УДК 004.912

МОДЕЛЬ СИСТЕМЫ ИДЕНТИФИКАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ В УНИВЕРСАЛЬНОЙ ИНФОРМАЦИОННОЙ СРЕДЕ

М. А. Севостьянюк, Ю. А. Скудняков
Белорусский государственный университет информатики и радиоэлектроники, Минск

Предложена модель системы идентификации текстовой информации в универсальной информационной среде с использованием байесовского классификатора с распределением Бернулли. Классификатор позволяет автоматически обрабатывать и идентифицировать текстовые данные с высокой скоростью и точностью, а также полностью заменить человеческий ресурс при распознавании различных документов. Приведены результаты тестирования разработанного общего алгоритма работы идентификационной системы, сформулированы подходы по развитию модели данной системы с точки зрения расширения ее функциональных возможностей и повышения качества идентификации.

В настоящее время поиск информации без использования каких-либо специальных ресурсов и устройств является весьма трудозатратным. Для корректной идентификации данных необходимо их классифицировать или провести кластерный анализ – преобразование множества объектов на группы. В каждой группе будут похожие объекты, а объекты различных групп должны быть как можно более отличающимися.

Задачами классификации для идентификации данных являются:

– получение представления о структуре данных, что позволяет упростить их дальнейшую обработку и применять конкретный алгоритм для каждой группы;

– обнаружение новизны, т. е. можно найти объект, который невозможно отнести ни к одной из групп.

В докладе рассмотрены возможности существующих нейронных сетей [1, 2] и на основе результатов их анализа предложена модель нечеткой экспертной системы, использование которой позволяет автоматизировать процесс диагностики выработки индивидуальных рекомендаций для лечения пациента с применением известных алгоритмов.

Крупные компании, такие как Apple, Google, Microsoft, Cisco, Яндекс, используют классификацию для идентификации данных.

Для реализации предложенной модели системы идентификации текстовой информации разработан алгоритм, который используется в компании, занимающейся медицинскими услугами и оборудованием. Он позволяет обрабатывать большой объем данных из личных карт пациентов автоматически, выделять и собирать только самую необходимую информацию.

Классификация текстов различных документов является задачей компьютерной лингвистики и заключается в отнесении какой-либо части документа к одной из нескольких категорий на основании содержания документа в целом.

В задаче идентификации каких-либо данных из карты пациента можно выделить этапы обработки текста, выделения признаков из текста, выбора и обучения классификатора и его практическое применение.

На этапе обработки текста все слова, которые имеют одинаковую смысловую нагрузку, приводятся к одному виду, удаляются лишние слова и символы, лишь мешающие классификации документа.

При выделении признаков производятся формирование пространства и перевод в него всего текста карты. Затем используется математический аппарат.

В докладе рассмотрен байесовский классификатор, для обучения которого был проанализирован набор данных из медицинских карт пациентов и личных карт персонала.

Байесовский классификатор является вероятностным классификатором, который основан на теореме Байеса. Он называется «наивным», потому как считает, что наличие одного из признаков никак не зависит от наличия другого [3].

Предложенная модель идентификации данных базируется на приведенном ниже математическом аппарате, программная реализация которой позволяет получить достаточно хорошие результаты при практическом ее использовании.

Теорема Байеса позволяет рассчитать апостериорную вероятность:

$$P(c_i|x) = P(x|c_i) \cdot P(c_i) / P(x),$$

где $P(c_i|x)$ – вероятность того, что вектор $x \in X$ относится к классу $c \in C$; $P(x|c_i)$ – вероятность того, что x примет значение при условии выбора класса c_i; $P(c_i)$ – априорная вероятность класса c_i; $P(x)$ – априорная вероятность вектора x.

Для общего представления класса вектора x можно найти такой класс, который максимизирует $P(c_i|x)$:

$$y_i = \arg\max_{c_i \in C} P(c_i|x) = \arg\max_{c_i \in C} P(x|c_i) \cdot P(c_i) / P(x),$$

$$P(x|c_i) = P(x_1|c_i) \cdot P(x_2|c_i, x_1) \cdot \ldots \cdot P(x_n|c_i, x_1, x_2, \ldots, x_{n-1}).$$

Вероятность $P(x|c_i)$ непросто посчитать. В большинстве случаев остается вопрос, как именно появление одного признака зависит от появления другого:

$$P(x_n|c_i, x_1, x_2, \ldots, x_{n-1}).$$

Алгоритм становится «наивным» в момент, когда можно воспользоваться предположением о независимости признаков:

$$P(x_2|c_i, x_1) = P(x_2|c_i).$$

Решаемую задачу можно привести к виду

$$y_i = \arg\max_{c_i \in C} P(c_i) \cdot P(x_1|c_i) \cdot P(x_2|c_i) \cdot \ldots \cdot P(x_n|c_i).$$

Если известны $P(c_i)$ и условные вероятности $P(x_i|c_i)$, можно найти вероятность принадлежности определенной части документа или медицинской карты x_i к каждому классу c_i. В дальнейшем класс, на котором достигается максимальная вероятность, и будет считаться результатом работы алгоритма. Применим этот алгоритм к решению основной задачи. Существует всего два класса документов: $C = \{c_i,\ i = 1, 2\}$. Множество документов можно обозначить $X \subset R^n$, где R^n– множество всех действительных чисел.

