

МРНТИ 20.53

А.А.Шойынбек¹, Д.Н. Куанышбай¹, К.Т. Кожамет¹

¹Университет имени Сулеймана Демиреля
Каскелен, Казахстан

МОДЕЛИ И АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИЙ ВИДА ПРЕСТУПЛЕНИЯ

Аннотация. В статье рассмотрено два алгоритма, которые классифицируют преступления на классы. Алгоритм Support vector machines (SVM) и Логистическая регрессия приведены в сравнительном анализе по критерию «точность». Преступления выглядят в виде текстовых сообщений. Рассмотрены методы tf-idf и GloVe для преобразования текстовых сообщений в машинный вид. Данный процесс также называется векторизацией. Взята выборка данных, по преступлениям имеющая свыше одного миллиона фабул для обучения моделей с участием учителя. Описаны действия препроцессинга данных для улучшения итогового и объективно верного результата. С этой целью были произведены такие действия как нормализация, случайная ротация и разделение данных на тестовую и обучающую выборку. Путем сравнительного анализа была определена наиболее точно классифицирующая модель для классификации преступлений.

Ключевые слова: логистическая регрессия, метод вспомогательных векторных машин, классификация, классификация преступлений, текстовая классификация, машинное обучение, векторизация текста.

Annotation. The article considers two algorithms that classify crimes into classes. Algorithm Support for vector machines (SVM) and Logistic regression in comparative analysis using the «accuracy» criterion. Crimes look like text messages. Methods tf-idf and GloVe for converting text messages to machine views are considered. This process is also called vectorization. A sampling of data is taken, for crimes, for obtaining models with a teacher. The actions of data preprocessing to improve the final and objectively correct result are described. For this purpose, actions such as normalization, random rotation, and separation of data into a test and training sample were performed. By means of a comparative analysis was a well-known classification model for classifying crimes.

Keywords: logistic regression, method of auxiliary vector machines, classification, classification of crimes, text classification, machine learning, text vectorization.

Андатпа. Мақалада қылмыстарды сыныптарға жатқызатын екі алгоритм қарастырылған. Қолдау векторы машиналары (SVM) алгоритмі және логистикалық регрессия «дәлдік» өлшемі бойынша салыстырмалы талдауда ұсынылған. Қылмыстар мәтіндік хабарларға ұқсас. Мәтіндік хабарламаларды машиналық көріністерге түрлендіру үшін tf-idf және GloVe әдістері қарастырылады. Бұл процесс векторизация деп аталады. Оқу сабақтарын оқытушымен бір миллионнан астам сюжеттік сызықтары бар қылмыстық деректер жиналды. Соңғы және объективті түрде дұрыс нәтижені жақсарту үшін деректерді алдын-ала өңдеу әрекеттері сипатталады. Ол үшін қалыпты режим, деректерді кездейсоқ бұру және тестілеу және оқу үлгісіне бөлу сияқты іс-шаралар орындалды. Салыстырмалы талдаудың көмегімен қылмыстарды жіктеуге арналған ең дәл классификациялық үлгі анықталды.

Түйінді сөздер: логистикалық регрессия, қосалқы векторлық машиналардың әдісі, жіктелуі, қылмыстарды жіктеу, мәтінді жіктеу, машина жасау, мәтінді векторизациялау.

Введение

Каждое государство сталкивается с внутривластным проблемой как коррупция и превышение должностных инструкций должностными лицами. Часто в развивающихся странах сотрудники правоохранительных органов манипулируют текстовыми сообщениями – «фабула», чтобы скрыть сигнал о правонарушении или изменить его. В рамках программы «Цифровой Казахстан»[1], разрабатывается система, которая позволяет выявлять сотрудников правоохранительных органов, превышающие свои должностные полномочия и не допускать изменения в текстовое сообщение – «фабулу».

Автоматизация классификации видов преступлений является одним из барьеров общей разрабатываемой системы. Это решение позволит классифицировать преступления в соответствии с текстом сообщения, после чего оно будет зашифровано и автоматически отправлено в соответствующий государственный орган посредством block chain сообщений. Это означает, что далее фабула будет не доступна к изменению.

Цель статьи - определить подходящую классификационную модель преступлений путем сравнения двух моделей логистической регрессии и SVM. Выбор этих моделей определяется определенными факторами. SVM считается одной из лучших моделей, подходящих для задач классификации текста [2]. Логистическая регрессия считается

простейшей моделью классификационных задач [3]. Сравнение будет основываться на показателе точности.

