КОМПЬЮТЕРНОЕ ОБЕСПЕЧЕНИЕ И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

DOI: 10.24143/2072-9502-2021-1-16-27
УДК 004.89

КЛАССИФИКАЦИЯ КОРОТКИХ ТЕХНИЧЕСКИХ ТЕКСТОВ С ПРИМЕНЕНИЕМ СИСТЕМЫ НЕЧЕТКОГО ВЫВОДА СУГЕНО

А. В. Боровский[1], Е. Е. Раковская[1], А. Л. Бисикало[2]

[1]Байкальский государственный университет,
Иркутск, Российская Федерация

[2]Иркутский государственный университет,
Иркутск, Российская Федерация

Важным этапом работы при проектировании технических систем специального назначения является подбор оборудования с учетом эксплуатационных характеристик. Необходимость категоризации технических коротких текстов, которые представляют собой краткие описания оборудования, аннотации, фрагменты баз данных, обусловлена тем, что информация об оборудовании, содержащаяся в тематических реферативных сборниках, технической и проектной документации, контекстной рекламе, зачастую не структурирована, имеется в разрозненных источниках. Дополнительной проблемой является наличие большого количества опечаток, некорректных словоупотреблений и обозначений в текстах. Приведены результаты классификации технических коротких текстов о назначении приборов с применением теории нечетких множеств и нечеткой логики. Большое внимание уделяется характеристике объектов исследования и учету их особенностей – наличию большого количества технических терминов, аббревиатур, специальных символов. Описана методика проведения классификации, обоснована целесообразность применения системы нечеткого вывода Сугено, связанная с «нечеткостью» естественно-языковой информации, простотой математических расчетов в ходе эксперимента. Модель Сугено сочетает в себе описание объектов исследования в виде лингвистических правил и функциональных зависимостей. Такой подход значительно облегчает интерпретацию результатов классификации.

Ключевые слова: технические короткие тексты, нечеткие множества, системы нечеткого вывода Сугено, классификация.

Для цитирования: *Боровский А. В., Раковская Е. Е., Бисикало А. Л.* Классификация коротких технических текстов с применением системы нечеткого вывода Сугено // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2021. № 1. С. 16–27. DOI: 10.24143/2072-9502-2021-1-16-27.

Введение

Переход к цифровой экономике в России, объявленный Президентом, способствует развитию компьютеризации во всех сферах жизни общества и государства [1]. Повсеместно внедряются и совершенствуются автоматизированные системы управления технологическими процессами и производством, диспетчерские системы управления, информационные системы различного назначения, системы компьютерного проектирования. Особо следует отметить развитие сети Интернет, переход на новые электронные способы коммуникации («Робот Анна» – пилотный проект роботизации контактного центра для клиентов Сбербанка), появление новых видов массовой коммерческой деятельности, связанной с развитием IT-сервисов (интернет-магазины, поиск и бронирование гостиниц онлайн, продажа электронных билетов, «Яндекс-такси» и т. д.).

На подходе появление роботов-шоферов, секретарей, гидов и др. В связи с этим резко возрастает интерес разработчиков к вопросам автоматической обработки текстовой информации. Лингвистический аспект важен для всех направлений индустрии обработки знаний: сбора, создания, хранения, систематизации, распространения, интерпретации информации.

В настоящее время многие исследования и разработки посвящены вопросам категоризации текстов. Это связано со значительной важностью и актуальностью прикладных задач, решаемых на основе классификации – работой с базами текстовой информации, фильтрацией спама, определением вредоносного контента, анализом настроений [2], обработкой звуковых сообщений [3]. Для классификации текстов успешно применяются разнообразные методы машинного обучения. Самые распространенные из них – метод Байеса (Naive Bayes, NB), метод К-ближайших соседей (K Neighbors, KNN), метод деревьев решений (Decision Trees, DT), метод опорных векторов (Support Vector Machine, SVM), методы на основе искусственных нейронных сетей [4]. Все эти методы работают с некоторой числовой моделью, которая переводит тексты в удобную для дальнейшей работы форму. Наиболее распространенные модели текста: «мешок слов» (Bag of Words, BW) – характеризуется представлением документа в виде вектора слов и частоты их появления в документе; модель Word2vec – представляет каждое слово в виде вектора, который содержит информацию о сопутствующих словах; модель, основанная на учете *n*-грамм, т. е. основной характеристикой текста принимается последовательность из соседних символов [5, 6].

Моделирование лингвистической информации для решения задач классификации и кластеризации коротких текстов

В классификаторах, основанных на традиционном машинном обучении, предполагается, что каждый класс текстов однозначно отделяется от других классов, т. к. классификация базируется на предположении, что разные классы являются взаимоисключающими и каждый объект не может принадлежать более чем к одному классу. Однако приведенные выше допущения не всегда выполняются при классификации реальных текстовых данных. Классы текстов могут иметь пересечения, что затрудняет проведение классификации с однозначным отнесением объекта к определенному классу [7].

Сложность систематизации естественно-языковой информации объясняется специфичностью объектов исследования. Естественный язык, т. е. язык, используемый как средство общения людей, характеризуется нечеткостью смысловых значений словесных выражений [8]. Человек мыслит нечеткими понятиями, метафорами, дает неточные оценки, делает заключения в неколичественных, нечетких терминах. Языку присущи противоположные тенденции: субъективность – объективность, устойчивость – изменчивость, отсутствие синонимии – бесконечность синонимических средств, дискретность – континуальность значения, информативность – избыточность, логичность – нелогичность и т. д.

