Сравнительный Анализ Метода Опорных Векторов и Использования Нейронных Сетей при Определении Авторства Текстов на Азербайджанском Языке

Камил Айда-заде[1,2], Сахават Талыбов[1,2]
[1]Бакинский Государственный Университет, Баку, Азербайджан
[2]Институт Систем Управления НАН Азербайджана, Баку, Азербайджан
Kamil_ayda-zade@rambler.ru, saxavat@yahoo.com

***Анотация*– В статье проведен сравнительный анализ результатов применения метода опорных векторов и использования математических моделей на базе нейронных сетей при распознавании авторства текстов. Применяемые признаки распознавания основаны на n-граммах при n=1. Приводятся результаты компьютерных экспериментов по распознаванию авторства текстов на азербайджанском языке.**

Ключевые слова— n-грамм, идентификация автора, распознавание, метод опорных векторов, нейронный сеть

I. ВВЕДЕНИЕ

В работах [1], [2] впервые для распознавания авторства азербайджанских текстов исследована частота использования букв и длины слов. Как было отмечено в [3],[4], работ, посвященных иследованию и разработке компьютерных систем распознавания авторства текстов на азербайджанском языке очень мало. Данная статья является продолжением работ, проводимых нами для определения авторства текстов небольших объемов на азербайджанском языке. Основная трудность распознавания авторства текстов (статей) малого объема на азербайджанском языке заключается в том, что в словах используется большое количество малоинформативных суффиксов, окончаний, а автоматический разбор слов на составные части для азербайджанского языка до сегодняшнего дня остается нерешенной проблемой. Кроме того для азербайджанского языка не разработан алгоритм стемминга для выделения корня слов.

II. ПОСТАНОВКА ЗАДАЧИ И ЕЕ РЕШЕНИЕ.

Формально постановку задачи идентификации авторства текстов можно описать следующим образом [4-5].

В базе данных имеются тексты n авторов и от каждого из них m_i текстов $D_{i,j}$ j = 1, ..., m_i , i = 1, ..., n.

Класс (группу) текстов i –того автора обозначим через Y_i. Рассматриваемая в статье задача состоит в том, что при появлении нового текста D требуется определить какому из n авторов или, другими словами, к какому классу Y_i эта текст принадлежит.

Введем следующие обозначения, определения и формулы.

Каждому из текстов $D_{i,j}$ и D сопоставим множество значений признаков $\left\{M_{i,j}^{s},\ s \in K_i\right\}$ и $\left\{d_s,\ s \in K,\ K = \bigcap_{s=1}^{n} K_s, i = 1,...,n,\ j = 1,...,m_i\right\}$, на основе которых происходит классификация текстов по авторам. Здесь K_i – множество признаков для определения авторства i-го автора, $i = 1,...,n$

Выделение признаков, основанное на n-грамах, для идентификации авторства проводится с применением следующих процедур.

1-ый шаг. Определяется частота встречаемости для каждой буквы(монограмы), входящих в класс Y_i i-того автора.

2-ый шаг. Объединяя все тексты каждого автора отдельно, определяются средние значения встречаемости по каждой букве (моноргам).

3-ий шаг. Определяется частота встречаемости каждой буквы(монограмы) или буквосочетаний(диграмы) нового произведения неизвестного автора, подлежащее идентификфции.

4-ый шаг. Выбирая структуру нейронной сети и применяя различные методы по ее обучению, определяется предполагаемый автор нового текста.

Алгоритмы, используемые для идентификации авторства текстов в основном можно разбить на три группы: алгоритмы, основанные на статических подходах; алгоритмы, основанные на методе опорных векторо; методы с использованием нейронных сетей.

DOI: 10.25045/NCSoftEng.2017.14

III. РЕЗУЛЬТАТЫ КОМПЬЮТЕРНЫХ ЭКСПЕРИМЕНТОВ

Для проверки и сравнения эффективности вышеизложенных алгоритмов, с целью обучения(параметрической идентификации) соответствующих математических моделей в базу данных были включены 50 газетных информационных статей на азербайджанском языке, взятых из Internet-а, случайно выбранных четырех авторов, условно названных *A1, A2, A3, A4*.

Для процеса обучения количество статей по авторам было взято соответственно для A1- 13 статей, для A2 -11 статей, для A3 -12 статей и для A4 -14 статей. Общее количество символов в каждой статье было в пределах от 3438 до 6859

Идентификация авторов проводилась соответственно по 8 дополнительным статьям одного из четырех авторов, авторство которых было скрыто Для распознавания были взяты две статьи $z^1 = (z_1^1, z_2^1)$ автора A1 (число букв 3926 и 5290), две статьи $z^1 = (z_1^2, z_2^2)$ автора A2 (число букв 3470 и 3740), две статьи $z^1 = (z_1^3, z_2^3)$ автора A3 (число букв 4067 и 4243) и две статьи $z^1 = (z_1^4, z_2^4)$ автора A4 (число букв 7463 и 4270).

В качестве признаков в случае применения монограммы (n=1) использовались 32 буквы азербайджанского языка.

IV. РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ МЕТОДА ОПОРНЫХ ВЕКТОРОВ

Метод опорных векторов (SVM – Support vector Machin) впервые был предложен В.Н. Вапником [8,9]. Наиболее популярная версия этого метода реализована в пакете LİBSVM [10]. Отметим, что благодаря современным разработкам многих исследователей можно констатировать, что SVM в настоящее время является одним их эффективных методов классификации (упорядочивания).

В качестве ядра для опорного вектора была выбрана функция gaussian radial basis function (RBF):

$$K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}, \gamma \geq 0.$$

Численные эксперименты проводились с помощью пакета LIBSVM с применением признаков, основанных на использовании монограм.

