Can punctuation marks be used as writer invariants? rough set-based approach to authorship attribution

Article · January 2008

CITATION
1

READS
158

2 authors:

Urszula Stanczyk
Silesian University of Technology
57 PUBLICATIONS **279** CITATIONS
SEE PROFILE

Krzysztof A. Cyran
Silesian University of Technology
69 PUBLICATIONS **259** CITATIONS
SEE PROFILE

Can punctuation marks be used as writer invariants? Rough set-based approach to authorship attribution

URSZULA STANCZYK, KRZYSZTOF A. CYRAN
Silesian University of Technology
Institute of Informatics
Akademicka 16, 44-100 Gliwice
POLAND

Abstract: Writer invariant is a stylometric notion corresponding to such unique characteristic that describes the writing style of a person, allowing for distinguishing texts authored by this person from all others and providing means for either discounting or confirming this person as the author of a text of unknown origin. It can be obtained in a variety of techniques usually belonging with either statistical analysis or machine learning methodologies and in this latter category there is included classical rough set theory and its elements. In the paper there is presented rough set-based approach to the problem of authorship attribution that falls within the scope of automated text categorisation.

Key–Words: Stylometry, Rough sets, Authorship attribution, Writer invariant, Decision table, Relative reduct

1 Introduction

Writer invariant (called also authorial or author's invariant) is considered as the primary stylometric concept. It is such a property of a text that is invariant of its author, which means that it is similar for all texts written by the same author and significantly different in texts by different authors. Textual analysis of written texts that yields information on linguistic style of their authors and study of these styles themselves is called stylometry.

Author's invariant can be used to discover plagiarism, recognise the real authors of anonymously published texts, for disputed authorship of literature in academic and literary applications, and even in criminal investigations in the area of forensic linguistics.

It is generally agreed that writer invariants exist, but the question what features of a text can constitute writer invariants is being argued for decades if not centuries. Some researches propose to use lexical properties, while others prefer syntactic, structural or content-specific.

In the early years of its history stylometric analysis was extremely tedious task of going through several texts by some author and by comparing them finding some similarities. It exploited human ability of perceiving some noticeable patterns or striking elements. In contrast, modern stylometry employs computational power of computers to continuously growing corpus of texts available via the Internet and can study common parts of speech, which is more reliable as they are used by writers subconsciously and such individual habits are less likely to be imitated by other authors.

Contemporary analytical techniques applied to stylometric tasks rely usually either on statistic-oriented computations, or artificial intelligence techniques. As the representatives of the first group there should be mentioned cluster analysis, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Markovian Models (MM), Cumulative Sum (CUSUM or QSUM), while from the latter there can be used Genetic Algorithms (GA), Artificial Neural Networks (ANN), Rough Set Theory (RST), decision trees, Support Vector Machines (SVM). Obviously these lists of techniques are not exhaustive.

In the paper there is presented application of Classical Rough Set-based methodology to the problem of author identification for literary texts. Rough Set Theory, developed by Zdzislaw Pawlak [6] in the early 1980s, deals with the problem of imperfect knowledge that has been studied by scientists for many years. Such imperfect or incomplete knowledge can be interpreted and manipulated in many ways, probably the most popular of which is provided by the fuzzy set theory due to Lotfi Zadeh [9]. Classical Rough Set Theory provides tools for succinct description of knowledge about the Universe by means of relative reducts and relative value reducts and resulting decision algorithms can be used for classification purposes with satisfying accuracy.

2 Stylometry

Stylometric analysis of written texts provides descriptors of linguistic style for their authors which can be used to study of these styles and to identification of authors of anonymous or disputed documents [1]. The applications of stylometry are academic and literary as well as legal in case of forensic linguistics employed in criminal investigations.

Textual features selected must be sufficiently distinct for each writer as to constitute the writer invariant, such characteristic that remains unchanged for all documents by this writer and different in texts by other authors. Since modern stylometry operates rather on electronic formats of documents rather than on handwritten manuscripts, in such context writer invariants are also called "cyber fingerprints" or "writerprints".