К фундаментальным качествам естественного языка относятся следующие:

– динамичность языковой системы;

– образность словесных выражений (основанная, прежде всего, на метафоричности);

– бесконечные творческие возможности в определении новых понятий;

– семантическое многообразие словарного состава, позволяющее выражать любую информацию с помощью конечного множества терминов;

– гибкость в передаче информации;

– разнообразие функций (включающее коммуникативную, когнитивную, планирующую, управляющую, обучающую, эстетическую, метаязыковую и др. функции);

– специфическая системность, под которой имеется в виду не только разделение языка на уровни – фонетический, морфологический, лексический, – но и разделение его на разновидности языка – относительно независимые функционально-тематические подсистемы (литературный язык, территориальные диалекты, язык делового общения).

По числу своих внутренних возможностей естественный язык может быть с полным основанием оценен как сложнейший объект для моделирования.

Для множественной классификации коротких текстов в настоящее время широко применяются технологии глубинного обучения (Deep Learning), основанные на использовании искус-

ственных нейронных сетей различной архитектуры, что позволяет значительно улучшить результаты исследования при выполнении практических задач, например анализ настроений в Твиттере [9], тематическая классификация новостей [10].

В том случае, когда имеется недостаточное количество документов в обучающем наборе или тексты очень короткие, применение глубинного обучения может быть затруднено [11]. Для классификации коротких текстов при условии недостаточности данных используются подходы, учитывающие семантический контекст терминов документа. Расширение признакового пространства модели представления коротких текстов за счет включения знаний из онтологий, применение контекстно-зависимых алгоритмов являются более предпочтительными и дают улучшенные результаты [12].

Онтологический подход совместно с методологией обработки естественного языка реализован Э. Д. Павлыгиным с соавторами [13] для определения социального портрета пользователей социальных медиа посредством классификации коротких текстовых фрагментов (посты социальной сети, комментарии) и открытых данных со страницы пользователя. Классами являются категории интересов пользователя: политика, бизнес, спорт, IT-технологии, музыка, кино и пр.

Фрагмент онтологии предметной области для определения предпочтений пользователей имеет вид $G = \langle C, R_C, I, R_I \rangle$, где C – множество категорий интересов пользователя; R_C – множество отношений, определяющее иерархию категорий интересов; $I = \langle f_1, ..., f_l, ..., f_z \rangle$ – множество признаков категории интересов (указанное множество признаков состоит из лексем, включающих слова и словосочетания, характеризующие категорию); R_I – множество отношений, определяющее связи между категориями и признаками. Задачей классификации является нахождение наиболее вероятной категории из множества C для текстового фрагмента d_i [13].

Для решения задачи классификации твитов, коротких сообщений новостных лент Blaž Škrej с соавторами применили контекстно-зависимый семантический подход tax2vec [14], в котором признаковое пространство документов определялось на основе меры tf-idf, характеризующей важность слова в документе, и информации, доступной в таксономиях – иерархических структурах классификаций определенного набора объектов. Подход tax2vec значительно повышает эффективность классификаторов за счет обогащения векторного пространства документов семантическими признаками из таксономий.

Контекстно-зависимые алгоритмы классификации коротких текстов были адаптированы для задач классификации химических веществ и протеинов по их свойствам. Молекулярные данные были векторизованы с применением эмбеддинговых моделей Mol2vec и Protvec [15], где химические соединения (модель Mol2vec) и белковые последовательности (модель Protvec) абстрактно представлялись в виде «предложений», а функциональные группы и сочетания функциональных групп как «слова». На основе этих псевдотекстов проводилось машинное обучение с применением различных методов (случайных лесов (Random forest), метода опорных векторов (SVM), методов глубинного обучения) и выполнялись задачи классификации по характерным химическим и физическим свойствам веществ (растворимость, биологическая активность и т. д.).

Перспективным направлением множественной классификации коротких текстов является применение мультимодельного подхода [16, 17]. В статье [17] описано функционирование информационной системы автоматизированного анализа неструктурированных документов на примере рубрикации текстов обращений граждан в государственные и общественные органы. В зависимости от характеристик текстов обращений – размера документов, степени пересечения тезаурусов рубрик, количества исследуемых текстов – использовались вероятностные модели или интеллектуальные методы анализа данных. Выбор метода анализа осуществляется с учетом комбинации критериев, определяющих условия применимости конкретной модели. По мнению авторов [17], при наличии взаимосвязанных рубрик предпочтительно использование модели рубрикации на основе нечетких деревьев решений. Если рубрики (классы) не взаимосвязаны между собой, то целесообразно применять нейро-нечеткий классификатор и вероятностные методы.

Для систематизации текстовой информации, связанной с субъективным восприятием человека и имеющей нечеткую природу, закономерно применение методов на основе нечеткой логики и теории нечетких множеств [18]. Нечеткая логика рассматривается как расширение детерминированной логики, т. е. в нечеткой логике рассматриваются непрерывные значения ис-

тинности от 0 до 1, а не бинарные значения 0 или 1 [19]. В контексте теории нечетких множеств каждый элемент имеет некоторую степень принадлежности к множеству, т. е. частично принадлежит множеству. Степень принадлежности элементов определяется функцией принадлежности, характеризующей нечеткое множество.

Функционирование нечетких классификаторов, представляющих собой системы нечеткого вывода, характеризуется следующими стадиями: формирование базы правил, фаззификация (введение нечеткости) входных переменных, вычисление степени выполнения условий отдельных правил, определение степени истинности заключений отдельных правил, определение результирующей функции принадлежности выходного значения всех правил (агрегирование), дефаззификация выходных переменных [20].

База нечетких правил вида «Если – то» («If – then») определяет причинно-следственные отношения (связи) между входными и выходными величинами и разрабатывается с использованием экспертных знаний [21] или посредством статистического изучения реальных данных [22]. При составлении базы правил применяются приближенные рассуждения, основанные на тавтологии типа обобщенный Modus Ponens (Generalized Modus Ponens, GMP). Этот подход позволяет употреблять в условиях и заключениях правил нечеткие размытые формулировки вида «более чем», «примерно равно», «более-менее» [20].

