Forensic Data Mining for Email Author Identification with R Language

Sofia Marshallowitz Apuzzo
Interdisciplinary Bachelor of Science and Technology Institute / Law School
Federal University of ABC / Presbyterian University Mackenzie

Introduction

In forensic analyzes, cases involving text analysis are not uncommon. Several crimes have anonymous e-mail, whose nebulous authorship can be remedied with the aid of data analysis. In particular, a Supporting Vector Machine can facilitate the discovery of the author of the text.

This work presents an approach (still in development) to categorize the written idiolect of emails, with R language usage.

Idiolect Concept

Idiolect is an individual's distinctive and unique use of language, including the writing and the speech.

A notorious case of the use of the idiolect in forensic linguistic is the prison of Ted Kaczynski, the "Unabomber“, in 1998. FBI profilers identified a number of linguistic idiosyncrasies, like the use of academic terms and literature references, in an essay published by the Unabomber, which led them to conclude that Kaczynski was the author.

Support Vector Machine Classifier

The concept of SVMs is based on the idea of structural minimization of risk that minimizes the generalization error (that is, true error in unseen examples) which is limited by the sum of the error of the training set and a term that depends on the Vapnik-Chervonenkis (VC) size of the classifier and the number of training examples. The use of structural risk minimization performance contrasts with the empirical minimization of risk approach used by conventional classifiers.

Conventional classifiers attempt to minimize the training set error which does not necessarily reach a minimum generalization error. In case of outlier, the SVM looks for the best possible form of classification and, if necessary, disregards the outlier. Therefore, SVMs theoretically have a higher capacity generalize.

Although it does not work well on very large data sets, it requires matrix inversion - increasing computational complexity with up to the cube of the data volume and also not working well on data set with large amount of noise. If the classes are overlapping, only independent evidence (due to the fact that it does not deal well with data with many noises) should be used.

Authorship Categorization

For the authorship categorization, the first use of a basic subset of structural and stylometric characteristics was chosen in a set of authors without considering possible characteristics of the author's idiolect (gender, language, etc.) nor of the topic and size of the email.

Anderson et al [1] used a larger set of stylometric characteristics and studied the effect of various parameters such as, for example, the type of feature sets, the text size, and the number of documents by author, on the performance of the author's categorization for e-mail and text documents. This method based the development of the categorization of the present work.

In addition, some types of features such as N-graphs (where N = 2 was used) resulted in good categorization results for different sizes of pieces of text, but these results were thought to be due to an inherent bias of some types of N-graphs towards content rather than style alone (N-graphs are contiguous sequences of characters, including whitespaces, punctuation etc…).

Average sentence and word length, frequency, vocabulary richness, white-space and *hapax legomenon* are some of the attributes considered in this step.

The second factor of the author's classification brought attributes of the idiolect, in fact. The same logic previously applied was added information such as questions of email topic, gender, age, local terms and incidence of dialects.

In this situation, it was difficult to define the parameters that would be most appropriately associated with a class.

A third step is being planned for mere Unicode analysis to detect the use of other non-Latin alphabets and that cause homographic confusion (for example, Cyrillic "a" (U + 0430 in Unicode) and Latin "a" (U + 0041) are visually identical, but they are also an indication of the origin of the author).

e1071 R Package

The package *e1071* for R is, in the words of it creators, a *Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien* (the Technical University of Vien).

The **SVM** function is used to train the support vector machine. It can be used to carry out general regression and classification (of nu and epsilon-type), as well as density-estimation.

Final Considerations

The project is currently under development. The code is responsive and the option for the Support Vector Machine Classifier has proven to be the best option.

Some supervised tests with a dataset of three distinct authors were performed and the result was positive, especially when using classes that consider idioletcs and personality traits. This led to the conclusion that some limitations may arise in forensic practice, whereby the recommendation is that the universe be limited to suspects and that personal information about them be provided in advance.

Finally, the objective for the code is to be available in GitHub and even as a package in the official R Language repository, *The Comprehensive R Archive Network*.

References

1. A. Anderson, M. Corney, O. de Vel, and G. Mohay. “Identifying the Authors of Suspect E-mail”. Communications of the ACM, 2001. (Submitted) Osborne, Martin J. An Introduction to Game Theory New York: Oxford University Press, 2004
2. A. Anderson, M. Corney, O. de Vel, and G. Mohay. “Multi-topic E-mail authorship attribution forensics”. In Proc. Workshop on Data Mining for Security Applications, 8th ACM Conference on Computer Security (CCS’2001), 2001 . Nash, J. F. (1951) Ph.D. thesis (Princeton University, Princeton)
3. D. Lowe and R. Matthews. “Shakespeare vs Fletcher: A stylometric analysis by radial basis functions”. Computers and the Humanities, pages 449–461, 1995
4. P. Oman and C. Cook. “Programming style authorship analysis”. In Proc. 17th Annual ACM Computer Science Conference, pages 320–326, 1989.