Linguistic descriptors are usually classified into four categories: lexical, syntactic, structural and content-specific. As lexical attributes there are used such statistics as total number of words, average number of words per sentence, distribution of word length, total number of characters (including letters, numbers and special characters such as punctuation marks), frequency of usage for individual letters, average number of characters per sentence, average number of characters per word. Syntactic features describe such patterns of sentence construction as formed by punctuation, structural attributes reflect the general layout of text and elements like font type or hyperlink, and as content-specific descriptors there are considered words of higher importance or with specific relevance to some domain [8].

Selection of features that is one of crucial factors in stylometric studies is not only task-dependent problem, but also to some degree determined by techniques employed.

2.1 Statistical approaches

Statistical analytical techniques used in stylometry rely on computations of probabilities and distribution of occurrences for single characters, words, word patterns, patterns of sentences [7].

In Markovian Model approach a text is considered as a sequence of characters corresponding to a Markov chain [3]. Since all constituent elements do not appear at random, in fact letters are dependent on these which precede them, in the simplest model only the immediately preceding letter is considered giving rise to 1st order Markov chain. For all pairs of letters there are calculated matrices of transition frequencies and statistics are obtained for all texts by known authors and the true author of an unknown text is found out as the one with the highest probability.

Another statistical method of author attribution, QSUM or CUSUM, was developed by Jill M. Farringdon [2]. It this method through cumulative sum there are calculated (and compared one against another) distributions of features for analysed texts of known and unknown authorship, for the first of which there is used average sentence length while for the second either the use of the 2 and 3 letter words, using words starting with a vowel, or the combination of these two together. If graphs match, the author is identified.

Linear Discriminant Analysis, Principal Component Analysis and cluster analysis are examples of multivariate methods that from their definition aim at reducing multidimensional data sets to lower dimension, by looking for linear combinations of variables that best explain data (LDA and PCA) or partitioning data into subsets (cluster analysis) described by some distance measure.

2.2 Machine learning approaches

Machine learning algorithms are characterised by their efficiency when dealing with large data sets. Not only do they achieve high accuracy in classification tasks, but are also popularly used in feature extraction process.

Artificial Neural Networks are often employed in classification tasks and authorship analysis certainly is an example of these. Application of ANN to stylometric tasks can be seen as the procedure the first step of which is to built the network with random weights associated with connections. Then the network is presented with training samples of texts of known authorship. As long as recognition is incorrect, weights are adjusted until the network can properly identify known texts. Then the network can be used for recognition of unknown texts [4].

The genetic algorithm approach starts out with definition of a set of rules expressing some characteristic of texts. Then these rules are tested against a set of known texts presented to the program and each rule is given a fitness score basing on which some rules are disregarded, leaving only these with highest scores (selection). These are slightly modified (mutation) and some new rules are added. The process is repeated until the rules that evolved correctly attribute authors to texts.

Decision trees represent a special type of classifier, which is trained by repetitive selection of individual features that stand out at each node of the tree. In classification procedure there are considered only those features that are required for studied pattern. Quite often decision trees are binary with feature selection built in the structure which makes them sub-optimal for most applications, yet they work fast.

Support Vector Machines are examples of a two-class classifier. As the criterion for optimisation there is considered the margin between the decision boundaries of the two classes, defined by the distance to the closest training patterns called support vectors. The classification function is defined by these patterns and their number is minimised by maximising the margin. The main drawback of this method is the computational complexity of the training procedure.

Rough Set Theory and its notions constitute yet another case of machine learning approaches and since they have been efficiently applied to the problem of authorship attribution presented in this paper, they are discussed in more detail in the next section.

3 Rough sets foundations

The fundamental concept of Rough Set Theory (RST) is the indiscernibility relation which using available information (values of attributes A) about objects in the Universe partitions the space into equivalence classes $[x]_A$ that are such granules of knowledge, within which single objects cannot be discerned [6].