На стадии фаззификации для четких значений входных переменных X_1 и X_2 вычисляются степени принадлежности нечетким множествам A_i и B_j. Для выполнения этой операции предварительно должны быть определены функции принадлежности $\mu_{A_i}(X_1)$ и $\mu_{B_j}(X_2)$ входных переменных.

Для корректной работы классификатора необходимо найти значение истинности условия для каждого правила. Чем выше степень выполнения условия, тем большее влияние правило оказывает на результат вывода [20]. В случае сложного условия, состоящего из простых подусловий со связкой «И» «Если$(x_1 = A_1)$ и $(x_2 = B_2)$», степень выполнения условия для числовых аргументов $x_1 = x_{1i}$ и $x_2 = x_{2i}$ определяется как степень принадлежности μ нечеткому отношению R:

$$\mu_R(x_{1i}, x_{2i}) = \mu_{A_1 \cap B_2}(x_{1i}, x_{2i}) = T\left(\mu_{A_1}(x_{1i}), \mu_{B_2}(x_{2i})\right),$$

где A_1, B_2 – нечеткие множества; T – оператор *t*-нормы, например MIN.

Степени выполнения условий отдельных правил используются в дальнейшем для определения степени активации заключений правил. Эта операция выполняется с использованием операторов нечеткой импликации.

Если для правила «Если$(x = C)$To$(y = D)$» нечеткую импликацию представить в виде отношения $R : C \to D$ и применить нечеткую импликацию Мамдани, то степень активации нечеткого правила будет равна $\mu_{C \to D}(x, y) = \text{MIN}(\mu_C(x), \mu_D(y))$, где C и D – нечеткие множества.

Для определения результирующей функции принадлежности вывода всех правил рассматриваются полученные на предыдущем этапе степени активации для каждого правила и производится их объединение.

Процесс дефаззификации выходных переменных проводится по одному из методов, в зависимости от требований эксперимента (метод центра тяжести, метод центра площади, метод левого модального значения, метод правого модального значения) [23].

В настоящее время существует необходимость категоризации технических коротких текстов, которые содержатся в тематических реферативных сборниках, технической и проектной документации, контекстной рекламе и представляют собой краткие описания оборудования, аннотации, фрагменты баз данных. При проектировании технических систем специального назначения важным этапом работы является подбор оборудования с учетом эксплуатационных характеристик. Информация об оборудовании часто не структурирована, имеется в разрозненных источниках. Проблемой поиска необходимой информации об оборудовании также является наличие большого количества опечаток, некорректных словоупотреблений и обозначений в текстах.

Целью настоящей статьи является проведение классификации технических коротких текстов о назначении приборов с применением систем нечеткого вывода Сугено.

Для достижения поставленной цели необходимо выполнить следующие задачи:

– теоретически обосновать возможность и целесообразность применения систем нечеткого вывода Сугено для классификации технических текстов;

– провести классификацию текстов;

– подобрать модель нечеткого вывода для классификации текстов о назначении приборов, методы фаззификации, агрегирования, импликации, дефаззификации значений входных (выходных) переменных;

– охарактеризовать значения выходной функции – номера классов текстов, равные 1, 2 или 3, в виде синглетонов (Singleton) – множеств с единственным элементом.

Эксперимент

Классификация коротких текстов с применением систем нечеткого вывода была проведена на примере текстов о назначении датчиков давления (табл. 1).

Таблица 1

Примеры текстов о назначении датчиков давления

№ п/п	Тексты
1	Датчики давления ИДД предназначены для измерения избыточного давления воды в магистралях систем теплоснабжения, горячего и холодного водоснабжения, при эксплуатации в составе автоматизированного индивидуального теплового пункта.
2	Датчики давления ТЖИУ406-М100-Вн предназначены для непрерывного измерения и преобразования значений измеряемого параметра, избыточного давления, абсолютного давления, разности давлений, избыточного давления-разрежения, разрежения нейтральных по отношению к нержавеющим сталям и сплавам титана, жидких, газообразных сред и пара в унифицированные выходные токовые сигналы и (или) цифровые сигналы в стандартах протоколов HART или MODBUS с интерфейсом RS-485.
3	Датчики давления ИВЭ-50-3 предназначены для измерений и преобразования значения измеряемой величины давления в унифицированный аналоговый электрический сигнал.
4	Датчики давления высокотемпературные ДДВС-РТМ предназначены для измерения избыточного давления в натриевых трубопроводах I и II контура, сосудах II контура и раздающем коллекторе ПГ РУ БН-800 и для формирования информационных сигналов, соответствующих измеряемому давлению. Датчики предназначены для использования в системах автоматического контроля параметров объектов.

Датчики применяются для оценки физических параметров исследуемых систем в стационарных или динамических условиях. Основными характеристиками датчиков давления являются измеряемый диапазон давления и ожидаемое измерение относительного или абсолютного давления.

При выполнении исследования учитывались особенности текстов – их размер, применение специфической терминологии, наличие особых символов, знаков, формул, аббревиатур [24]. На этапе предварительной обработки из текстов были удалены термины, не имеющие смысловой нагрузки – «стоп-слова», цифры, знаки препинания, короткие слова, многочисленные аббревиатуры. Тексты были преобразованы для удаления окончаний слов. Также учитывалась синонимия слов, т. е. близкие по значению речевые обороты заменялись предварительно заданными синонимическими (терминами). В качестве модели представления текстовых данных была использована модель «мешок слов», в которой каждый документ рассматривается как совокупность содержащихся в нем терминов. В этой модели не учитываются порядок расположения слов и семантические связи между ними.

