On employing elements of Rough Set Theory to stylometric analysis of literary texts

Urszula Stańczyk, Krzysztof A. Cyran

***Abstract*—**Rough Set Theory deals with imperfect knowledge about the Universe and objects that comprise it, and such knowledge can be interpreted and manipulated in many ways. The paper describes application of rough set-based approach to the problem of stylometric analysis of texts that by the notion of writer invariant enables to identify authors of disputed or unattributed works. Writer invariant is the primary stylometric concept which corresponds to such unique characteristic that expresses the writing style of a person, allowing for distinguishing texts authored by this person from all others. It can be obtained in a myriad of techniques usually belonging with either statistical analysis or machine learning methodologies, with RST addressed in this paper being the example of the latter category.

***Keywords*—**Stylometry, rough sets, authorship attribution, relative reduct, decision algorithm.

I. Introduction

HUMAN ability to express oneself by speaking or writing is distinctive to such degree that allows for not only distinguishing humans from other species, but also different nationalities and even individuals with perfectly the same background and lifetime experiences. Textual analysis that brings characterisation of an author's writing style is called stylometry. Stylometry belongs with automated text categorisation and information retrieval tasks [1].

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.

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 for example to verify ransom notes.

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 and it still stands open. 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 an extremely tedious task of going through several texts by some author (the more the better) and by comparing them finding some similarities andshared properties. It exploited human ability of perceiving some noticeable patterns or striking elements.

In contrast, modern stylometry employs computational power of contemporary computers to continuously growing corpus of texts available via the Internet and can study even common parts of speech, which is much more reliable as they are used by writers subconsciously and such individual habits are less likely to be imitated by other authors. Such scientific investigation into numerical measurement of style goes back to the late nineteenth century and the works of Mendenhall [2] who as the first proposed to use quantitative as opposed to qualitative text descriptors.

Contemporary analytical techniques applied to stylometric tasks rely usually either on statisticoriented computations, or artificial intelligence techniques [3]. As the representatives of the first group there should be mentioned cluster analysis, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Markovian Models (MM), or 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, or Support Vector Machines (SVM). Obviously for all these techniques there can be studied several solutions and neither of these lists is 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 Polish scientist Zdzislaw Pawlak [4] 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 [5].

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

K. A. Cyran is with the Institute of Informatics, Silesian University of Technology, Gliwice, 44-100 Poland, phone: +48-32-237-2500; fax: +48-32- -237-2733; e-mail: krzysztof.cyran@ polsl.pl.

U. Stańczyk is with the Institute of Informatics, Silesian University of Technology, Gliwice, 44-100 Poland, phone: +48-32-237-2969; fax: +48-32- -237-2733; e-mail: urszula.stanczyk@ polsl.pl.

decision algorithms can be used for classification purposes with satisfying accuracy [6].

II. Stylometry

The notion underlying stylometric research is that works by different authors can be distinguished by quantifiable features of the text. Stylometric analysis 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 [7]. The applications of stylometry are academic and literary as well as legal in case of forensic linguistics employed in criminal investigations.

The three aims of stylometry are:

- author characterisation that brings conclusions about social background, education and even gender,
- similarity detection which allows to find some properties shared by different authors,
- and authorship attribution considered to be of the primary importance, which answers the question of author identity.

Textual features selected as descriptors for analysis 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", "cyberprints" or "writerprints". This is true also for electronic formats of historic documents in Greek for example, which in original are too valuable to be handled [8].

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 (organization into headings or paragraphs) and elements like font type, embedded pictures or hyperlink, and as content-specific descriptors there are considered words of higher importance or with specific relevance to some domain [2].

Many measures used are strongly dependent on the length of the text being studied and so are difficult to apply reliably. Thus selection of features is one of crucial factors in stylometric studies and it is not only task-dependent problem, but also to some degree determined by techniques employed.

A. Statistical Approaches

Statistical analytical techniques used in stylometry rely on computations (with the use of computers due to high computational complexity) of probabilities and distribution of occurrences for single letters, characters, words, word patterns, sentences, patterns of sentences [9].

In Markovian Model approach a text is considered as a sequence of characters corresponding to a Markov chain [10]. In natural language model all constituent elements do not appear at random, in fact letters are dependent on these which precede and succeed them. In the simplest model only the immediately preceding letter is considered giving rise to 1st order Markov chain. For all pairs of letters and other characters there are calculated matrices of transition frequencies from one into another 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 [11].

