УДК 004.91:004.8

О. В. Золотухин

КЛАССИФИКАЦИЯ ПОЛИТЕМАТИЧЕСКИХ ТЕКСТОВЫХ ДОКУМЕНТОВ С ПРИМЕНЕНИЕМ НЕЙРО-ФАЗЗИ ТЕХНОЛОГИЙ

В статье рассматривается проблема интеллектуальной обработки текстовой информации. Представлена архитектура нечеткой нейросетевой системы для классификации текстовых документов и on-line алгоритм обучения сети адаптивного нечеткого векторного квантования.

Ключевые слова: *текстовый документ, нечеткая классификация, AFLVQ.*

1. Введение

В данной работе рассматриваются результаты научных исследований автора, относимые к областям искусственного и вычислительного интеллекта, а именно, задачи интеллектуальной обработки и анализа естественно-языковых текстов (ЕЯТ), и разработка нейро-фаззи технологий для решения задач классификации сложноструктурированных объектов, относящихся сразу к нескольким классам, каковыми являются политематические тексты.

Стремительное развитие Интернет и Web-технологий обусловили возможность широкого доступа пользователей к различного рода текстовым документам. При этом документы, подлежащие обработке, зачастую характеризуются разнородностью, широким охватом сразу нескольких тем, т. е. политематичностью. On-line классификация такого рода текстовых документов не является тривиальной задачей, поскольку в небольшом фрагменте текста может содержаться весьма ценная информация, и отнесение к соответствующему классу нельзя игнорировать, а близко расположенные классы могут пересекаться и/или сливаться. К такого рода документам могут относиться новостные потоки в сети Интернет, обзоры, дайджесты, формируемые новостными агентствами, научные публикации, посвященные нескольким областям исследований, причем как близким, так и далеким (например, медико-биологические, физико-химические, искусственный интеллект и информационные технологии, онтологические инжиниринг и автоматическая обработка текстов). Примеров подобного рода становится все больше. Таким образом, чрезвычайно актуальным направлением исследований является разработка методов классификации политематических текстовых документов.

2. Постановка проблемы

Задачей исследования является разработка методов классификации политематических текстовых документов, поступающих на обработку в реальном времени, с использованием нейро-фаззи технологий, обеспечивающих возможность отнесения одного документа сразу к нескольким темам.

3. Основная часть

3.1. Литературный обзор публикаций. Проблемная область интеллектуального анализа и обработки текстовой информации характеризуется множеством взаимосвязанных задач, которые, с одной стороны, могут решаться независимо (например, для конкретных предметных областей), с другой стороны, часто используют результаты предыдущих этапов обработки ЕЯТ для достижения конечного результата. Следует заметить, что, очевидно, в силу сложности такой проблемной области (ПО) как естественный язык, пока нет готовых коммерческих решений по реализации всего спектра задач интеллектуального анализа и обработки ЕЯТ. На сегодняшний день наиболее перспективным подходом для решения этой проблемы является онтологический подход [1—3], позволяющий с помощью средств Semantic Web специфицировать модель ПО в виде иерархически связанной системы концептов, снабженных соответствующими свойствами и отношениями. Возможные подходы к решению задачи on-line классификации политематических документов с возможностью отнесения одного документа сразу к нескольким темам, были предложены в [4, 5]. Введена нечеткая модификация вероятностной нейронной сети, отличающаяся крайне простой численной реализацией и малым объемом необходимой для ее реализации памяти.

3.2. Результаты личных исследований. В основу классификации положена искусственная нейронная сеть (ИНС) обучаемого векторного квантования, имеющая крайне простую однослойную архитектуру, настройка семантических весов которой производится в режиме обучения с учителем

ISSN 2226-3780

с элементами конкуренции по типу «победитель получает все». Основными преимуществами этой ИНС по сравнению с другими нейронными системами является простота архитектуры, незначительное количество входящих в нее нейронов, малый объем обучающей выборки и возможность on-line обучения, что крайне важно в задачах обработки текстовых документов.

Архитектура предлагаемой нейро-фаззи системы адаптивного обучаемого векторного квантования (AFLVQ) приведена на рис. 1. Система содержит два слоя обработки информации, при этом нейроны первого скрытого слоя связаны между собой латеральными связями, с помощью которых реализуются процессы конкуренции. Исходной информацией для обучения является последовательность векторов-образов $x(1), x(2), \ldots, x(k), \ldots, x(N), \ldots$; где $x(k)=(x_1(k), x_2(k), \ldots, x_n(k))^T \in R^n$ с известной классификацией, при этом входные сигналы предварительно нормируются так, что $\|x(k)\|=1$. Нейроны первого скрытого слоя N_j^c ($j=1, 2, \ldots, m$; m — априори задаваемое количество возможных классов) предназначены для нахождения прототипов (центроидов) классов $c_j(k)=(c_{j1}(k), c_{j2}(k), \ldots$ $\ldots c_{ji}(k), \ldots c_{jn}(k))^T$, при этом компоненты $c_{ji}(k)$ являются по сути настраиваемыми синаптическими весами нейрона N_j^c. Нейроны выходного слоя N_j^u вычисляют уровни принадлежности $u_j(k)$ предъявленного образа $x(k)$ к j-му классу.