В дальнейшем тексты были преобразованы в числовые данные. Для этого в каждом тексте подсчитывалось количество употреблений слов. На основе данных о частоте встречаемости слов была построена матрица «документ – термин» размерностью $m \times n$, строки которой соответствуют документам, а столбцы – терминам. Характеристиками терминов в документе являются частоты встречаемости терминов или связанные с ней величины: обратная частота слов в документе *tf – idf*, бинарные значения («1» – слово встречается в документе, «0» – не встречается) [24, 25]. Размерность матрицы была уменьшена при помощи сингулярного разложения (SVD) $C = UWV^T$, где C – матрица «документ – термин» размерностью $m \times n$; U – $m \times m$-матрица, столбцы которой являются собственными ортогональными векторами матрицы CC^T; V^T – $n \times n$-матрица, столбцы которой являются собственными ортогональными векторами матрицы C^TC; W – диагональная $m \times n$-матрица с диагональю из невозрастающих чисел σ_r, где r – ранг матрицы C; C^T – транспонированная матрица C [14, 15].

Для нахождения матрицы C_p, являющейся малоранговой аппроксимацией матрицы C, по матрице W строилась матрица W_p с заменой нулями $(r - p)$ наименьших значений. По полученным матрицам (U, W_p, V^T) вычислялась матрица $C_p = U\, W_p\, V^T$.

В результате проделанных операций размер исходной матрицы $C_{200\times78}$ был снижен до размерности 200 × 10. Для анализа были взяты первые два столбца матрицы C_p – параметры x_1 и x_2. После «оцифровки» текстов и снижения размерности матрицы «документ – термин» были получены данные (фрагмент) (табл. 2).

Таблица 2

Данные для анализа (входные параметры x_1 и x_2)

№ документа	x_1	x_2
1	0,045	0,026
2	0,045	0,038
3	0,025	0,036
4	0,039	0,031
5	0,039	0,031

Для классификации использовалась система, содержащая 2 входа (параметры x_1 и x_2) и 1 выход (обозначение номеров классов датчиков давления). Диапазоны изменения входных параметров: x_1 от 0 до 0,24, x_2 от –0,15 до 0,35, множество значений выходного параметра $y \in \{1, 2, 3\}$. Взаимосвязь параметров x_1 и x_2 с обозначением классов текстов о назначении датчиков давления представлена на рис. 1.

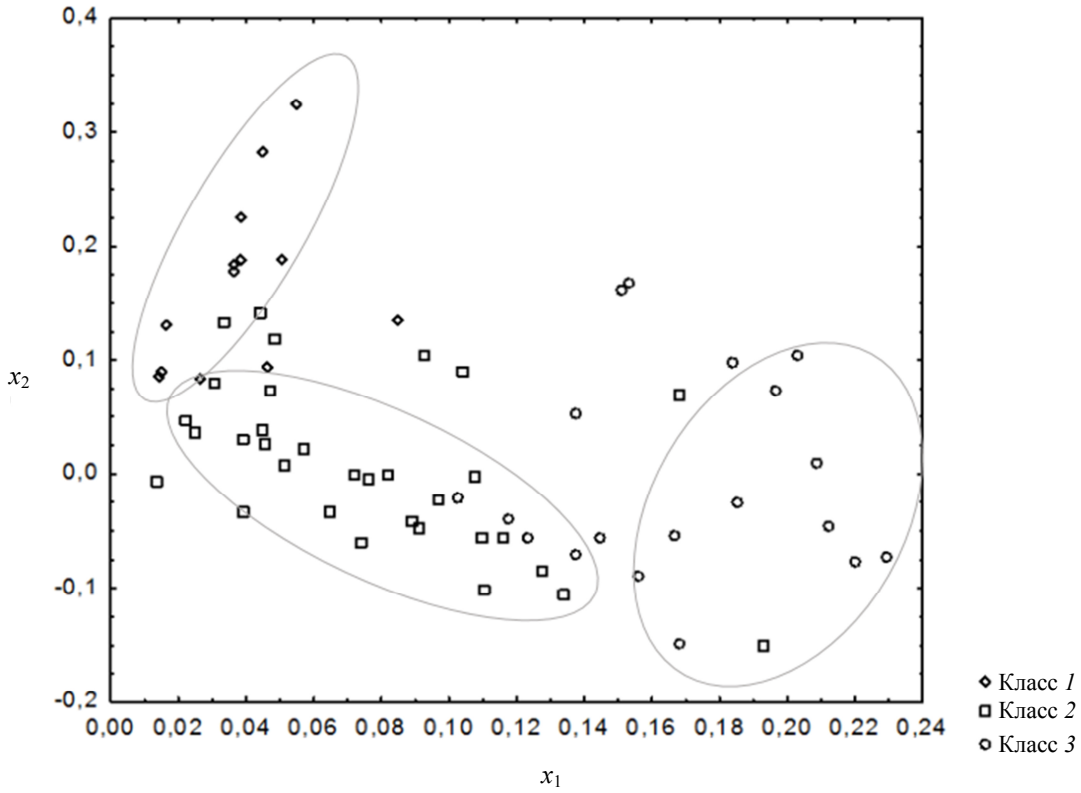

Рис. 1. Область значений входных параметров модели классификации

Классы текстов определялись по описанию условий применения датчиков давления и характеристике измеряемых величин: *1* – тексты о датчиках для определения быстроменяющегося давления (ромбы); *2* – тексты о датчиках для определения абсолютного и избыточного давления (квадраты); *3* – тексты об универсальных датчиках, пригодные для измерения давления разрежения, разности давлений (круги).

На рис. 1 наблюдается выраженная локализация точек в зависимости от классов отображаемых ими текстов. Это дает возможность применить для классификации текстов методы на основе нечеткой логики и теории нечетких множеств. Для оценки переменных x_1 и x_2 использовались лингвистические переменные «малый», «большой» с заданием функций принадлежности треугольного и трапециевидного типов.

