УДК 519.68:007.5

Об идентификации языка искаженных текстов методом опорных векторов

А. В. Ермилов

Национальный исследовательский институт «Высшая школа экономики»
кафедра Управления разработкой программного обеспечения
ул. Мясницкая, д. 20, Москва, 101000, Россия

Рассматривается задача автоматического определения языка текстовых сообщений для случая, когда текст, язык которого нужно определить, подвергается случайным искажениям называемых «замена символа» с различными вероятностями. Приводятся результаты экспериментов по идентификации языка методом опорных векторов.

Ключевые слова: идентификация языка, алгоритм опорных векторов, n - граммы.

1. Введение

Метод опорных векторов (SVM — Support Vector Machines) широко применяется в естественно-языковых приложениях, например, при обработке речевых сигналов, а именно: определение языка, идентификация дикторов и других задач, в которых проводится классификация объектов [1].

В работах [2] рассматривается задача идентификации языка искаженного текста в случае искажений типа «замена символа», производимых с вероятностью от 0 до 1.

Цель данной статьи — исследование практической эффективности метода опорных векторов при решении задачи идентификации языка искаженного текста в зависимости от степени искажения и сравнение с результатами [2], полученными с использованием мультиграммных моделей.

2. Описание экспериментов

Рассматривается закрытая задача идентификации английского, испанского, польского и французского языков (предполагается, что новый текст написан на одном из рассматриваемых языков). Эксперименты проводятся с текстами рассказов А. Конан-Дойля о Шерлоке Холмсе, записанными в латинице без пробелов и знаков препинания в одном регистре. Тексты разбиваются на два подмножества: обучающее (около 150000 символов для каждого языка) и тестовое (около 120 векторов длиной 1000 символов для каждого языка).

Тексты подвергались искажениям типа «замена». С вероятностью P_z буква исходного текста заменяется любой другой буквой латинского алфавита (выбор замещающей буквы осуществляется случайно равновероятно).

Эксперименты проводились для различного количества обучающих векторов, которые формируются из обучающего множества. В качестве пространства признаков используются относительные частоты встречаемости символов и биграмм текста.

Методология использования SVM подробно изложена в [3]. При проведении экспериментов использовалось ядро RBF (Radial Basis Functions) [4], задаваемое в виде $K(X,Y) = \exp^{-\beta\|X-Y\|}$, где параметр β подвергался оптимизации методом

Статья поступила в редакцию 9 февраля 2012 г.

Автор приносит благодарность Кулаю Александру Юрьевичу и Мельникову Сергею Юрьевичу за предоставленные материалы и неоценимую помощь в написании статьи. Отдельную благодарность автор выражает Гостеву Ивану Михайловичу за помощь в редактировании и публикации статьи.

градиентного спуска. Сравнение эффективности при решении задачи классификации текстов SVM с RBF и полиномиальными ядрами приводится в работе [5], а с линейными — [6]. При сравнении классификаторы с RBF ядрами продемонстрировали лучшие результаты.

В качестве параметров метода SVM были использованы параметр β ядра и величина штрафа в задаче обучения SVM. Переход к многоклассовой классификации осуществлялся методом «каждый против каждого» (one vs. one) [7]. В пограничных случаях, когда несколько классов набрали одинаковое количество голосов распознавателей, победителем признается класс с наименьшим номером. Такой подход не является самым лучшим, но очень прост и часто используется на практике.

Для проведения экспериментов была использована реализация SVM утилитой pcp-2.2 [8]. Подбор параметра β ядра SVM осуществлялся на обучающем множестве с помощью метода кросс-валидации (cross-validation) [9].

3. Результаты экспериментов

Под точностью идентификации понимается отношение количества текстов с правильно идентифицированным языком к общему количеству текстов. Результаты экспериментов приведены на графиках зависимости точности идентификации от вероятности искажения.

