DOI 10.37882/2223-2966.2023.6-2.30

РАЗРАБОТКА ПРОГРАММНОГО СРЕДСТВА ДЛЯ КЛАССИФИКАЦИИ ПИСЕМ ЭЛЕКТРОННОЙ ПОЧТЫ НА ОСНОВЕ МЕТОДА ОПОРНЫХ ВЕКТОРОВ

DEVELOPMENT OF A SOFTWARE TOOL FOR CLASSIFYING E-MAIL MESSAGES BASED ON THE SUPPORT VECTOR METHOD

P. Savin
A. Zhuravlev
O. Lukyanchikov
A. Rusakov
V. Filatov

Summary: The article describes the principle of creating an application in the Python programming language using the support vector machine, which performs the classification of letters. On the prepared data set, this method, as well as in other scientific papers, showed high accuracy. A large number of emails accumulate in the e-mail of corporate employees during work correspondence, and every day employees spend a lot of time filtering and processing emails in their mailboxes, which affects their productivity. The developed software tool allows you to organize emails in the best possible way according to the specified user criteria.

Keywords: text mining, machine learning, support vector machine, e-mail.

Савин Павел Олегович
РТУ МИРЭА
zet20600@yandex.ru
Журавлев Артем Дмитриевич
Российская академия народного хозяйства и государственной службы
artem11102000@gmail.com
Лукьянчиков Олег Игоревич
к.т.н. доцент, Руководитель группы разработки CommuniGate Systems, РТУ МИРЭА
lukoilo@communigate.ru
Русаков Алексей Михайлович
старший преподаватель, РТУ МИРЭА
rusal@bk.ru
Филатов Вячеслав Валерьевич
Доцент, РТУ МИРЭА
filv@mail.ru

Аннотация: В статье описан принцип создания приложения на языке программирования Python, использующее метод опорных векторов, который выполняет классификацию писем. На подготовленном наборе данных этот метод также, как и в других научных работах показал высокую точность. В электронной почте сотрудников корпораций в ходе рабочих переписок накапливается большое количество писем, и ежедневно сотрудники тратят много времени на фильтрацию и обработку электронных писем в своих ящиках, что отражается на их производительности. Разработанное программное средство позволяет выполнить упорядочивание писем наиболее лучшим образом по заданным пользовательским критериям.

Ключевые слова: интеллектуальный анализ текстов, машинное обучение, метод опорных векторов, электронная почта.

Одним из первых способов коммуникаций в интернете является электронная почта, которая до сих пор повседневно используется для личных и деловых переписок. На нее приходят различные уведомления из Web-сервисов, чеки, результаты медицинских исследований и прочая важная информация. Среди всех этих писем зачастую оказывается нежелательная рассылка или же спам. Спам — массовая рассылка корреспонденции рекламного характера лицам, не выражавшим желания её получить [1]. В почте сотрудников корпораций в ходе рабочих переписок накапливается большое количество писем, и ежедневно сотрудники тратят много времени на фильтрацию и обработку электронных писем в своих ящиках, что отражается на их производительности. Поэтому весь этот поток входящих электронных писем необходимо как-то автоматически обрабатывать. Одной из практик является классификация электронных писем, перемещая их в разные папки или присваивая им тэги. В ряде работ [1-6] специалисты в области информационных технологий также занимаются классификацией электронных писем, что показывает актуальность данной задачи. Однако разные методы, используемые при классификации писем, дают разную точность результатов поэтому исследования данной проблемы всё ещё остается актуальным.

Классификация электронных писем сводится к классификации текста, содержащегося в нем. Данная задача весьма затруднительна, связано это с разнообразием признаков и большим объемом обрабатываемой текстовой информации.

