Обробка інформації в складних організаційних системах

УДК 004.91:004.8

О.В. Золотухин

Харьковский национальный университет радиоэлектроники, Харьков

КЛАССИФИКАЦИЯ ПОЛИТЕМАТИЧЕСКИХ ТЕКСТОВЫХ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ НЕЧЕТКИХ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ

В статье рассмотрен нейросетевой подход, применяемый в задачах нечеткой классификации политематических текстовых документов.

Ключевые слова: *нечеткая классификация, политематический текстовый документ, нейронная сеть, текст на естественном языке.*

Введение

С каждым годом увеличивается объем доступных пользователю массивов текстовой информации, и поэтому становится все более актуальной задача поиска необходимых пользователю документов в таких массивах. Для решения этой задачи часто применяются различные тематические классификаторы, рубрикаторы и т.д., которые позволяют искать (автоматически или вручную) документы в небольшом подмножестве документной базы, соответствующем интересующей пользователя тематике.

Классификация текстовых документов для электронных библиотек рассматривается как один из возможных вариантов решения проблемы использования информационных ресурсов. Коротко она характеризуется следующим образом. К настоящему моменту различными хранилищами знаний (в том числе и библиотеками) накоплены огромные информационные массивы. Проблема заключается в сложности ориентирования в этих массивах, адекватной их размерам. Отсутствие возможности получить наиболее актуальную и полную информацию по конкретной теме делает бесполезной большую часть накопленных ресурсов. Поскольку исследование конкретной задачи требует все больших трудозатрат на непосредственный поиск и анализ информации по теме, многие решения принимаются на основе неполного представления о проблеме. Использование классификаторов позволяет сократить трудозатраты на поиск нужной информации, представленной электронными текстами. Использование нечеткой нейронной сети упрощает процедуру построения классификатора.

1. Формализация задачи

Классификацию текстовых документов на естественном языке называют рубрицированием, поэтому в статье эти термины считаются идентичными. Рубрикаторы подразделяют на три основных класса: плоские, иерархические и сетевые. Плоские рубрикаторы двухуровневые, на первом уровне размещается корневая, а на втором – дочерние к корневой рубрики. Иерархические и сетевые рубрикаторы могут быть представлены в виде сочетания нескольких плоских рубрикаторов.

Большинство существующих систем, работающих с документами, представляют собой электронный вариант архива документов со стандартным набором технических средств: пользовательский интерфейс, поддержка большого количества форматов, надёжное хранение, система ограничений прав доступа и т.п. Однако часто это оказывается недостаточным для современных информационных систем: во-первых, постоянный рост потока документов приводит к тому, что многие документы не доходят до фактического адресата, т.е. до тех, кто заинтересован в получении данного документа; во-вторых, в постоянно разрастающемся архиве становится трудно (практически невозможно) найти нужные документы. Актуальность задачи интеллектуальной обработки документов, в частности, состоит в преодолении этих проблем. Современные классификаторы документов должны решать задачу, связанную с управлением потоком входящих документов в режиме реального времени, – их автоматическую классификацию и последующий поиск в нем документов по содержанию.

Исходя из позиций Text Mining – это задача классификации, когда каждый документ может быть отнесен к одному из априори заданных классов, при этом предполагается, что априори задана обучающая выборка с известной классификацией, на основании которой формируются границы между этими классами. Классические методы распознавания образов в этой задаче малоэффективны, поскольку их

использование связано с гипотезой компактности и линейной разделимости классов. Для построения нелинейной разделяющей гиперповерхности между разными классами текстовых документов с успехом могут быть использованы искусственные нейронные сети (ИНС) [1,5,9], при этом предпочтение, естественно, отдается ИНС, обучение которых может производиться в on-line режиме, когда тексты на обработку поступают последовательно одним за одним. Задача существенно усложняется, когда один и тот же документ с различными уровнями принадлежности может одновременно относиться сразу к нескольким классам. В данной ситуации наиболее эффективными представляются методы нечеткой (фаззи) классификации [2], предназначенные для обработки данных, однозначная классификация которых в принципе невозможна.