Функции принадлежности нечетких множеств для входных параметров x_1 и x_2 показаны на рис. 2 и 3.

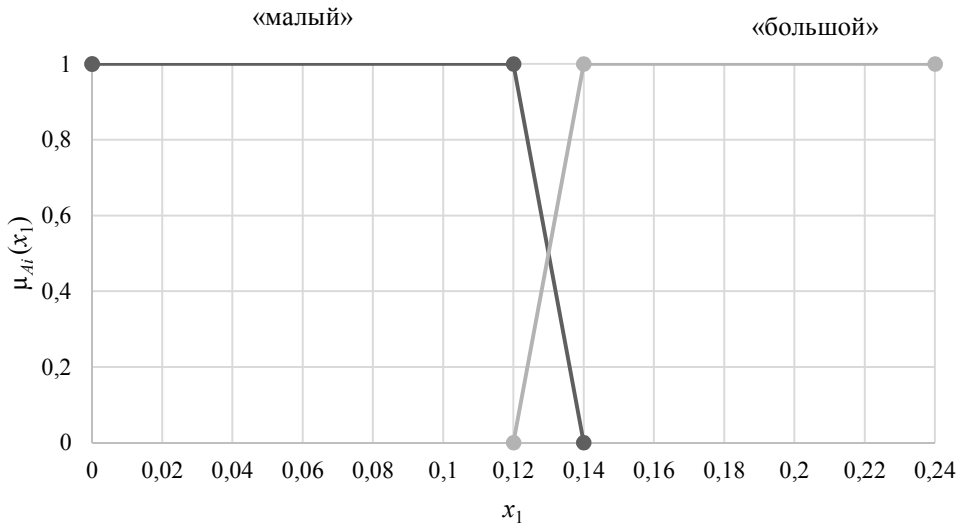

Рис. 2. Функции принадлежности нечетких множеств «малый» и «большой» входного параметра x_1

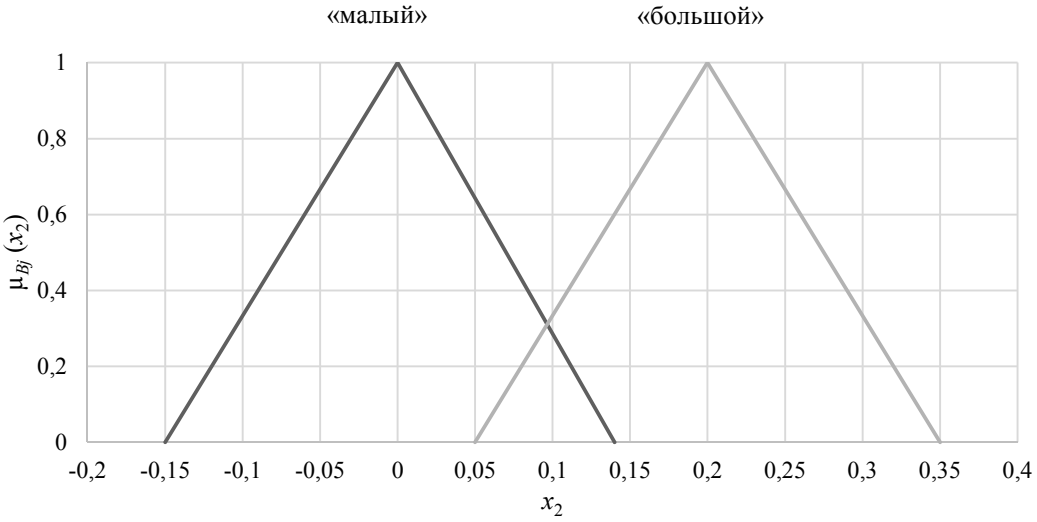

Рис. 3. Функции принадлежности нечетких множеств «малый» и «большой» входного параметра x_2

При установлении функций принадлежности проводился анализ зависимости $x_1 - x_2$ с разбиением пространства X – значений входных параметров модели, на участки (кластеры), соответствующие каждому классу. Максимальные значения функций принадлежности совпадают с центрами полученных кластеров. Минимальные значения функции принадлежности соответствуют областям на границе кластеров или переходным областям, где возможно отнесение текстов к одному из двух классов.

По условиям эксперимента поверхность отображения $X \to Y$, где X и Y – множества значений входных и выходных параметров, является ступенчатой. Каждая «ступень» соответствует определенному классу – значения 1, 2 или 3.

Для классификации текстов применялись системы на основе нечетких правил, в которых в качестве консеквентов использовались одноэлементные множества (синглетоны). Модель нечеткого вывода с синглетонами в заключениях нечетких правил может быть интерпретирована как модель нечеткого вывода Сугено, типичное правило которой имеет вид «Если $x_1 = A$ и $x_2 = B$, то $y = f(x_1, x_2)$», где y – четкая функция, x_1 и x_2 – входные параметры модели, A и B – нечеткие множества, которым соответствуют функции принадлежности $\mu_{A_i}(x_1)$ и $\mu_{B_j}(x_2)$.

$Y = f(x_1, x_2)$ является функцией от входных переменных, она может быть выражена полиномом n-го порядка. Y = const (полином нулевого порядка) соответствует модели нечеткого вывода Сугено нулевого порядка с синглетонами в заключениях правил вывода.

При проведении эксперимента была сформирована база правил:

«Если x_1 = «малый» и x_2 = «большой», то класс «1» (y = 1)».

«Если x_1 = «малый» и x_2 = «малый», то класс «2» (y = 2)».

«Если x_1 = «большой» и x_2 = «малый», то класс «3» (y = 3)».