While in classic set theory elements are either included or not included in a set, in RST the indiscernibility relation leads to lower $\underline{A}X$ and upper approximations $\overline{A}X$ of sets, the first comprised of objects whose whole equivalence classes are included in the set, the second consisting of objects whose equivalence classes have non-empty intersections with the set. If the set difference between the upper and lower approximation of some set is not empty then the set is said to be rough.

Sometimes among the attributes describing objects of the Universe there are distinguished two classes, called *conditional* attributes C and *decision* attributes D. Then information about the Universe can be expressed in the form of Decision Table.

3.1 Decision tables

Decision Table (DT) is defined as 5-tuple

$$DT =< U, C, D, v, f > \tag{1}$$

where U, C, and D are finite sets (U being the Universe, C set of conditional attributes and D set of decision attributes), while v is a mapping which to every element $a \in C \bigcup D$ assigns its finite value set V_a (domain of attribute a), and f is the information function $f : U \times (C \bigcup D) \rightarrow V$, where V is a union of all V_a and $f(x, a) = f_x(a) \in V$ for all x and a.

For each decision table there is defined its consistency measure $\gamma_C(D^*)$ which answers the question whether the table is deterministic. All decision rules provided by rows of DT are compared and if there are at least two that have the same values of conditional attributes but different for decision attributes D, the table is not deterministic.

The consistency measure $\gamma_C(D^*)$ of Decision Table is equal to the C-quality of the approximation of the family D^*

$$\gamma_C(D^*) = \frac{card\,(POS_C(D^*))}{card(U)} \tag{2}$$

where the C-positive region of the family D^*, with $POS_C(D^*)$ defined as

$$POS_C(D^*) = \bigcup_{X_i \in D^*} \underline{C} D_i \tag{3}$$

3.2 Relative reducts and value reducts

It may often happen that information contained in a Decision Table is excessive in this sense that not all values for all conditional attributes are necessary for correct classification indicated by decision attributes. Rough Set Theory provides tools for finding, if they exist, such functional dependencies between conditional attributes which may lead to reduction of their number without any loss of information.

A set of attributes $R \subseteq C$ is called relative reduct of C with respect to D or D-reduct of C ($RED_D(C)$) if R is the maximum independent subset of C with respect to D. If R is D-reduct then $POS_R(D^*) = POS_C(D^*)$ and $C \stackrel{k}{\rightarrow} D$ implicates $R \stackrel{k}{\rightarrow} D$. Attribute $c \in C$ is redundant in C with respect to D (D-redundant) if $POS_C(D^*) = POS_{C-\{c\}}(D^*)$, otherwise the attribute c is irremovable from C with respect to D (D-irremovable).

A relative core of C with respect to D (D-core of C) is the set of all D-irremovable attributes of C

$$\begin{aligned} &CORE_D(C) = \{c \in C: \\ &POS_C(D^*) \neq POS_{C-\{c\}}(D^*)\} \end{aligned} \tag{4}$$

The relation between D-reduct and D-core is given by the following formula

$$CORE_D(C) = \bigcap_{R \in RED_D(C)} R \tag{5}$$

Further reduction of the Decision Table is achieved by such elimination of some values of an attribute for some elements of the Universe (without eliminating the attribute itself) that does not diminish the classification abilities of DT for this set of attributes. That leads to the concept of relative value reduct (D-value reduct) and the core of value reducts (D-value core) [5].

4 Experiments

In stylometric research there were used texts from the 4 novels by famous Polish writers, Henryk Sienkiewicz ("Potop" and "Krzyżacy") and Bolesław Prus ("Lalka" and "Faraon"). The training set consisted of 36 rules (4×9 samples from each novel) and following the same guidelines also 36 testing rules were chosen from another set of 4 novels ("Rodzina Połanieckich" and "Quo vadis" by Sienkiewicz, and "Emancypantki" and "Placówka" by Prus). The choice of novels to short works is explained by the wider corpora that enables not only higher cardinality of both training and testing data sets but also ensures that text samples are long enough to be representative. For short texts frequencies of neither function words nor punctuation marks are reliable descriptors and thus could not be considered as writer invariants.