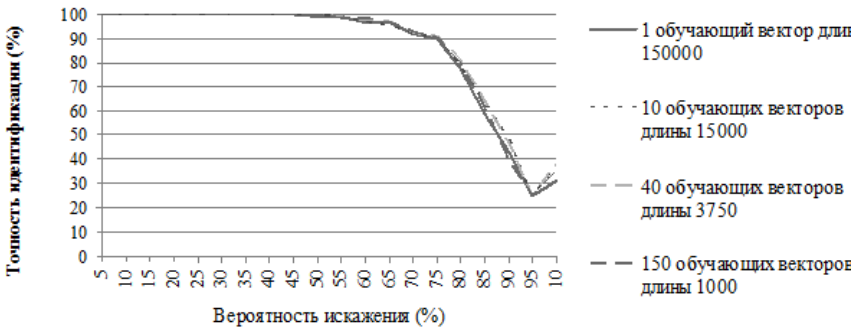

Рис. 1. Точность идентификации языка фрагментов текста длиной 1000 символов в зависимости от вероятности искажения P_z и количества обучающих векторов. SVM с RBF ядром. Значковые статистики

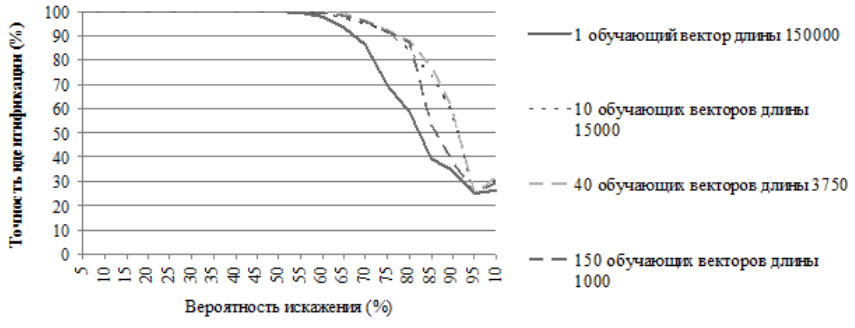

Рис. 2. Точность идентификации языка фрагментов текста длиной 1000 символов в зависимости от вероятности искажения P_z и количества обучающих векторов. SVM с RBF ядром. Биграммные статистики

Как видно из рис. 2, в рамках проведенных экспериментов на точность идентификации оказывает существенное влияния выбор количества обучающих векторов (и, соответственно, их длины). Объяснение этого результата следующее. Рассмотрим обучающие вектора как точки в пространстве признаков и предположим для простоты, что классов всего два и они линейно разделимы. Точки образуют области сложной формы. Малое количество обучающих векторов не позволяет судить о форме области, поэтому разделяющая гиперплоскость может быть проведена не оптимальным образом как показано на рис. 3.

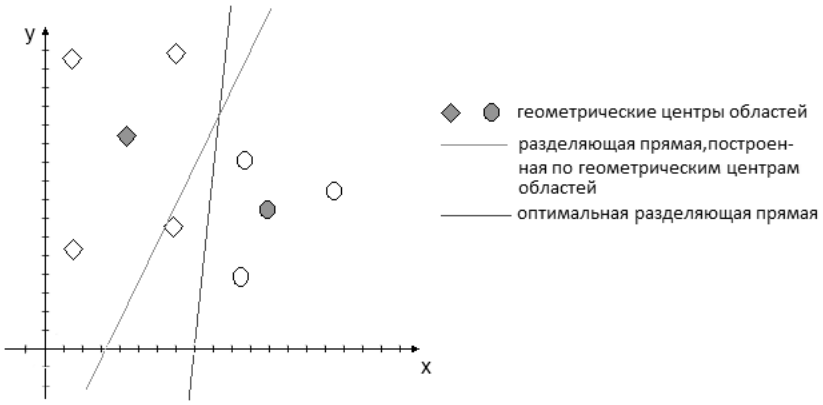

Рис. 3. Разделяющие прямые, построенные по обучающим векторам и геометрическим центрам областей

Увеличение числа обучающих векторов предоставляет больше информации о формах областей, позволяя проводить разделяющую гиперплоскость с их учётом, что увеличивает точность идентификации. При этом следует иметь в виду, что при ограниченном обучающем множестве такой подход приводит к уменьшению размера фрагментов текста, по которым формируются обучающие векторы. На маленьких фрагментах текста слабее проявляются статистические свойства, присущие данному языку. Это в свою очередь приводит к падению точности идентификации, что наблюдается в результатах экспериментов, показанных на рис. 2.

Рис. 4. Точность идентификации языка фрагментов текста длиной 1000 символов в зависимости от вероятности искажения P_z для SVM с RBF ядром (значковые статистики) и 1-граммной модели

Сравнивая результаты, полученные с использованием SVM и мультиграммных моделей [2], можно заметить, что на биграммных статистиках использование SVM может привести к увеличению точности идентификации (см. рис. 5), на значковых статистиках результаты сопоставимы (см. рис. 1 и рис. 4).