1.1. Задача классификации текстов

Задача классификации определяется следующим образом: есть множество объектов $T = \{t_i\}$, не обязательно конечное, а так же множество $C = \{c_i\}$ $i = 1...N_c$, состоящее из N_c классов объектов. Каждый класс c_i представлен некоторым описанием F_i, имеющим некоторую внутреннюю структуру. Процедура классификации f объектов $t \in T$ заключается в выполнении преобразований над ними, после которых либо делается вывод о соответствии t одной из структур F_i, что означает отнесение t к классу c_i, либо вывод о невозможности классификации t. Применительно к текстам на естественном языке, элементами множества T являются электронные версии текстовых документов. Общая модель плоского текстового рубрикатора может быть представлена алгебраической системой вида:

$$R = \langle T, C, F, R_c, f \rangle, \qquad (1),$$

где T – множество текстов, подлежащих рубрицированию, C – множество классов, F – множество описаний, R_c – отношение на $C \times F$, f – операция рубрицирования вида $T \to C$. Отношение R_c имеет свойство: $\forall c_i \in C \exists F_i \in F : (c_i, F_i) \in R_c$, то есть классу соответствует единственное описание. Обратное требование необязательно. Отображение f не имеет никаких ограничений, так что возможны ситуации, когда $\exists t \in T : f(t) = C_t \subset C \wedge |C_t| > 1$, то есть некоторый текст может быть отнесен к нескольким классам одновременно.

Кроме сформулированной задачи классификации определяется задача обучения рубрикатора, под которой подразумевается частичное или полное формирование C, F, R_c и f на основе некоторых априорных данных [3].

1.2. Полнота и точность классификации

Существует несколько характеристик оценки качества работы текстового классификатора, их описание приведено в [4]. Наибольшее распространение получили точность (V) и полнота (U), применяемые так же при оценке качества естественно-языкового поиска, например, в поисковых машинах сети Интернет.

Для количественной оценки полноты и точности рубрикатора используются следующие измерения: a – число правильно рубрицированных документов, b – число неправильно рубрицированных документов, c – число неправильно отвергнутых документов. Под правильной и неправильной рубрикацией понимается случай, когда классификатор приписывает анализируемый документ некоторой рубрике, что расценивается некоторым экспертом соответственно, как верное и неверное решение. Под неправильным отвержением документа понимается случай, когда классификатор не приписывает документ рубрике, что, по мнению эксперта, неверно. На рис. 1 представлена иллюстрация этих случаев.

Рис. 1. Соотношение экспертной оценки и оценки классификатора

С учетом этого, оценка V и U имеет вид:

$$V = \frac{a}{a+b}; U = \frac{a}{a+c}$$

Согласно этой формуле, точность оценивает долю правильно классифицированных документов во всех документах, отнесенных к некоторой рубрике. Полнота оценивает долю правильно классифицированных документов во всех документах, которые необходимо было отнести к некоторой рубрике.

2. Искусственные нейронные сети в задачах классификации

Нейронные сети могут применяться при решении многих задач обработки информации, в частности в задачах распознавания образов. Как известно, искусственный нейрон выполняет следующие преобразования входного вектора

$$X = \{x_i\} : y = I(S); S = \sum w_i x_i,$$

где w_i – весовой вектор нейрона (веса синаптических связей), S – результат взвешенного суммиро-

вания, I – нелинейная функция активации нейрона. В терминах классификатора (1) X – соответствует внутренним описаниям $\{F_i\}$, а функции S и I – компоненты процедуры классификации f .

Функциональность нейрона проста, поэтому для решения конкретных задач нейроны объединяются в сети. Обучение классификатора, при условии, что выбрана топология сети и выбрана функция активации I, сводится к подбору весовых коэффициентов каждого нейрона. В данной работе рассматривается применение топологии адаптивного нечеткого обучаемого векторного квантования (AFLVQ).