Для нечетких множеств входных переменных x_1 «большой» и x_2 «большой» значение выходной переменной не определено, т. к. по условиям эксперимента отсутствуют точки соответствующих входных параметров модели (см. рис. 1).

Степени выполнения условий нечетких правил и степени активации нечетких правил определялись с применением оператора MIN. Результирующие выходные значения нечеткого классификатора (классы текстов) рассчитывались на основе степеней активации $\mu_{A_i}(x)$ заключений отдельных правил f_i = const, где i = 1, 2, 3, по формуле $y = \dfrac{\sum_{i=1}^{3} \mu_{A_i}(x) f_i(x)}{\sum_{i=1}^{3} \mu_{A_i}(x)}$. Дробные значения на выходе модели округлялись до целых. В результате проведенного эксперимента точность классификации текстов составила 82 %.

Заключение

В работе проверена возможность применения теории нечетких множеств и нечеткой логики для классификации коротких технических текстов. Такой подход может быть применен для автоматизации работ с базой данных коротких технических текстов. Примером может служить база данных оборудования и приборов для систем инженерного проектирования. Классы текстов определялись по описаниям датчиков давления в зависимости от условий эксплуатации приборов и характеристик измеряемых величин.

При классификации учитывались особенности объектов исследования. На предварительной стадии из всех слов были удалены окончания, из текстов были исключены «стоп-слова», которые несут мало значимой информации, а также аббревиатуры, цифровые обозначения, обозначения физических величин, знаки препинания. Также были удалены редко встречающиеся слова. По частотам употребления терминов в документе была построена матрица «документ – термин», строки которой соответствуют документам, а столбцы – терминам. На пересечениях строк и столбцов указывалась частота встречаемости термина в определенном документе. Уменьшение размерности матрицы было выполнено с применением сингулярного разложения.

Для классификации текстов применялась система нечеткого вывода Сугено, которая используется для приблизительных рассуждений и дает возможность проводить классификацию объектов, информация о которых является неопределенной, размытой. Текстовые данные определяются как нечеткие данные. Нечеткость объектов исследования связана с нечеткостью естественного языка.

Целесообразность применения модели Сугено определяется простотой математической обработки данных без потери точности анализа. Нелинейные зависимости входных и выходных переменных преобразуются в кусочно-линейные функции. Каждый линейный сегмент соответствует одному правилу. В случае модели Сугено нулевого порядка аналитические расчеты еще более упрощаются – в заключениях правил модели функции выражены $y_i = f(x_i) = \text{const}$. В алгоритме Сугено отсутствует дефаззификация выходных данных модели, т. к. на стадии аккумулирования заключений правил получаются четкие значения. Модель Сугено сочетает в себе описание объектов исследования на основе лингвистических правил и традиционного пред-

ставления в виде функциональных зависимостей. Такой подход значительно упрощает интерпретацию полученных результатов, делает их понятными и логически обоснованными.

Для классификации технических текстов модель нечеткого вывода Сугено может применяться как альтернативная классическим методам машинного обучения – Байеса, К-ближайших соседей, деревьев решений, опорных векторов. Однако она является значительно проще с точки зрения математической обработки текстов. Данная модель для описаний приборов (на примере датчиков давления) позволяет добиться точности классификации выше 80 %.

СПИСОК ЛИТЕРАТУРЫ

1. *О развитии* искусственного интеллекта в Российской Федерации (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 года»): Указ Президента РФ от 10.10.2019 г. № 490. URL: http://www.consultant.ru/document/cons_doc_LAW_335184/ (дата обращения: 15.10.2020).

2. *Pedrycz W., Chen S-M.* Sentiment analysis and ontology engineering: an environment of computational intelligence. Heidelberg: Springer, 2016. 456 p.

3. *Lane I. R., Kawahara T., Matsui T.* Dialogue Speech Recognition by Combining Hierarchical Topic Classification and Language Model Switching // IEICE – Transactions on Information and Systems. 2005. V. E88-D. Iss. 3. P. 446–454.

4. *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск. М.; СПб.; Киев: Вильямс, 2011. 520 с.

5. *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval // Information Processing & Management. 1998. V. 24. N. 5. P. 513–523.

6. *Mikolov T., Sutskever I., Chen K.* Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. 2013. P. 3111–3119.

7. *Карпович С. Н.* Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI // Тр. СПИИРАН. 2016. № 47 (4). С. 92–104.

8. *Полиниченко Д. Ю.* Естественный язык как лингвокультурный семиотический концепт: автореф. дис. … канд. филол. наук. Волгоград, 2004. 22 с.

9. *Tang D., Qin B., Liu T.* Document modeling with gated recurrent neural network for sentiment classification // Proceedings of the 2015 Conference on Empirical Methods in Natural Language processing. Lisbon, Portugal, 2015. P. 1422–1432.

10. *Kusner M., Sun Y., Kolkin N.* From word embeddings to document distances // Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015. V. 37. P. 957–966.

11. *Hartmann J., Huppertz J., Schamp C.* Comparing automated text classification methods // IJRM International Journal of Research in Marketing. 2019. V. 36. P. 20–38.

12. *Cagliero L., Garza P.* Improving classification models with taxonomy information // Data & Knowledge Engineering. 2013. V. 86. P. 85–101.

13. *Павлыгин Э. Д., Подлобошников А. Г., Савинов Р. А.* Разработка программного комплекса для интеллектуального анализа социальных медиа // Автоматизация процессов управления. 2019. № 2 (56). С. 23–36.

14. *Škrlj B., Martinc M., Kralja J., Lavrač N., Pollaka S.* Tax2vec: Constructing Interpretable Features from Taxonomies for Short Text Classification // Journal Pre-proof. Computer Speech & Language, Computer Speech & Language. 2020. V. 65. P. 101104. URL: https://doi.org/10.1016/j.csl.2020.101104 (дата обращения: 15.10.2020).