Another statistical method of author attribution, Cumulative Sum (QSUM or CUSUM),was developed by Jill M. Farringdon [12]. In this method through cumulative sum there are calculated and plotted as graphs (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 the two 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.

B. 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 [13].

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 some 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 authors of known texts. Then the network can be used for recognition of unknown texts [14].

The genetic algorithm approach starts out with definition of a set of rules expressing some characteristics of texts [3]. 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 next 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 a 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 the studied pattern. Quite often decision trees are binary with feature selection built in the structure which makes them suboptimal 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 [15] 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.

III. 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 input space into some number of equivalence classes $[x]_A$ that are such granules of knowledge, within which single objects cannot be discerned [4].

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 $\bar{A}X$ 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, called the boundary region of this set, is not empty then the set is said to be rough, otherwise it is crisp.

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.

A. 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 $a \in C \cup D$ assigns its finite value, set V_a (domain of attribute *a*), and *f* is the information function f: $U \times (C \cup D) \rightarrow V$, where *V* is a union of all V_a and $f(x, a) = f_x(a) \in U$ for all *x* and *a*.

Thus Decision Table in its columns specifies all attributes defined for objects within the Universe, both conditional and decision ones, while rows provide values of these attributes for all objects. Each row constitutes also the decision rule as for specified values of condition attributes the values of decision attributes are provided.

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, one by one against all others, 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^*, $POS_C(D^*)$, is defined as

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

B. Relative Reducts and Relative Value Reducts

It may often happen that information contained in a Decision Table is excessive in this sense that either not all conditional attributes or not all their values are necessary for correct classification indicated by decision attributes. Rough Set Theory provides tools for finding, if they exist, such functional dependencies [16] between conditional attributes which may lead to reduction of their number without any loss of information and they involve the concept of a reduct.

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

$$POS_R(D^*) = POS_R(D^*) \quad and \quad C \xrightarrow{k} D \Rightarrow R \xrightarrow{k} D \tag{4}$$

Attribute $c \in C$ is redundant in *C* with respect to *D* (*D*-redundant) if

$$POS_C(D^*) = POS_{C-\{c\}}(D^*) \tag{5}$$

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*.

$$CORE_D(C) = \{c \in C : POS_C(D^*) \neq POS_{C-\{c\}}(D^*)\} \tag{6}$$

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

following formula

$$CORE_D(C) = \bigcap_{R \in RED_D(C)} R \quad (7)$$

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) [17].

It is said that a value of attribute $c \in C$ is *D*-dispensable for $x \in U$ if

$$C(x) \subseteq D(x) \Rightarrow C_c(x) \subseteq D(x) \quad (8)$$

otherwice the value of attribute *c* is *D*-dispensable for *x*. If for every attribute $c \in C$ value of *c* is *D*-indispensable for *x*, then *C* is called *D*-independent for *x*.

Subset $C' \subseteq C$ is a relative value reduct of *C* if and only if *C'* is *D*-independent for *x* and

$$C(x) \subseteq D(x) \Rightarrow C'(x) \subseteq D(x) \quad (9)$$

The set of all *D*-indispensable for *x* values of attributes in *C* is called the relative value core of *C* for *x* and denoted by COExD(C), with the property

$$CORE_D^x(C) = \bigcap RED_D^x(C) \quad (7)$$

where $RED_D^x(C)$ is the family of all *D*-reducts of *C* for *x*.

Relative reducts can be perceived as masks put on decision rules included in the decision table, indicating for each rule these attributes whose values are sufficient to perform correct classification. It is quite common that for a decision rule several distinct relative value reducts can be used and this results in the necessity of choice among them.

IV. Experiments

In stylometric research described in this paper there were used punctuation marks to work as writer invariants. Such choice of syntactic textual descriptors over others, for example lexical, even though they seem to be more natural to employ, is explained by the fact that while punctuation marks undeniably express the structure of the text, construction of sentences, they are applied in less conscious way by writers than for example some function words which can be more easily imitated. Such individual habits as of adding some emphasis expressed with a question or exclamation mark are less likely to be copied.