There were counted frequencies of 8 punctuation marks: a comma, a semicolon, a full stop, a bracket (assuming that when we have "(" also ")" follows, such occurrence is counted as single), a quotation mark, an exclamation mark, a question mark, and a colon. Obviously counting frequencies returned continuous values for attributes which are not directly applicable in classic rough set methodology. Thus the issue of discretisation needed to be considered.

The simplest imaginable discretisation is thresholding that returns binary data yet firstly the threshold value has to be selected. For this purpose there were used 2-quantiles for each of conditional attributes independently on others, as specified by the Table 1.

Table 1: 2-quantiles of frequencies

Attribute	Attribute median frequency
,	$MF_{\{,\}} = 0.101128$
;	$MF_{\{;\}} = 0.003055$
.	$MF_{\{.\}} = 0.110114$
(	$MF_{\{(\}} = 0.000128$
"	$MF_{\{"\}} = 0.003881$
!	$MF_{\{!\}} = 0.012082$
?	$MF_{\{?\}} = 0.010168$
:	$MF_{\{:\}} = 0.006575$

The next step was to specify decision attributes, their number and values. In the presented experiments with text samples to be attributed to one out of two writers one decision attribute D was enough and its values are used to denote which author is recognised, $D = 1$ indicates Prus while $D = 0$ points to Sienkiewicz. Hence the Decision Table 2 for the D being set describes works by Prus while the Table 4 corresponds to the reset state of D and works by Sienkiewicz.

Table 2: Decision Table for $D = 1$

	Conditional attributes							
R	,	;	.	(	"	!	?	:
1	0	0	1	1	1	0	1	0
2	1	1	0	1	0	0	0	0
3	1	0	1	1	0	0	1	1
4	1	0	1	1	0	0	1	0
5	0	0	1	1	1	0	0	0
6	0	0	1	1	0	1	1	0
7	0	0	1	1	0	0	1	1
8	0	0	1	1	0	1	1	0
9	0	1	1	1	0	0	0	0
10	0	1	1	1	1	1	1	1
11	1	1	1	0	0	1	1	0
12	0	0	1	0	0	0	1	0
13	0	1	1	1	1	0	1	0
14	0	1	1	0	1	0	1	0
15	0	1	1	1	1	1	0	1
16	0	1	1	1	1	1	1	1
17	0	1	1	1	1	0	0	0
18	0	1	1	1	1	0	0	1

When DT is specified it is necessary to answer the question whether it is deterministic. Fortunately the consistency measure $\gamma_C(D^*)$ equals 1, thus the table is deterministic and it is possible to continue the procedure to the phase of finding relative reducts.

By rough set analysis for this Decision Table there were obtained several relative reducts, comprised of conditional attributes as specified by the Table 3. By comparing them it is clear that the core is composed of a comma and a bracket as these attributes are present in all relative reducts.

Table 3: Generated relative reducts

	Conditional attributes				
RED_1	,	;	.	(	"
RED_2	,	;	(	"	?
RED_3	,	;	(	!	?
RED_4	,	.	(	!	
RED_5	,	(	!	?	:

Since the 4th relative reduct on the list is the only one with 4 conditional attributes instead of 5 as it is in all other cases, it is the one that was chosen.

After limiting DT to include only these conditional attributes included in the selected relative reduct, the next step was to apply the notion of the relative value reducts to all decision rules in the Decision Table which returned 6 subsets of conditional attributes, 5 with cardinality of 2

Table 4: Decision Table for $D = 0$

	Conditional attributes							
R	,	;	.	(	"	!	?	:
19	1	0	0	0	1	0	0	1
20	1	0	0	1	0	1	0	1
21	1	0	0	0	1	0	0	1
22	1	0	0	0	1	1	1	1
23	1	0	0	0	0	1	1	1
24	1	0	0	0	1	0	0	1
25	0	0	0	0	0	0	0	1
26	1	0	0	0	1	0	0	1
27	1	0	0	0	1	1	1	1
28	1	1	0	0	0	0	0	0
29	0	1	0	0	0	1	1	1
30	1	1	0	0	0	1	0	0
31	1	1	0	0	1	1	0	1
32	1	0	0	0	0	1	0	0
33	0	1	1	0	0	1	1	0
34	0	1	0	0	0	1	0	0
35	1	1	0	1	1	1	0	0
36	1	1	0	1	1	1	1	1