Описание предметной области

Работа электронной почты осуществляется по стандартизированным протоколам. В процессе приема и отправки электронных писем задействованы следующие протоколы: SMTP, POP3, IMAP. В данных протоколах

осуществляется обмен сущностью электронное письмо формата eml, которое соответствует стандарту RFC 2822. Поэтому для классификации писем по содержимому необходимо сначала выполнить парсинг eml формата, и затем провести лексический и/или семантический анализ текста

В формате данных eml содержатся данные отправителя и получателя, текст самого письма и его тема, а также различные вложения. Задача парсера в данном случаем является получение текстового содержимого письма. Также необходимо очистить данный текст, от данных препятствующих корректной работе алгоритма машинного обучения. В ходе данного процесса из текста должны удаляться ссылки, так как они могут снижать точность алгоритма. После чего полученный текст сохраняется в нужной нам структуре данных.

Дальнейший лексический анализ текста сводится к преобразованию входной последовательности символов в последовательность слов, или лексем. С точки зрения лексического анализа входные символы можно разделить на две группы: символы разделители и символы, не относящиеся к разделителям, именно из таких подряд идущих символов собираются лексемы. Необходимым критерием для лексического анализа является выбор естественного язык, поскольку он влияет на набор разделители и лексем для анализа входных символов.

Семантический анализ — процесс формирования семантических отношений для автоматического «понимания» текста. Семантический анализ является трудоемкой математической задачей. Использование семантического анализ текста в задача машинного обучений зависит от задач, поставленных перед алгоритмом. Для решения задачи классификации электронных писем данный анализ считаю необязательным.

Обзор методов машинного обучения применительно для классификации писем электронной почты

Машинное обучение — особый раздел искусственного интеллекта, изучающего методы построения моделей, способных обучаться, и алгоритмов для их построения и обучения [7].

Машинное обучение также можно определить, как процесс решения практической задачи путем:

1) формирования набора данных
2) алгоритмического построения статистической модели на его основе.

Предполагается, что эта статистическая модель будет каким-то образом использоваться для решения практической задачи [8].

Под обучаемость в данном определении понимается процесс накопления опыта по ряду задач, в ходе которого происходит улучшение результата, получаемого при решении данных задач.

В настоящее время принято выделять 3 категории машинного обучения:

— обучение с учителем,
— обучение без учителя
— обучение с подкреплением

Процесс обучения с учителем заключается в выявлении взаимосвязей между входными данными и данными, ожидаемыми на выходе. Обучение происходит на заранее подготовленной выборке, которая содержит входные данные и необходимый результат на выходе. Точность данного алгоритма зависит от количества записей в обучающей выборке.

Во время обучения без учителя, алгоритм не располагает верными выходными значениями. Основная задача заключается в выявлении зависимостей между входными данными, упорядочивание их и создание паттернов. В какой-то степени происходит группировка (систематизация) записей по определенным признакам.

Обучение с подкреплением основано на взаимодействии алгоритма с некой средой. Алгоритм совершает определенные действия с данной средой, а в ответ получает от нее положительны или отрицательный отзыв. Другими словами, обучение с подкреплением можно описать как метод проб и ошибок. Задача алгоритма в данном случае сводится к избеганию отрицательных отзывов.

В нашем случае, необходимо решить задачу классификации электронных писем. Стоит отметить, что для данного типа задач используется алгоритм, основанный на обучении с учителем, это следует из наличия возможности произвести маркировку входных данных. Маркировка данных — процесс, при котором входным данным назначается соответствующие выходное значение.

Существует ряд алгоритмов позволяющих производить классификацию, все они являются подклассом алгоритма, основанного на обучении с учителем:

— деревья решений;
— машины опорных векторов;
— байесовский классификатор;
— линейный дискриминантный анализ;
— метод k-ближайших соседей;

На выбор алгоритма машинного обучения влияет ряд критериев. К ним относятся точность вычислений и объемы обрабатываемой информации. Для рассмотрения сравнительных характеристик методов машинного

обучения можно обратиться к работе Бурлаевой Е.И., Зори С.А. [9]. В данной работе был произведен сравнительный анализ методов машинного обучения для решения задач анализа текстовых документов.