2.1. Способы представления текста

Нейронные сети приспособлены обрабатывать только информацию, представленную числовыми векторами, поэтому для применения в обработке текстов на естественном языке их необходимо представлять в векторном виде. Используются два способа представления: полиграммная модель и модель терм – документ. В модели терм – документ [3] текст описывается лексическим вектором $\{\tau_i\}$ $i = 1...N_w$, где τ_i – важность (информативный вес) термина w_i в документе, N_w – полное количество терминов в документальной базе (словаре). Вес термина, отсутствующего в документе, принимается равным 0. Для удобства веса нормируются, так что $\tau_i \in [0,1]$. В данной работе использовались дискретные значения, так что присутствующий термин в тексте имеет вес 1, а отсутствующий – вес 0. Достоинствами данной модели являются:

— возможный учет морфологии, когда все формы одного слова соответствуют одному термину;

— возможный учет синонимии, так что слова - синонимы, объявляются одним термином словаря;

— возможность учета устойчивых словосочетаний, так что в качестве термина может выступать не отдельное слово, а несколько связанных слов, образующих единое понятие.

В качестве недостатков выделим следующее:

— при отсутствии простейшей дополнительной обработки, такой как морфологический анализ, существенно снижается качество классификатора, поскольку разные формы одного слова считаются разными терминами, вместе с тем морфологический анализ – весьма нетривиальная задача, требующая для ее решения привлечения лингвистов;

— размерность векторов $\{\tau_i\}$ зависит от общего количества терминов в обучающей выборке текстов, что в реальных задачах приводит к необходимости разрабатывать альтернативные структуры данных, отличные от векторов;

— словарь терминов может не охватывать всех документов, подлежащих классификации, так что анализируемые документы могут содержать значимые термины, не вошедшие в обучающую выборку, что отрицательно сказывается на адекватности модели.

В полиграммной модели со степенью n и основанием M текст представляется вектором $\{f_i\}, i = 1...M_n$, где f_i – частота встречаемости i -й n -граммы в тексте. n -грамма является последовательностью подряд идущих n – символов вида $a_1...a_{n-1}a_n$, причем символы a_i принадлежат алфавиту, размер которого совпадает с M . Непосредственно номер n -граммы определяется как

$$M^n \cdot r(a_n) + M^{n-1} \cdot r(a_{n-1}) + ... + r(a_1),$$

где $r(a_i) \in [1, M]$ – порядковый номер символа a_i в алфавите. Предполагается, что частота появления n -граммы в тексте несет важную информацию о свойствах документа. Предлагается рассматривать модель со степенью $n = 3$ (триграммная модель) и основанием $M = 33$, при этом применяется русский алфавит с естественной нумерацией символов $r("А") = 1)$, $r("Б") = 2), ..., r("Я") = 32$. Все остальные символы считаются пробелами с нулевыми номерами. Несколько подряд идущих пробелов считаются одним. С учетом этого размерность вектора для произвольного текста жестко фиксирована и составляет $33^3 = 35937$ элемента. Достоинствами полиграммной модели являются:

— отсутствие необходимости дополнительной лингвистической обработки;

— фиксированная размерность векторов и простота получения векторного описания текста;

К недостаткам отнесем следующее:

— отражение векторами $\{f_i\}$ содержания текста не всегда адекватно (такой моделью плохо отражается содержание небольших текстов; модель больше подходит для определения языка текста, чем для классификации по тематике),

— в соответствии с предыдущим пунктом возникает необходимость более тщательного подбора обучающей выборки текстов.

2.2. Адаптивная нечеткая нейронная сеть обучаемого векторного квантования

В основу предлагаемой системы положена искусственная нейронная сеть обучаемого векторного квантования (LVQ), имеющая однослойную архитектуру, настройка семантических весов которой производится в режиме обучения с учителем с элементами конкуренции по типу «победитель получает все». Основными преимуществами этой ИНС по сравнению с другими нейросистемами является простота архитектуры, незначительное количество входящих в нее нейронов, малый объем обучающей выборки и возможность on-line обучения [1], что крайне важно в задачах обработки текстовых документов. К настоящему времени известно множество

вариантов LVQ-нейросетей [6, 8], отличающихся выбором параметра шага обучения, используемой метрикой, необходимым объемом обучающей выборки. Эти системы подтвердили свою эффективность во многих приложениях, связанных с четкой классификацией и распознаванием образов.