{, .} {. !} {(!} {, (} {. (}
and one including 3 elements
{, (!}

For majority of decision rules there were several possible relative value reducts, but for some of the rules (2, 11, 12, 14 and 33) there was only one choice and these relative value reducts had to be chosen. There were 4 such necessary relative value reducts and to complete the list one more had to be added which led to the Table 5 of selections.

Table 5: Selected relative value reducts

VR	Rule numbers
, .	11, 25
. !	1, 3, 4, 5, 7, 9, 12, 13, 14, 17, 18, 22, 23, 27, 30, 31, 32
(!	2
, (!	19, 20, 21, 24, 26, 28, 29, 33, 34, 35, 36
, (	6, 8, 10, 15, 16

As a result of application of the selected relative value reducts there was obtained the new Decision Table 6, from which multiplied rows were eliminated (leaving only first occurrences from the list).

Automatic knowledge processing technique applied to the Table 6 results in Decision Algorithm in which there were incorporated medians of frequencies previously used in the discretisation of the continuous input space. With such approach testing examples in fact do not have to be discrete. The Decision Algorithm consists of two "If ... then ..." sentences, one per each value of the decision attribute D. The conditional sentences are composed of inequalities checking frequencies of attributes indicated by relative value reducts.

Table 6: DT limited to relative value reducts

	Attributes				
R	,	.	(	!	D
1		1		0	1
2			1	0	1
6	0		1		1
11	1	1			1
19	1		0	0	0
20	1		1	1	0
22		0		1	0
25	0	0			0
29	0		0	1	0

PRUS ($D = 1$) If:
$(F_{\{,\}} \geq MF_{\{,\}}$ AND $F_{\{!\}} < MF_{\{!\}})$ OR
$(F_{\{(\}} \geq MF_{\{(\}}$ AND $F_{\{!\}} < MF_{\{!\}})$ OR
$(F_{\{,\}} < MF_{\{,\}}$ AND $F_{\{(\}} \geq MF_{\{(\}})$ OR
$(F_{\{,\}} \geq MF_{\{,\}}$ AND $F_{\{.\}} \geq MF_{\{.\}})$
SIENKIEWICZ ($D = 0$) If:
$(F_{\{,\}} < MF_{\{,\}}$ AND $F_{\{!\}} \geq MF_{\{!\}})$ OR
$(F_{\{,\}} < MF_{\{,\}}$ AND $F_{\{.\}} < MF_{\{.\}})$ OR
$(F_{\{,\}} \geq MF_{\{,\}}$ AND $F_{\{(\}} < MF_{\{(\}}$ AND $F_{\{!\}} < MF_{\{!\}})$ OR
$(F_{\{,\}} \geq MF_{\{,\}}$ AND $F_{\{(\}} \geq MF_{\{(\}}$ AND $F_{\{!\}} \geq MF_{\{!\}})$ OR
$(F_{\{,\}} < MF_{\{,\}}$ AND $F_{\{(\}} < MF_{\{(\}}$ AND $F_{\{!\}} \geq MF_{\{!\}})$

The Decision Algorithm was then subjected to testing for verification of classification accuracy.

5 Results and discussion

For validation purposes there was used the same number of testing samples as training ones, that is 36. The obtained results are given in the Table 7 into 3 categories of verdict: as correct classification, incorrect classification and undecided, in relation to the total number of testing examples.