В таблице 1 приведены результаты распознавания авторства с использованием в качестве признаков 1-грамм.

Для обучения и распознавания использовался вариант второй обучающей выборки. Значение параметра γ было взято γ =1. Величина параметра штрафа *C* выбиралась равной 1, 10 и 100. При малых значениях параметра штрафа *C* точность распознавания составила 60-70% .

ТАБЛИЦА 1

	C=10				**C=100**			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
z_1^1	+				+			
z_2^1	+				+			
z_1^2		+				+		
z_2^2		+				+		
z_1^3			+				+	
z_2^3			+				+	
z_1^4				+				+
z_2^4				+				+

V. РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ МЕТОДОВ НА БАЗЕ НЕЙРОННЫХ СЕТЕЙ

Структура нейронной сети выбиралось двухслойной с 32 входами и одним выходам. В качестве функции активации на первом и втором слоях выбраны соответственно логарифимический сигмоид и положительная линейная функция.

В таблице 2 привелены результаты распознавания с применением нейронных сетей, обученных различными методами оптимизации. Все эксперименты проводились с использованием системы прикладных програм Matlab.

Как видно из результатов, приведенных в таблице 2, наиболее эффективным при обучении нейронных сетей являлся алгоритм метода упругого обратного распространения ошибки. Качество распознавания для всех остальных методов по сравнении с методами опорных векторов составило от 50 до 65 процентов.

ЗАКЛЮЧЕНИЕ

В статье проведено сравнение полученных результатов с применением обоих подходов.

Основными признаками являлись n-грамм при n=1. Алгоритмы распознавания строились с использованием метода опорных векторов и математических моделей на базе нейронных сетей. Объектом распознавания были авторы газетных информационных статей, характеризуемые малым объемом.

Проведенные компьютерные эксперименты позволили установить, что при использовании нейронных сетей на качество автоматической идентификации авторства для небольших объемов текстов существенно влияет выбор метода оптимизации, использованного при обучении нейронной сети.

DOI: 10.25045/NCSoftEng.2017.14

Кроме того, результаты распознавания зависят удачного выбора признаков, используемых для идентификации авторства.

На качество распознавания, очевидно, влияет и правильный выбор структуры нейронной сети.

ТАБЛИЦА 2

Методы оптимизаци при обучении	АВТОРЫ							
	A1		A2		A3		A4	
	z_1^1	z_2^1	z_1^2	z_2^2	z_1^3	z_2^3	z_1^4	z_2^4
Квази–Ньютоновский метод	+	+	+			+	+	+
Метод Пауэлла-Билла	+	+	+	+	+		+	+
Метод Флетчера-Пауэлла	+	+	+		+			+
Метод Полака-Рибьера	+			+		+		
Метод градиентного спуска	+	+	+		+			+
Метод градиентного спуска с адаптивным обучением	+	+	+		+			+
Метод градиентного спуска с учетом моментов	+	+	+	+	+			+
Метод градиентного спуска с учетом моментов и с адаптивным обучением	+	+	+		+			+
Метод Левенберга-Маркварта	+		+		+	+		+
Одноступенчатый метод секущих	+	+	+			+		+
Алгоритм упругого обратного распространения ошибки	+	+	+	+	+	+	+	+
Метод шкалированных связанных градиентов	+	+	+		+	+		+

Данная работа выполнена при финансовой поддержке Фонда Развития Науки при Президенте Азербайджанской Республики -Грант N EİF-KETPL-2-2015-1(25)-56/53/5

ЛИТЕРАТУРА

[1] S Gasimov., I.Ibrahimov. Analysis of sentences and words used in azerbaijani texts // The Second International Conference “Problems of Cybernetics and Informatics”, September 10-12, 2008, Baku, pp 117-119

[2] K.R.Aida-zade, S.G Talibov. Analysis of the effectiveness of the methods of recognition of authorship of texts in the Azerbaijani language // The 5th International Conference on Control and Optimization with Industrial Applications (COIA-2015), 27-29 August, 2015, Baku, Azerbaijan, pp. 183

[3] K.R. Aida-zade, S.Q. Talibov Authorship identification of the azerbaijani texts using n-grams. The 10th ieee intern. Conf. On application of information and communication technologies (aict2016), baku, 2016, 12-14 october

[4] R.P Айда-заде, С. Г. Талыбов Анализ методов определения авторства текстов на азербайджан-ском языке. İtb № 1, 2017, c. 15-22

[5] S. Doğan, B.Diri Türkçe Dokümanlar için N-gram Tabanlı Yeni Bir Sınıflandırma // Yazar, Tür ve Cinsiyet. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2010, 3, s.11–20

[6] G. Biricik, B. Diri, A. Sönmez A New Method For Attribute Extraction with Application on Text Classification / 5th International Conference on Soft Computing, Computing with Words, ICSCCW, North Cyprus, Famagusta, 2009, p 4

[7] V.N.Vapnik Statistical Learning Theory, New York: Wiley, 1998, 732 p.

[8] Vapnik V.N. The nature of statistical learning theory, New York: Springer-Verlag, 2000, 332 p.

[9] C.-W. Hsu, C.-C. Chan, C.-J. Lin A practical guide to support vector classification.// http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[10] А. С. Романов, Р.В. Мещеряков Идентификация автора текста с помощью аппарата опорных векторов в случае двух возможных альтернатив (Authorship identification with support vector machine in case of two possible alternatives) // http://www.dialog-21.ru/digests/dialog2009/materials/pdf/67.pdf

[11] D. I Holmes, Authorship attribution, computers and the humanities, vol.28, 1994, p 87-106

DOI: 10.25045/NCSoftEng.2017.14