафедры экспериментальной физики, Крымский федеральный университет им. В.И. Вернадского (КФУ), г. Симферополь, Российская Федерация, ORCID: https://orcid.org/0000-0001-9760-5839, e-mail: uliasmalko73543@gmail.com

The Use of Artificial Neural Networks for Solving Classification Problems in the Processing of Scientific Texts (Using the Example of Weka)

Yulia V. Shmalko
V.I. Vernadsky Crimean Federal University (KFU)
Simferopol, Russian Federation
ORCID: https://orcid.org/0000-0001-9760-5839
e-mail: uliasmalko73543@gmail.com

With the advent of deep learning technologies and their application in natural language processing, the accuracy of these methods has been improved in two main directions: using a neural network with a teacher to train a classifier and without a teacher to optimize data preprocessing and selection of characteristics. Over the past few years, neural networks have re-emerged as powerful machine learning models, and have shown better results in areas such as pattern recognition and speech processing. More recently, neural network models have also been applied to various natural language processing tasks with very good results. The study involves the consideration of the method of training a neural network with a teacher to classify scientific articles by belonging to one or another scientific journal.

Keywords: artificial neural networks, scientific text, machine learning, classification.

For citation: Shmalko Yu.V. The Use of Artificial Neural Networks for Solving Classification Problems in the Processing of Scientific Texts (Using the Example of Weka)// *Digital humanities and technologies in education (DHTE 2023): collection of articles of the IV International Scientific and Practical Conference. November 16–17, 2023* / Edited by V.V. Rubtsov, M.G. Sorokova, N.P. Radchikova. M.: Publishing House of the Moscow State Pedagogical University, 2023. 591–596 p.

Information about the authors
Yulia V. Shmalko, 2nd year student of the Master's degree of the Department of Experimental Physics, V.I. Vernadsky Crimean Federal University (KFU), Simferopol, Russian Federation, ORCID: https://orcid.org/0000-0001-9760-5839 , e-mail: uliasmalko73543@gmail.com