Метод

В ходе работы предлагается разделить задачу на три части, которые могут повлиять на точность: предварительная обработка и подготовка данных, выбор метода путем векторизации текста и выбор модели классификации. Сбор данных проводился годами и был собран в правоохранительной базе, доступ к данным был предоставлен Генеральной прокуратурой после подписания меморандума NDA. В базе свыше 1 миллиона записей, уже классифицированных преступлений вручную, сотрудниками правоохранительных органов. На рисунке 1 показана структура исходных данных

ID	Status	Date	Time	Location	Description	Crime Type
1021693	0	01.01.18	187540031000011	1975	ПРОСИТ ПРИНЯТЬ МЕРЫ В ОТНОШЕНИИ ЛИЦА КОТОРЫЙ...	Кража
1021694	0	01.01.18	187541031000001	1975	[REDACTED]	Кража
1021695	0	01.01.18	187541031000002	1975	[REDACTED]	Кража
1021696	0	01.01.18	187541031000003	1975	[REDACTED]	Кража
1021697	0	01.01.18	187550031000001	1975	[REDACTED] ТОМ ЧТО 01...	Кража
1021698	0	01.01.18	187550031000002	1975	[REDACTED] ТОМ ЧТО 01.0...	Кража
1021699	0	01.01.18	187550031000003	1975	[REDACTED] ТОМ ЧТО 0...	Кража
1021700	0	01.01.18	187550031000004	1975	[REDACTED] ПРОСИТ ПРИНЯТЬ ...	Кража
1021701	0	01.01.18	187550031000005	1975	[REDACTED] ТОМ ЧТО 01...	Кража
1021702	0	01.01.18	187550031000006	1975	РАПОРТ ОКАЗЫВАЮЩИХСЯ УЛИЦА РОЗЫБАКИЕВА ...	Кража
1021703	0	01.01.18	187570031000001	1975	По заявлению гр [REDACTED] укр...	Грабёж
1021704	0	01.01.18	187570031000002	1975	По заявлению [REDACTED] ...	Кража
1021705	0	01.01.18	187570031000003	1975	По заявлению [REDACTED] прож. ...	Мошенничество
1021706	0	01.01.18	187580031000002	1975	мкр. Алтай-1, д.26, кафе "донер" кража.	Кража
1021707	0	01.01.18	187580031000005	1975	Репорт оу л-нт полиции [REDACTED]	Мошенничество
1021708	0	01.01.18	187580031000006	1975	По заявлению [REDACTED] прож. у...	Кража
1021709	0	01.01.18	187580031000007	1975	[REDACTED] завод "Квант" кража кабеля	Кража
1021710	0	01.01.18	187702031000001	1977	01.01.18г., в ЛОВД ст. Кандыгаш обратился Б...	Кража
1021711	0	01.01.18	187702031000002	1977	01.01.18г., в ЛОВД ст. Кандыгаш обратился ...	Кража
1021712	0	01.01.18	187702031000003	1977	01.01.18г. в ЛОВД на ст.Кандыгаш рапортом до...	Наркотики
1021713	0	01.01.18	187801031000001	1978	31.12.2017 г. в ЛУВД с заявлением обратился Ше...	Кража
1021714	0	01.01.18	187812031000001	1978	В ЛОВД обратился [REDACTED] ...	Кража

Рис.1. Структура данных по преступлениям

Поле С1 - везде 0, неиспользуемые данные

Поле С2 - дата регистрации правонарушения

Поле С3- не известный показатель, неиспользуемые данные

Поле С4 не известный показатель, неиспользуемые данные

Поле С5- текстово сообщение, тоесть тело фабулы

Поле С6 - идентификационный номер региона правонарушения

Поле С7 – классифицированное правонарушение

Поля C1, C2, C4, C3, C6 не содержат ценной информации способствующей классификации правонарушений. C5 - это текстовое поле, которое необходимо преобразовать в цифровой формат для создания математических моделей машинного обучения. Поле C7 уже классифицировано сотрудниками правоохранительных органов.

В результате мы имеем 7 классов имущественных преступлений, которые указаны на рисунке 2:

- Вымогательство;
- Грабеж;
- Кража;
- Мошенничество;
- Наркотики;
- Разбой;
- Хулиганство.

```
#categories = np.unique(np.concatenate(yyy))# если выборка большая заменить на ->
categories = ['Вымогательство', 'Грабеж', 'Кража', 'Мошенничество', 'Наркотики', 'Разбой', 'Хулиганство']
```

Рис.2. Классы правонарушений.

Далее необходимо предварительно обработать текст и перевести их в цифровой тип данных. Данный процес называется векторизацией. Существует ряд методов, таких как tf-idf, word2vec, GloVe. Рассмотрим tf-idf и GloVe.

При поиске информации, tf-idf или TFIDF, короткий для частоты частотно-обратных документов, является числовой статистикой, которая предназначена для отражения того, насколько важно слово для документа в коллекции или корпусе. Оно часто используется в качестве весового коэффициента в поисках поиска информации, интеллектуального анализа текста и моделирования пользователей. Значение tf-idf увеличивается пропорционально количеству раз, когда слово появляется в документе и компенсируется частотой слова в корпусе, что помогает приспособиться к тому, что некоторые слова чаще появляются в целом. Tf-idf - одна из самых популярных схем взвешивания слов сегодня; 83% текстовых рекомендаций в цифровых библиотеках используют tf-idf [4].