Для решения задач нечеткой классификации в условиях пересекающихся классов был введен целый ряд модификаций LVQ-систем. Так, было введено нечеткое обучаемое векторное квантование FLVQ, представляющее собой по сути гибрид метода нечетких С-средних (FCM) и LVQ-сети и предназначенное для работы только в пакетном режиме. Были предложены нечеткие алгоритмы обучаемого векторного квантования (FALVQ), в которых с каждым вектором-прототипом класса связывается та или иная функция принадлежности, определяющая подобие каждого прототипа с предъявляемым вектором-образом. Можно отметить громоздкость этого подхода и субъективизм при выборе конкретной функции принадлежности. Введено нечеткое мягкое векторное квантование (FSLVQ), основанное на использовании мягкой конкуренции, ядерных функций соседства-принадлежности и, опять-таки, пакетной обработки данных. Весьма перспективным представляется подход, предложенный в [7] и представляющий собой гибрид нейронных сетей адаптивного резонанса (ART) и обучаемого векторного квантования. Данная система предназначена для работы в on-line режиме, однако весьма громоздка с вычислительной точки зрения.

Архитектура предлагаемой нейро-фаззи системы адаптивного обучаемого векторного квантования (AFLVQ) приведена на рис. 2. 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 $c_j(k) = (c_{j1}(k), c_{j2}(k), \ldots c_{ji}(k) \ldots c_{jn}(k))^T$, 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.

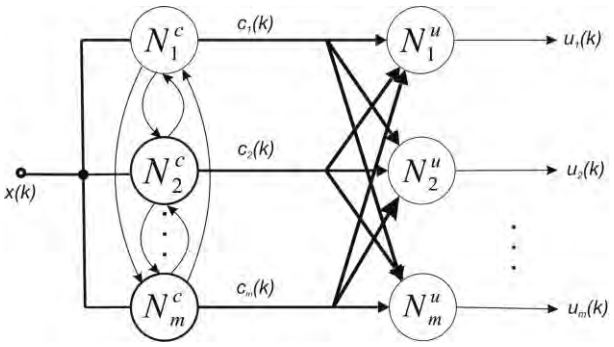

Рис. 2. Нейронная сеть адаптивного нечеткого обучаемого векторного квантования (AFLVQ).

При подаче на вход системы образа $x(k)$ в процессе конкуренции определяется нейрон-победитель j^*, синаптические веса которого $c_{j^*}(k-1)$ в смысле принятой метрики (в нашем случае евклидовой) наиболее близки к входному сигналу.

Поскольку обучение является контролируемым, то принадлежность вектора $x(k)$ к конкретному классу известна, что позволяет рассмотреть две возможные ситуации, возникающие в обучаемом векторном квантовании:

— входной вектор $x(k)$ и нейрон-победитель $N_{j^*}^c$ принадлежат одному классу;

— входной вектор $x(k)$ и нейрон-победитель $N_{j^*}^c$ принадлежат разным классам.

$$c_j(k) = \begin{cases} \dfrac{c_{j^*}(k-1) + \eta(k)\left(x(k) - c_{j^*}(k-1)\right)}{\left\| c_{j^*}(k-1) + \eta(k)\left(x(k) - c_{j^*}(k-1)\right)\right\|} - \\ \quad \text{åñëè } x(k) \text{è } c_{j^*}(k-1) \in \text{î äí î ì ó êëàññó}, \\ \dfrac{c_{j^*}(k-1) - \eta(k)\left(x(k) - c_{j^*}(k-1)\right)}{\left\| c_{j^*}(k-1) - \eta(k)\left(x(k) - c_{j^*}(k-1)\right)\right\|} - \\ \quad \text{åñëè } x(k) \text{è } c_{j^*}(k-1) \in \text{ðàçí ûì êëàññàì}, \\ \eta(k) = r^{-1}(k), r(k) = \alpha r(k-1) + 1, 0 < \alpha \le 1, \\ c_j(k-1) - \text{åñëè } j - \text{é íåéðîí íå ïîáåäèë}. \end{cases}$$

Рассчитанные с помощью правила обучения прототипы $c_j(k)$ ($c_j(N)$ в случае, если обучающая выборка имеет фиксированный объем) подаются на входной слой, где вычисляются уровни принадлежности

$$u_j(k) = \frac{\left\| x(k) - c_j(N) \right\|^{-2}}{\sum_{l=1}^{m} \left\| x(k) - c_l(N) \right\|^{-2}},$$

Эти соотношения задают on-line алгоритм обучения адаптивной нечеткой нейронной сети обучаемого векторного квантования.