Если обучающая выборка сформирована достаточно точно и описывает реальное распределение частей документов по классам, то $P(c_i)$ можно вычислить по формуле

$$P(c_i) = D(c_i),\ i = 1, 2,$$

где $D(c_i)$ – количество документов класса c_i во всем множестве документов D.

Принадлежность документа классу c_i можно вычислить путем использования распределения Бернулли:

$$P(x_i|c_i) = \prod_{i=1}^{n} P(w_i|c_i)^{x_i} \cdot (1 - P(w_i|c_i))^{1-x_i}.$$

Распределение используется в случае дискретных признаков, которые могут принимать значения 0 или 1. При этом следует учитывать, что все документы переведены в *n*-мерное пространство.

Обучение классификатора можно осуществить, используя значение

$$P(w_i|c_i) = \sum_{x \in X_1} x_i \,/\, \sum_{i=1}^{n} \sum_{x \in X_1} x_i \,,$$

где $X_1 \subset X$ – множество тех документов, которые принадлежат классу c_i; w_i– слово, которое соответствует x_i.

Оценка результатов использования предложенной модели осуществлялась путем ее автоматизированного тестирования на основе алгоритма, реализованного на языке Python. Ниже приведены результаты его работы с использованием разных вариантов распределения и размеров обучающей выборки.

Количество документов для классификации – от одного до 20 тыс. Данная оценка показывает, насколько точно происходит классификация, т. е. сколько из предсказанных правильных текстов или документов реально оказались медицинскими картами. Чем ближе это значение к единице, тем больше истинно положительных (true positive, *TP*) результатов в сравнении с ложноположительными (false positive, *FP*).

Расчет точности *A* выполняется по формуле

$$A = t_p / (t_p + f_p),$$

где t_p– количество истинно положительных результатов при классификации; f_p – количество ложноположительных результатов при классификации.

Значение *TP* показывает, сколько текстов или документов из общего числа реальных медицинских карт были отнесены к истинным. Чем ближе значение TP к единице, тем больше истинно положительных результатов в сравнении с ложноотрицательными (false negative, *FN*):

$$TP = t_p / (t_p + f_n),$$

где f_n – количество ложноотрицательных результатов при классификации.

Значение *FP* показывает, сколько текстов или документов из общего числа реальных были отнесены к истинным:

$$FP = f_p / (f_p + t_n),$$

где t_n – количество истинно отрицательных (true negative, *TN*) результатов при классификации.

F-мера, оценивающая параметры A и TP, описывается формулой

$$F = \frac{1}{\alpha \cdot \frac{1}{A} + (1-\alpha) \cdot \frac{1}{TP}}, \quad \alpha \in (0,1).$$

Регулируя коэффициент α, можно увеличить или уменьшить вес параметров. Например, если $\alpha > 0{,}5$, то *F*-мера будет сбалансированной.

Конкретные результаты классификации приведены в таблице.

Результаты классификации оценки принадлежности текста документа к нужному классу

Количество документов	*A*	*TP*	*FP*	*F*-мера
$\|X_1\| = 1000$	0,725 66	0,8445	0,1653	0,781 15
$\|X_2\| = 10\ 000$	0,818 13	0,8373	0,0711	0,823 10
$\|X_3\| = 20\ 000$	0,803 25	0,8285	0,0761	0,815 70

При обработке большого количества данных в документах наиболее важно и приоритетно решать задачи скорости обработки информации иидентификации правильных документов.

Тестирование предложенной модели показало, что использование байесовского классификатора дает хорошие результаты, которые полностью подошли под решаемые задачи.

Список литературы

1. Хайкин, С. Нейронные сети: полный курс Neural Networks: A ComprehensiveFoundation / С. Хайкин.– М. : Вильямс, 2006. – 2-е изд.– 1104 с.

2. Николенко, С. И. Глубокое обучение / С. И. Николенко, А. А. Кадурин, Е. О. Архангельская. – СПб. : Питер, 2018. – 480 с.

3. Воронцов, К. В. Лекции по статистическим алгоритмам классификации / К. В. Воронцов [Электронный ресурс]. – 2005. – Режим доступа: http: // www.machine-learning.ru/wiki/images/e/ed/Voron-ML-Bayes.pdf. – Дата доступа: 20.08.2020.