GloVe, придуманная из Global Vectors, является моделью для распределенного представления слов. Модель представляет собой неконтролируемый алгоритм обучения для получения векторных представлений для слов. Обучение проводится по совокупной статистике совпадения словесного словаря из корпуса, а полученные представления демонстрируют интересные линейные подструктуры векторного пространства слов. Он разработан как проект с открытым исходным кодом в Стэнфорде [5].

Результаты

После векторизации текстовых сообщений, произведена ротация текстовых сообщений в случайном порядке. Перед началом обучения моделей, данные были разделены на обучающую выборку и тестовую, что соответствует в процентном соотношении 80% на 20%.

В результате, применяя несколько типов векторизации и две модели классификации, достигается следующая точность, как показано в таблице 1.

Таблица 1

Сравнение моделей

	tf-idf	GloVe
SVM	83%	85,5%
Логистическая регрессия	70%	73%

Вышеприведенные результаты были получены с использованием моделей с атрибутами по умолчанию, показанных на рисунках 3 и 4.

```
>>> from sklearn.pipeline import Pipeline
>>> text_clf = Pipeline([('vect', CountVectorizer()),
...                       ('tfidf', TfidfTransformer()),
...                       ('clf', MultinomialNB()),
... ])
```

Рис. 3. Параметры по умолчанию. Модель логистическая регрессия

```
>>> from sklearn.linear_model import SGDClassifier
>>> text_clf = Pipeline([('vect', CountVectorizer()),
...                       ('tfidf', TfidfTransformer()),
...                       ('clf', SGDClassifier(loss='hinge', penalty='l2',
...                                             alpha=1e-3, random_state=42,
...                                             max_iter=5, tol=None)),
... ])
```

Рис. 4 . Параметры по умолчанию. Модель SVM

После тестирования путем выбора и изменения атрибутов, как показано на рисунках 5 и 6, в таблице 2 показаны следующие результаты.

Таблица 2. Сравнение моделей после изменения атрибутов

	tf-idf	GloVe
SVM	95.5%	95.5%
Логистическая регрессия	94%	94.3%

```
from sklearn.pipeline import Pipeline
text_clf = Pipeline([('vect', CountVectorizer()),
                     ('tfidf', TfidfTransformer()),
                     ('clf', MultinomialNB(alpha=0.09, )),])
```

Рис. 5. Модель логистической регрессии после изменения атрибутов

```
text_clf = Pipeline([('vect', CountVectorizer()),
                     ('tfidf', TfidfTransformer()),
                     ('clf', SGDClassifier
                      (loss='hinge', penalty='l2', alpha=0.00006,
                       random_state=1, max_iter=5, tol=None)),])
```

Рис.6. Модель SVM после изменения атрибутов

На основании результата, представленного в таблице 2, можно сделать вывод, что тип векторизации не влияет на точность.

Заключение

В статье рассмотрено два алгоритма, которые классифицируют преступления на классы. Алгоритм Support vector machines (SVM) и Логистическая регрессия приведены в сравнительном анализе по критерию «точность». Рассмотрены методы tf-idf и GloVe для преобразования текстовых сообщений в машинный вид. Взята выборка данных, по преступлениям имеющая свыше одного миллиона фабул для обучения моделей с учителем. Описаны действия препроцессинга данных для улучшения итогового и объективно верного результата. Произведены такие действия как нормализация, случайная ротация и разделение данных на тестовую и обучающую выборку. Путем сравнительного анализа модель SVM была определена как модель, классифицирующая преступления с наибольшей точностью 95,5%. Также в ходе работы было доказано, что метод векторизации текста не влияет на конечную точность.

Список использованной литературы:

- 1 Госпрограмма «Цифровой Казахстан», утверждена Постановлением Правительства РК №827 от 12.12.2017 г.
- 2 Victoria Bobicev., Marina Sokolova. An Effective and Robust Method for Short Text Classification. // Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. – 2008. – P. 1444-1445.
- 3 Lucila Ohno-Machadob., Stephan Dreiseitla. Logistic regression and artificial neural network classification models: a methodology review. // Journal of Biomedical Informatics. – Volume 35. – Issues 5-6, October 2002. – P. 352-359.

4 Rajaraman, A., Ullman, J.D. Data Mining. // Mining of Massive Datasets (PDF). – 2011. – P. 1-17. doi: 10.1017 / CBO9781139058452. 002. ISBN 978-1-139-05845-2.

5 Jeffrey Pennington, Richard Socher, Christopher D. Manning “GloVe: Global Vectors for Word Representation” Computer Science Department. – Stanford University: Stanford. – CA 94305 – 2014. – 1543 p.