Выводы

Рассмотрена задача автоматической классификации текстовых документов, поступающих на обработку в реальном времени. Предложена архитектура адаптивной нечеткой нейронной сети обучаемого векторного квантования (AFLVQ) и on-line алгоритм ее обучения, отличающийся вычислительной простотой и высоким быстродействием.

Список литературы

1. Umer M.F. Classification of textual documents using learning vector quantization / M.F. Umer, M.S.H. Khiyal // Information Technology Journal. – 2007. – 6(1). – P. 154-159.

2. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing / J.C. Bezdek, J. Keller, R. Krishnapuram, N.R. Pal. – N.Y.: Springer Science + Business Media, Inc., 2005. – 776 p.

3. Андреев А.М. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа [Электронный ресурс] / А.М. Андреев, Д.В. Березкин, В.В. Морозов, К.В. Симаков. – Электрон. текст. дан. – Режим доступа: http://www.inteltec.ru/publish/articles/textan/57_simakov.shtml, свободный.

4. Андреев А.М. Модели и методы автоматической классификации текстовых документов / А.М. Андреев, Д.В. Березкин, В.В. Сюзев, В.И. Шабанов // Вестн. МГТУ. Сер. Приборостроение. – М.: Изд-во МГТУ.- 2003.- №3. – С. 96-108.

5. Ciarelli P.M. An enhanced probobalistic neural network approach applied to text classification / P.M. Ciarelli, E. Oliveira // Lecture Notes on Computer Science.-V.5856. – Berlin- Heidelberg: Springer-Verlag, 2009. – P. 661-668.

6. Sanches J.S. An LVQ-based adaptive algorithm for learning from very small codebooks / J.S. Sanches, A.I. Marques // Neurocomputing. – 2006. – 69. – P. 922-927.

7. Kim Y.-S. Fuzzy neural network model using a fuzzy learning vector quantization with the relative distance / Y.-S. Kim, S.-I. Kim // Proc.7th Int. Conf. on Hybrid Intelligent System "HIS 2007". – Kaiserlautern, Germany, 2007. – P. 90-94.

8. Kohonen T. Improved version of learning vector quantization / T. Kohonen // Proc. Int. Joint Conf. on Neural Networks.- San Diego, CA, 1990.-1. –p.545-550.

9. Руденко О.Г. Штучні нейронні мережі / О.Г. Руденко, Є.В. Бодянський. – Х.: ТОВ «СМІТ», 2006. – 404 с.

Поступила в редколлегию 31.10.2012

Рецензент: д-р техн. наук, проф. В.А. Филатов, Харьковский национальный университет радиоэлектроники, Харьков.

КЛАСИФІКАЦІЯ ПОЛІТЕМАТИЧНИХ ТЕКСТОВИХ ДОКУМЕНТІВ ІЗ ВИКОРИСТАННЯМ НЕЧІТКИХ НЕЙРОМЕРЕЖЕВИХ ТЕХНОЛОГІЙ

О.В. Золотухін

У статті розглянуто нейромережевої підхід, застосовуваний в задачах нечіткої класифікації політематичних текстових документів..

Ключові слова: *нечітка класифікація, політематичний текстовий документ, нейронна мережа, текст на природній мові.*

POLYTHEMATIC CLASSIFICATION OF TEXT DOCUMENTS USING FUZZY NEURAL NETWORK TECHNOLOGY

O.V. Zolotukhin

This article presents a neural network approach is used in the fuzzy classification polythematic text documents.

Keywords: *fuzzy classification, polythematic text document, a neural network in natural language text.*