Таблица 1.

Результаты научной работы Бурлаевой Е.И., Зори С.А.

Название композиции	Точность, %
T – I + SVM	88
T – I + LSA	85
T – I + D	80
T – I + SVM + D	93
T – I + LSA + D	91

В этой таблице:

tf-idf — статистическая мера, вес некоторого слова, который пропорционален количеству употребления этого слова в документе и обратно пропорционален частоте употребления слова в других документах коллекции;

SVM — метод опорных векторов;

LSA — латентно-семантический анализ;

D — дерево принятия решений;

Исходя из этого выбран метод **машины опорных векторов**. Основной идеей этого алгоритма является построение гиперплоскости, разделяющей объекты выборки оптимальным способом так, чтобы расстояние между ними было как можно больше, что ведет к меньшей средней ошибке классификатора:

$$wx^T + b = b + \sum_{i=1}^{m} a_i x^T x^{(i)} = b + \sum_{i=1}^{m} a_i k\left(x, x^{(i)}\right),$$

$$F(x) = sign\left(w^T x + b\right),$$

где $w = (w_1, w_2, \ldots, w_n)$ — весовые коэффициенты;

$x^{(i)}$— обучающий пример;

a — вектор коэффициентов;

b — базис;

$k\left(x, x^{(i)}\right) = \phi(x) \times \phi\left(x^{(i)}\right)$— скалярное произведение функций признаков $\phi(x)$, называемое ядром.

Разработанные программные средства

Для классификации писем реализован аналитический модуль на языке Python [10]. Данный программный продукт состоит из трех файлов:

Скрипт Main.py включает в себя основной алгоритм работы и графический интерфейс.

Скрипт parsefile.py — реализованный парсер файлов с форматом eml.

Скрипт my_parser.py — парсер, который осуществляет загрузку писем из каталога, адрес которого служит входным параметром.

Непосредственно для выполнения классификации необходим файл с подготовленным набором данных (data.csv). Данный набор данных может быть получен, если послать на вход разработанного приложения путь до папки с классифицированными электронными письмами eml формата (например, в общей папке находятся папки: A, B, C внутри которых письма в формате eml заданной тематики). Алгоритм 1 считывает данные и создают структуру класс-значение, которые преобразуются в файл формата csv и он сохраняется во внутренней папке программного обеспечения.

Следующим шагом начинается процесс обучения алгоритма на данных их csv файла. После процесса об-