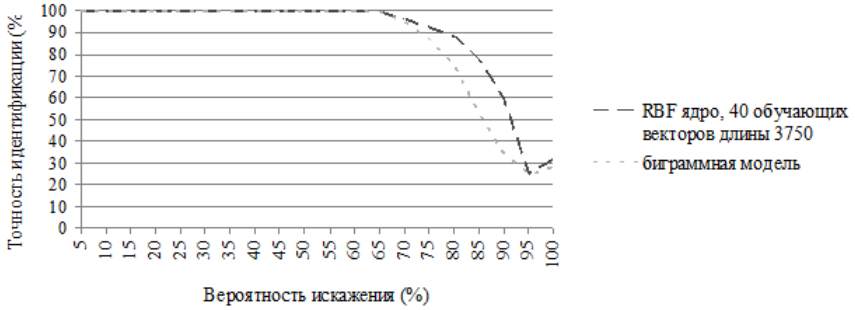

Рис. 5. Точность идентификации языка фрагментов текста длиной 1000 символов в зависимости от вероятности искажения P_z для SVM с RBF ядром (биграммные статистики) и биграммной модели

4. Заключение

В результате проведённой работы было установлено, что на биграмных статистиках SVM более эффективен, чем мультиграммные модели.

Кроме того, было установлено, что при идентификации языка искаженного художественного текста использование SVM с RBF ядром на биграммных статистиках даёт большую точность, чем мультиграммные модели.

Литература

1. Support Vector Machines for Speaker and Language Recognition / W. M. Campbell, J. P. Campbell, D. A. Reynolds et al. // Computer Speech and Language. — 2006. — Vol. 20. — Pp. 210–229.
2. *Кулай А. Ю., Мельников С. Ю.* О точности идентификации языка искаженного текста в зависимости от степени искажения // Концептуальный спектр изысканий в современном речеведении (Вестн. Моск. Гос. Лингвист. Ун-та, сер. Языкознание. — Вып. 575). — М.: ИПК МГЛУ «Рема». — 2009. — С. 200–209. [*Kulayj A. Yu., Meljnikov S. Yu.* O tochnosti identifikacii yazihka iskazhennogo teksta v zavisimosti ot stepeni iskazheniya // Konceptualjnihyj spektr izihskaniyj v sovremennom rechevedenii (Vestn. Mosk. Gos. Lingvist. Un-ta, ser. Yazihkoznanie. — Vihp. 575). — M.: IPK MGLU "Rema". — 2009. — S. 200–209.]
3. *Boser B. E., Guyon I. M., Vapnik V. N.* A Training Algorithm for Optimal Margin Classifiers // Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. — ACM Press, 1992. — Pp. 144–152.
4. *Buhmann M. D.* Radial Basis Functions: Theory and Implementations. Cambridge Monographs on Applied and Computational Mathematics. — Cambridge University Press, 2009. — ISBN 9780521101332. — `http://books.google.co.uk/books?id=-v2GPAAACAAJ`.
5. *Joachims T.* Text Categorization with Support Vector Machines: Learning with Many Relevant Features. — 1998.
6. *Teytaud O., Jalam R.* Kernel-Based Text-Categorization // In International Joint Conference on Neural Networks (IJCNN'2001. — 2000. — P. 1.
7. *Hsu C.-W., Lin C.-J.* A Comparison of Methods for Multiclass Support Vector Machines. — 2002.
8. *Buturović L. J.* PCP: a Program for Supervised Classification of Gene Expression Profiles // Bioinformatics. — 2006. — Vol. 22, No 2. — Pp. 245–247. — `http://bioinformatics.oxfordjournals.org/content/22/2/245.abstract`.
9. *Kohavi R.* A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. — Morgan Kaufmann, 1995. — Pp. 1137–1143.

UDC 519.68:007.5

About Language of Distorted Text Identification Using Support Vector Machines

A. V. Ermilov

National Research University "Higher School of Economics"
Department of Control of System Development
Myasnickaya str, 20, Moscow, 101000, Russia

In this article we consider a problem of language identification in a text message in case where the message is under stochastic distortion called "symbol change" with different probabilities. We provide experimental results in language identification using support vector machines.

Key words and phrases: language identification, support vector machines, n-gramms.