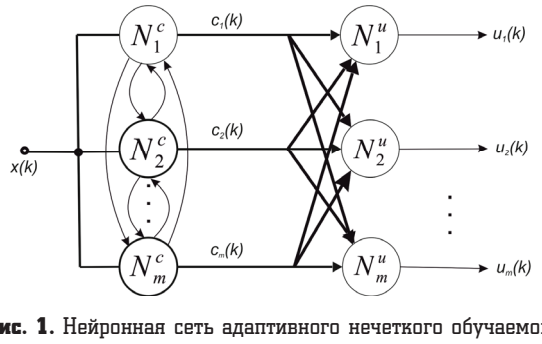

Рис. 1. Нейронная сеть адаптивного нечеткого обучаемого векторного квантования

Таким образом, предложена архитектура адаптивной нечеткой нейронной сети обучаемого векторного квантования и on-line алгоритм ее обучения, отличающийся вычислительной простотой и высоким быстродействием.

Литература

1. Рябова Н. В. Об онтологическом подходе к представлению лингвистических знаний в системах интеллектуальной интеграции информации [Текст] / **Н. В. Рябова** // Проблемы бионики. — 2001. — № 54. — С. 96—100.
2. Рябова Н. В. Методы и модели интеллектуальной обработки текстов в задачах онтологического инжиниринга [Текст] : материалы II Международного радиоэлектронного форума / **Н. В. Рябова**, В. В. Волкова, Я. В. Дыдыкина // Прикладная радиоэлектроника. Состояние и перспективы развития. — Харьков. — 2005. — Т. 3. — С. 87—90.
3. Бодянский Е. В. Интеллектуальная нейросетевая on-line технология обработки информации в автоматизированных системах принятия решений и управления [Текст] : зб. наук. пр. / Е. В. Бодянский, **Н. В. Рябова**, В. В. Волкова // Вісник СевДТУ. — Автоматизація процесів та управління. — Вип. 95. — 2009. — С. 20—23.
4. Бодянский Е. В. Классификация текстовых документов с помощью нечеткой вероятностной нейронной сети [Текст] / Е. В. Бодянский, **Н. В. Рябова**, О. В. Золотухин // Восточно-Европейский журнал передовых технологий. — 2011. — Вып. 6/2(54). — С. 16—18.
5. Бодянский Е. В. Обработка текстовых документов с помощью адаптивного нечеткого обучаемого векторного квантования [Текст] : зб. наук. пр. / Е. В. Бодянский, **Н. В. Рябова**, О. В. Золотухин // Вісник НТУ «Харківський політехнічний інститут». — Тематичний випуск: Нові рішення в сучасних технологіях. — 2011. — № 53. — С. 109—115.

КЛАСИФІКАЦІЯ ПОЛІТЕМАТИЧНИХ ТЕКСТОВИХ ДОКУМЕНТІВ ІЗ ЗАСТОСУВАННЯМ НЕЙРО-ФАЗЗІ ТЕХНОЛОГІЙ

О. В. Золотухін

У статті розглядається проблема інтелектуальної обробки текстової інформації. Представлена архітектура нечіткої нейромережевої системи для класифікації текстових документів та on-line алгоритм навчання мережі адаптивного нечіткого векторного квантування.

Ключові слова: текстовий документ, нечітка класифікація, AFLVQ.

Олег Вікторович Золотухін, *аспірант кафедри штучного інтелекту Харківського національного університету радіоелектроніки,* ***тел.: (067) 317-49-69,*** ***e-mail: zolotukhin.ov@gmail.com***.

MULTI-TOPIC TEXT DOCUMENT CLASSIFICATION BASED ON THE NEURO-FUZZY TECHNOLOGIES

O. Zolotukhin

This article discusses a problem of an intelligent text processing. Architecture of the neuro-fuzzy system is presented for classification of text documents and on-line learning algorithm for fuzzy network adaptive vector quantization.

Keywords: text document, fuzzy classification, AFLVQ.

Oleg Zolotukhin, *graduate student of Artificial Intelligence Department Kharkiv National University of Radioelectronics,* ***tel.: (067) 317-49-69, e-mail: zolotukhin.ov@gmail.com***.