Раковская Елена Евгеньевна

аспирант

ФГБОУ ВО «Байкальский государственный университет»

г. Иркутск, Иркутская область

ПРИМЕНЕНИЕ СИСТЕМ НЕЧЕТКОГО ВЫВОДА ДЛЯ КЛАССИФИКАЦИИ ТЕКСТОВ

Аннотация: *в статье приводится теоретическое обоснование применения систем нечеткого вывода для классификации текстовой информации, описаны основные этапы анализа. Показана необходимость определения специфических признаков текста, применения одноэлементных множеств на выходе системы, а также перечислены методы кодирования обозначений классов текстов.*

Ключевые слова: *нечеткая логика, классификация текстов, фаззификация, одноэлементное нечеткое множество.*

В связи с ростом объемов доступных данных и увеличением скорости их передачи, в последнее время возрос интерес к проблеме интеллектуальной обработки информации, в том числе к различным видам классификации текстов разрозненных интернет-источников. Одно из направлений, связанных с решением этой проблемы, состоит в использовании аппарата нечетких систем: нечетких множеств, нечеткого моделирования, нечеткой логики и т. п. [1].

Поход на основе нечеткой логики, использующий нечеткие признаки текста и модели на основе нечеткого логического вывода, позволяет значительно улучшить качество классификации документов.

Определение признаков текста для классификации, предварительная обработка текста-оригинала с использованием лексического, морфологического анализа слов является очень важной ступенью в процессе анализа [2].

Предварительная обработка текста включает в себя:

– выделение специфических компонентов – формул, наименований валют и пр.;

– назначение увеличенных весов признакам текста в зависимости от расположения признаков – начале текста, в заголовке, в первом абзаце и пр.;

– назначение увеличенных весов ключевым словам, словам из названия.

Важная способность механизмов нечеткого логического вывода – принимать решения в условиях неопределенности, что делает их особенно подходящими для приложений, связанных с рисками, двусмысленностью.

Системы нечеткого вывода имеют возможность обрабатывать нечеткую естественно-языковую информацию, т.к. они обладают гибкостью и толерантностью к неточным значениям [3].

Разработка и применение систем нечеткого вывода включает в себя ряд этапов, базирующихся на основных положениях нечеткой логики. Информация, которая поступает в систему нечеткого вывода, определяется характеристиками текстов. Информация на выходе систем вывода соответствует выходным переменным, которые являются значениями, определяющими классы текстов («Политика», «Спорт», «Экономика»). Выходные переменные могут быть выражены числовыми значениями, обозначающими, например, номера классов.

В системах нечеткого вывода входные переменные преобразуются в выходные на основе использования нечетких правил продукций. Для этого системы вывода должны содержать базу правил нечетких продукций и реализовывать нечеткий вывод заключений на основе посылок, выраженных в форме нечетких лингвистических высказываний.

Таким образом, основными этапами нечеткого вывода при классификации текстов являются:

– формирование базы правил систем нечеткого вывода;

– фаззификация входных переменных;

– агрегирование подусловий в нечетких правилах продукций;

– активизация подзаключений в нечетких правилах;

– аккумулирование заключений в нечетких правилах продукций;

– дефаззификация выходных переменных.

Наиболее часто используемые модели нечеткого вывода – модели Мамдани и Сугено. Эти модели различаются только тем, как они получают выходные данные.

Система нечеткого вывода Мамдани является наиболее распространенной из-за простоты применяемых операций (например, max и min). В этой модели предполагается, что выходные переменные являются нечеткими множествами.

Метод Мамдани влечет за собой значительные вычислительные трудности. Эффективность вычислений снижается за счет того, что после процесса агрегации каждой выходной переменной имеется нечеткое множество, которое нуждается в дефаззификации.

Во многих случаях при проведении процесса классификации гораздо эффективнее использовать на выходе одноэлементное множество. Одноэлементное множество (singleton) – специальный случай нечеткого множества, состоящего только из одного элемента, значение функции принадлежности которого равно 1. Для всех остальных элементов универсума функция принадлежности этого множества равна 0.

В случае применения одноэлементных множеств для выходных переменных, процесс дефаззификации упрощается двух – трех простых операций, в зависимости от числа активных правил в конкретном случае классификации.

Применение одноэлементных множеств в системе классификации правомерно, т.к. классы текстов могут быть выражены категориальными значениями (переменными), а они, в свою очередь, преобразованы в уникальные числовые коды [1; 4]. В простейшем случае кодирование осуществляется с применением порядковых номеров классов. Данный метод используется, если «значения» классов допускают порядковую интерпретацию – «малый бизнес» – 1, «средний бизнес» – 2, «бизнес» – 3. При кодировании выходной переменной произвольным образом затрудняется решение задачи классификации, т.к. вносится несуществующая упорядоченность переменных. Оптимальное кодирование не должно искажать структуру соотношений между классами. Если классы не упо-

рядочены, то должна применяться схема кодирования для неупорядоченных категориальных признаков с помощью маски из двоичных цифр. В этом случае каждому уникальному значению ставится в соответствие двоичное число, например, 100, 010, 001 и т. д. При этом количество битов (то есть нулей и единиц) должно быть достаточным для обеспечения такого количества состояний маски, чтобы их хватило для кодирования всех уникальных «значений» классов. Еще один способ двоичного кодирования для классификации разнородных групп – применение фиктивных переменных, которые определяют граничные значения выходных признаков.

Список литературы

1. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям: Учеб. пособие / Н.Б. Паклин, В.И. Орешков. – СПб.: Питер, 2013. – 704 с.

2. Леонтьева Н.Н. Автоматическое понимание текстов. Системы, модели, ресурсы: Учеб. пособие / Н.Н. Леонтьева. – М.: Academa, 2006.

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

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