Table 7: Classification results

Classification verdict	Ratio
correct	30/36
incorrect	5/36
undecided	1/36

The overall classification accuracy is satisfactory 30/36%=83.3%, yet it is interesting to study which testing samples where incorrectly classified, especially when considered in the context of coverage of input space provided by training and testing data. Since each conditional attribute is binary and there are 4 of them, there are 16 points in the discrete input space. Table 8 consists of rows described by coordinates present either during training (columns denoted by "Tr") or testing (columns denoted by "Ts").

Table 8: Coverage of the discrete input space by training and testing data

Values of attributes				Prus		Sienkiewicz		Result of
,	.	(	!	Tr	Ts	Tr	Ts	class.
0	0	0	0			1		
0	0	0	1			2		
0	0	1	0		1			c
0	0	1	1		1			n
0	1	0	0	2	1			c
0	1	0	1			1		
0	1	1	0	7	9			c
0	1	1	1	5	4			c
1	0	0	0			5	13	c
1	0	0	1			6		
1	0	1	0	1			4	n
1	0	1	1			3		
1	1	0	1	1				
1	1	1	0	2	1			c
1	1	1	1		1			cn

Training data is present only in 12 rows of the table, which means that coverage of the input space is 75% and for 25% there are no representatives within the training set. On the other hand, testing data is contained only in 9 rows, which means the coverage of the input space in 56.25% which is even less than in case of training, yet there are examples of incorrect classification denoted by "n" in the right-most column. The result denoted by "cn" means that while part of the Decision Algorithm gave correct answer, the other returned incorrect thus resulting in undecided verdict.

The cases of incorrect or undecided classification happened for samples that were either absent or poorly represented in the training data set and thus the classifier had insufficient information for creating some decision rule with correct classification dedicated to them. In this kind of situation the Decision Algorithm certainly can fail yet not necessarily always, as can be seen in the third row of the table. It is noteworthy that all training facts that appeared several times within the set were later properly recognised.

6 Conclusions

Presented results of authorship attribution obtained with rough set-based methodology were satisfactory but recognition accuracy can be enhanced in the future by applying different discretisation approaches, another choice of descriptors to work as writer invariants, i.e. incorporating the usage of function words, and widening the set of training data to contain texts from short stories that are likely to have less uniform distributions of selected features which can help to tune-in the rough set-based classifier.

Acknowledgements: The software used in the research to obtain frequencies of punctuation marks was implemented by P. Cichoń under supervision of K. Cyran, in fulfilment of requirements for MSc thesis.

References:

[1] S. Argamon and J. Karlgren and J.G. Shanahan, eds., Stylistic analysis of text for information access, *Proceedings of the 28th International ACM Conference on Research and Development in Information Retrieval*, Brazil, 2005.

[2] W. Buckland, Forensic semiotics, *The Semiotic Review of Books* 10(3), 1999.

[3] D.V. Khmelev and F.J. Tweedie, Using Markov chains for identification of writers, *Literary and Linguistic Computing* 16(4), 2001, pp. 299–307.

[4] R.A.J. Matthews and T.V.N. Merriam, Distinguishing literary styles using neural networks, in E. Fiesler and R. Beale, eds., *Handbook of neural computation,* Oxford University Press, 1997, pp. G8.1.1–6.

[5] M.J. Moshkow and A. Skowron and Z. Suraj, On Covering Attribute Sets by Reducts, in M. Kryszkiewicz and J.F. Peters and H. Rybinski and A. Skowron, eds., *Lecture Notes in Artificial Intelligence* 4585, Springer-Verlag, Singapore, 2007, pp. 175–180.

[6] Z. Pawlak, Rough Set Rudiments, *Institute of Computer Science Report, Warsaw University of Technology, Poland*, 1996, pp. 1–47.

[7] R. Peng, Statistical Aspects of Literary Style, *Bachelor's Thesis, Yale University*, 1999.

[8] R.D. Peng and H. Hengartner, Quantitative analysis of literary styles, *The American Statistician* 56(3), 2002, pp. 15–38.

[9] L.A. Zadeh, A fuzzy-algorithmic approach to the definition of complex or imprecise concepts, *International Journal on Man-Machine Studies* 8, 1976, pp. 249-291.