15. *Jaeger S., Fulle S., Turk S.* Mol2vec: Unsupervised machine learning approach with chemical intuition // Journal of Chemical Information and Modeling. 2018. V. 58 (1). P. 27–35.

16. *Kang M., Ahn J., Lee K.* Opinion mining using ensemble text hidden Markov models for text classification // Expert Systems with Applications. 2018. V. 94. P. 218–227.

17. *Дли М. И., Булыгина О. В., Козлов П. Ю.* Разработка экономической информационной системы автоматизированного анализа неструктурированных текстовых документов // Прикладная информатика. 2018. № 5 (77). С. 51–57.

18. *Zadeh L. A.* From computing with numbers to computing with words – from manipulation of measurements to manipulation of perceptions // IEEE Transactions on Circuits and Systems, I: Fundamental Theory and Applications. 1999. V. 4. P. 105–119.

19. *Zadeh L. A.* Fuzzy Sets // Information and Control. 1965. V. 8. № 3. P. 338–353.

20. *Пегат А.* Нечеткое моделирование и управление. М.: БИНОМ. Лаборатория знаний, 2013. 798 с.

21. *Mamdani E., Assilian S.* An experiment in linguistic synthesis with a fuzzy logic controller // Int. J. Hum. Comput. Stud. 1999. V. 51 (2). P. 135–147.

22. *Bergadano F., Cutello V.* Learning membership functions // European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. Granada, Spain, 1993. P. 25–32.

23. *Леоненков А. В*. Нечеткое моделирование в среде MATLAB и fuzzyTECH. СПб.: БХВ-Петербург, 2005. 736 с.

24. *Боровский А. В., Раковская Е. Е., Бисикало А. Л.* Дискриминантный анализ технических коротких текстов // Вестн. Астрахан. гос. техн. ун-та. Сер.: Управление, вычислительная техника и информатика. 2018. № 2. С. 53–60.

25. *Барсегян А. А., Куприянов М. С., Степаненко В. В.* Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. СПб.: БХВ-Петербург, 2007. 384 с.

Статья поступила с редакцию 09.12.2020

ИНФОРМАЦИЯ ОБ АВТОРАХ

Боровский Андрей Викторович – Россия, 664003, Иркутск; Байкальский государственный университет; д-р физ.-мат. наук; профессор кафедры математических методов и цифровых технологий; andrei-borovskii@mail.ru.

Раковская Елена Евгеньевна – Россия, 664003, Иркутск; Байкальский государственный университет; аспирант кафедры математических методов и цифровых технологий; rakovskaya19@mail.ru.

Бисикало Артем Леонидович – Россия, 664003, Иркутск; Иркутский государственный университет; канд. хим. наук; доцент кафедры аналитической химии; bisikalo.a@yandex.ru.

CLASSIFICATION OF SHORT TECHNICAL TEXTS USING SUGENO FUZZY INFERENCE SYSTEM

A. V. Borovskii[1], E. E. Rakovskaia[1], A. L. Bisikalo[2]

[1]Baikal State University,
Irkutsk, Russian Federation

[2]Irkutsk State University,
Irkutsk, Russian Federation

Abstract. The paper presents the results of classification of the short technical texts on the purpose of instruments using fuzzy sets theory and fuzzy logic. An important stage in designing special-purpose technical systems is the choice of equipment with specific operational characteristics. The need to categorize short technical texts, which present a brief description of equipment, annotations, fragments of databases, appears due to the fact that information about the equipment found in thematic abstract collections, technical and design documentation or in contextual advertising is often not structured and scattered. The other problems are a large number of typos, incorrect word usage and definitions in the texts. Much attention is paid to the characteristics of the objects of research and to recording their specific features – a large number of technical terms, abbreviations, symbols. The classifying technique is described, the expediency of application of fuzzy inference of Sugeno system associated with fuzziness of the natural language, the simplicity of mathematical calculations in the course of the experiment. A Sugeno model combines the description of the objects of research in the form of linguistic rules and functional dependencies. This approach greatly facilitates the interpretation of classification results.

Key words: short technical texts, fuzzy sets, Sugeno fuzzy inference system, classification.

For citation: Borovskii A. V., Rakovskaia E. E., Bisikalo A. L. Classification of short technical texts using Sugeno fuzzy inference system. *Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and Informatics*. 2021;1:16-27. (In Russ.) DOI: 10.24143/2072-9502-2021-1-16-27.

REFERENCES

1. *O razvitii iskusstvennogo intellekta v Rossiiskoi Federatsii (vmeste s «Natsional'noi strategiei razvitiia iskusstvennogo intellekta na period do 2030 goda»). Ukaz Prezidenta RF ot 10.10.2019 g. № 490* [On development of artificial intelligence in the Russian Federation (along with National Strategy for the Development of Artificial Intelligence for the Period up to 2030): Decree of the President of the Russian Federation of 10.10.2019, No. 490]. Available at: http://www.consultant.ru/document/cons_doc_LAW_335184/ (accessed: 15.10.2020).

2. Pedrycz W., Chen S-M. *Sentiment analysis and ontology engineering: an environment of computational intelligence.* Heidelberg, Springer, 2016. 456 p.

3. Lane I. R., Kawahara T., Matsui T. Dialogue Speech Recognition by Combining Hierarchical Topic Classification and Language Model Switching. *IEICE – Transactions on Information and Systems*, 2005, vol. E88-D, iss. 3, pp. 446-454.

4. Manning K. D., Ragkhavan P., Shiuttse Kh. *Vvedenie v informatsionnyi poisk* [Principles of information search]. Moscow, Saint-Petersburg, Kiev, Vil'iams Publ., 2011. 520 p.

5. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 1998, vol. 24, no. 5, pp. 513-523.

6. Mikolov T., Sutskever I., Chen K. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013, pp. 3111-3119.

7. Karpovich S. N. Mnogoznachnaia klassifikatsiia tekstovykh dokumentov s ispol'zovaniem veroiatnostnogo tematicheskogo modelirovaniia ml-PLSI [Multivalued classification of text documents using probabilistic thematic modeling ml-PLSI]. *Trudy SPIIRAN*, 2016, no. 47 (4), pp. 92-104.

8. Polinichenko D. Iu. *Estestvennyi iazyk kak lingvokul'turnyi semioticheskii kontsept. Avtoreferat dissertatsii ... kand. filol. nauk* [Natural language as linguocultural semiotic concept. Diss.Abstr.... Cand.Phil. Sci.]. Volgograd, 2004. 22 p.

9. Tang D., Qin B., Liu T. Document modeling with gated recurrent neural network for sentiment classification. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language processing*. Lisbon, Portugal, 2015. Pp. 1422-1432.

10. Kusner M., Sun Y., Kolkin N. From word embeddings to document distances. *Proceedings of the 32nd International Conference on Machine Learning.* Lille, France, 2015. Vol. 37. Pp. 957-966.

11. Hartmann J., Huppertz J., Schamp C. Comparing automated text classification methods. *IJRM International Journal of Research in Marketing*, 2019, vol. 36, pp. 20-38.

12. Cagliero L., Garza P. Improving classification models with taxonomy information. *Data & Knowledge Engineering*, 2013, vol. 86, pp. 85-101.

13. Pavlygin E. D., Podloboshnikov A. G., Savinov R. A. Razrabotka programmnogo kompleksa dlia intellektual'nogo analiza sotsial'nykh media [Development of software package for intellectual analysis of social media]. *Avtomatizatsiia protsessov upravleniia*, 2019, no. 2 (56), pp. 23-36.

14. Škrlj B., Martinc M., Kralja J., Lavrač N., Pollaka S. Tax2vec: Constructing Interpretable Features from Taxonomies for Short Text Classification. *Journal Pre-proof. Computer Speech & Language, Computer Speech & Language*, vol. 65, January 2021, April 2020. Pp. 101-104. Available at: https://doi.org/10.1016/j.csl.2020.101104 (accessed: 15.10.2020).

15. Jaeger S., Fulle S., Turk S. Mol2vec: Unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling*, 2018, vol. 58 (1), pp. 27-35.

16. Kang M., Ahn J., Lee K. Opinion mining using ensemble text hidden Markov models for text classification. *Expert Systems with Applications*, 2018, vol. 94, pp. 218-227.

17. Dli M. I., Bulygina O. V., Kozlov P. Iu. Razrabotka ekonomicheskoi informatsionnoi sistemy avtomatizirovannogo analiza nestrukturirovannykh tekstovykh dokumentov [Development of economic information system for automated analysis of unstructured text documents]. *Prikladnaia informatika*, 2018, no. 5 (77), pp. 51-57.

18. Zadeh L. A. From computing with numbers to computing with words – from manipulation of measurements to manipulation of perceptions. *IEEE Transactions on Circuits and Systems, I: Fundamental Theory and Applications*, 1999, vol. 4, pp. 105-119.

19. Zadeh L. A. Fuzzy Sets. *Information and Control*, 1965, vol. 8, no. 3, pp. 338-353.

20. Piegat A. *Fuzzy modeling and control*. Berlin, Physica-Verlag Heidelberg, 2001. 361 p. (In Rus.: Pegat A. Nechetkoe modelirovanie i upravlenie [Fuzzy modeling and control]. Moscow, BINOM. Laboratoriia znanii Publ., 2013. 798 p.).

21. Mamdani E., Assilian S. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Human-Computer Studies*, 1999, vol. 51 (2), pp. 135-147.

22. Bergadano F., Cutello V. Learning membership functions. *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty.* Granada, Spain, 1993. Pp. 25-32.

23. Leonenkov A. V. *Nechetkoe modelirovanie v srede MATLAB i fuzzyTECH* [Fuzzy modeling in MATLAB and fuzzyTECH environments]. Saint-Petersburg, BKhV-Peterburg Publ., 2005. 736 p.

24. Borovskii A. V., Rakovskaia E. E., Bisikalo A. L. Diskriminantnyi analiz tekhnicheskikh korotkikh tekstov [Discriminant analysis of short technical texts]. *Vestnik Astrakhanskogo gosudarstvennogo tekhnicheskogo universiteta. Seriia: Upravlenie, vychislitel'naia tekhnika i informatika*, 2018, no. 2, pp. 53-60.

25. Barsegian A. A., Kupriianov M. S., Stepanenko V. V. *Tekhnologii analiza dannykh: Data Mining, Visual Mining, Text Mining, OLAP* [Technologies for data analysis: Data Mining, Visual Mining, Text Mining, OLAP]. Saint-Petersburg, BKhV-Peterburg Publ., 2007. 384 p.

The article submitted to the editors 09.12.2020

INFORMATION ABOUT THE AUTHORS

Borovskii Andrei Viktorovich – Russia, 664003, Irkutsk; Baikal State University; Doctor of Physics and Mathematics; Professor of the Department of Mathematical Methods and Digital Technologies; andrei-borovskii@mail.ru.

Rakovskaia Elena Evgenievna – Russia, 664003, Irkutsk; Baikal State University; Postgraduate Student of the Department of Mathematical Methods and Digital Technologies; rakovskaya19@mail.ru.

Bisikalo Artem Leonidovich – Russia, 664003, Irkutsk; Irkutsk State University; Candidate of Chemical Sciences; Assistant Professor of the Department Analytical Chemistry; bisikalo.a@yandex.ru.