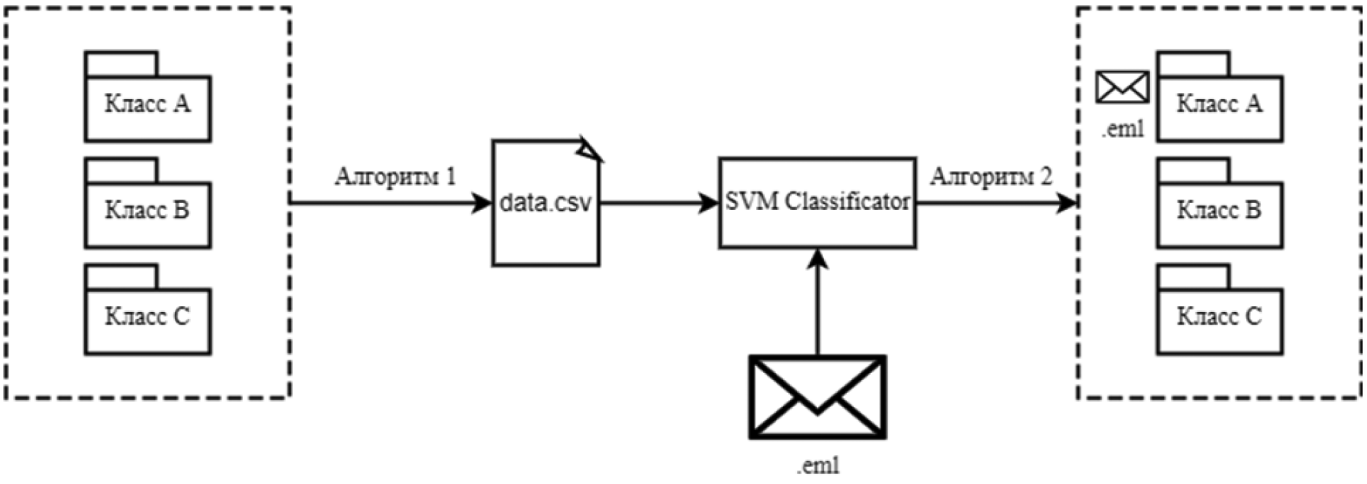

Рис. 1. Общая схема разработанного программного продукта

учения у пользователя появляется возможность выбора электронного письма и алгоритм 2 спрогнозирует категорию выбранного электронного письма.

Анализ результатов работы классификатора писем

В качестве источника электронных писем был выбран личный почтовый ящик. Письма были разбиты на 4 различные смысловые группы со следующими названиями:

— Научные работы
— Скидки
— Вебинары
— Чеки

Полученная данные были разбиты на две выборки: обучающая и тестовая. После процесса обучение, было произведено тестирование со следующими результатами. Не смотря на различие представленных категорий, стоит заметить, что иногда в прогнозировании случались ошибки. Наибольший процент ошибок наблюдался между категорией «Вебинары» и «Научные работы», это можно объяснить смежными часто употребляемыми словами, которые влияли на точность прогнозирования. Но тем не менее итоговая точность данного программного средства на представленных данных составила от 89 до 98 процентов, результат зависит от наполнения тестовой и обучающей выборки.

Заключение

В данной статье описано приложение на языке программирования Python, использующее метод опорных векторов машинного обучения, который выполняет классификацию писем. На подготовленном наборе данных этот метод также, как и в других научных работах показал высокую точность, однако алгоритм не всегда гарантирует верный результат. Поэтому данный метод можно использовать в качестве помощника для пользователя при получении новых писем. То есть при получении письма выводить оповещение с предложением переместить автоматически письмо в другую папку, а пользователь принимает окончательное решение выполнять это действие или нет.

ЛИТЕРАТУРА

1. Youn S., McLeod D. A comparative study for email classification //Advances and innovations in systems, computing sciences and software engineering. — Springer, Dordrecht, 2007. — С. 387–391.
2. Катасёв А.С., Катасёва Д.В., Кирпичников А.П. Нейросетевая технология классификации электронных почтовых сообщений // Вестник Казанского технологического университета. — 2015. — Т. 18. — №. 5.
3. Wang X.L. et al. Learning to classify email: a survey // 2005 International conference on machine learning and cybernetics. — IEEE, 2005. — Т. 9. — С. 5716–5719.
4. Alsmadi I., Alhami I. Clustering and classification of email contents //Journal of King Saud University-Computer and Information Sciences. — 2015. — Т. 27. — № 1. — С. 46–57.
5. Alghoul A. et al. Email classification using artificial neural network. — 2018.
6. Klimt B., Yang Y. The enron corpus: A new dataset for email classification research // European Conference on Machine Learning. — Springer, Berlin, Heidelberg, 2004. — С. 217–226.
7. Бирозин С., Ракян С. Internet у вас дома 3-е изд. // СПб.: BHV— СанктПетербург. — 2010.
8. Бурков А. Машинное обучение без лишних слов // Питер. —2020. —192.
9. Бурлаева Е.И., Зори С.А. Сравнение некоторых методов машинного обучения для анализа текстовых документов // Проблемы искусственного интеллекта. — 2019. — №. 1 (12). — С. 42–51.
10. Исходный код проекта «Разработка программного средства для классификации писем электронной почты на основе метода опорных векторов» [Электронный ресурс] — URL — https://github.com/Artem11102000/VKR (дата обращения: 22.05.2023)