In experiments within the training phase there were used texts from the 4 novels by famous Polish writers, Henryk Sienkiewicz - a winner of the Nobel Prise ("Potop" and "Krzyżacy") and Bolesław Prus ("Lalka" and "Faraon") and the training set consisted of thirty six rules (4×9 samples from each novel).

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. Thus both training and testing samples ware created as files of comparable length, using the chapter structures whenever possible.

By using dedicated software 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 always counted as single and not double), a quotation mark, an exclamation mark, a question mark, and a colon. Obviously the software counting frequencies returned continuous values for all attributes which are not directly applicable in classic rough set methodology that works on discrete data. Thus the issue of discretisation needed to be considered.

Discretisation is not a trivial problem with two contradicting goals: one of reflecting the input data in the closest possible way even at the cost of in-depth study of distributions of values and introducing many ranges of them, and other of limiting the representation to some few values. Since this problem was not of primary importance to the research addressed in this paper only the simplest discretisation was applied.

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

Table 1. 2-quantiles of occurrence frequencies for conditional attributes.

Attribiute	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 just one decision attribute D was enough and its values are used to denote which author is recognised, D = 1indicates 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 part for decision attribute 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.

The close look at the specified Decision Table revealsthat the knowledge contained in it is excessive because for example rules with numbers 19, 21 and 24 are exactly the same. Such repetitions are not erroneous yet can be considered as advantageous only in more sophisticated approaches when the multiple instances of decision rules work as additional confirmation for these rules that increases their classification power.

Also rules 28 and 30 differ only in the value of attribute {!}. If this attribute is disregarded the rules become the same, and no other rule within the whole Decision Table is in contradiction. Such reasoning obviously gives only a hint how systematic procedures of Rough Set Theory employed to the DT find all possibilities for diminishing the size of the table yet still maintaining the full classification properties of the original table.

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 four conditional attributes instead of five as it is in all other cases, it is the one that was chosen for the following computations.

Table 4. Decision Table part for decision attribute 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

In the study presented in this paper it was not the case, but it may happen that for some additional reasons some of conditional attributes are more important than others. In such case not only the cardinality of a relative reduct to be chosen should be taken into considerations but also which conditional attributes are contained in it. In the stylometric experiments performed none particular punctuation mark can be perceived as of foremost importance thus the only criterion for selecting a relative reduct is the minimal number of attributes it is comprised of.

After limiting the Decision Table to include only these conditional attributes included in the selected relative reduct (that is a comma, a full stop, a bracket and an exclamation mark), 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

{, .}
{. !}
{(!}

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

For majority of decision rules there were several possible relative value reducts found (from which any could be chosen), as specified by the Table 5, where decision rules are grouped by their relative value reducts.

Table 5. Generated relative reducts.

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

For some of the decision rules (2, 11, 12, 14 and 33) only single value reducts could be used (belonging to the value core), while for others some selection was necessary. There were 4 such necessary relative value reducts and to complete the list just one more had to be added which led to the Table 6 of selections.

It should be noticed that some decision rules in the decision table reduced to the selected relative reduct were repeated - that is appeared more than once in the table. Whether a decision rule occurs once or many times, each occurrence results in the same set of possible value reducts to be selected for it. Such repetitions from the point of view of generated relative value reducts can be considered together and not separately one by one. Yet on the other hand for two rules having the same relative value reducts not necessarily the same has to be selected.

Actually, the choice of relative value reducts and how it influences the outcome decision algorithm can be considered as a separate optimisation problem yet to be solved.

Table 6. 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 7, from which multiplied rows were eliminated (leaving only first occurrences from the list and their rule numbers). Thus the final decision table contains just four rows for decision attribute D = 1and five for D = 0.

Table 7. 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

Automatic knowledge processing technique applied to the Table 7 results in Decision Algorithm consisting of two "If . . . then . . . " sentences, one per each value of the decision attribute *D*. The first sentence is comprised of four conditional clauses and five are included in the second sentence giving the total of nine, the number of rows in the reduced decision table. The Decision Algorithm can be presented as follows

D=1 If
({.}=1 AND {!}=0) OR
({(}=1 AND {!}=0) OR
({,}=0 AND {(}=1) OR
({,}=1 AND {.}=1)
D=0 If
({,}=1 AND {(}=0 AND {!}=0) OR
({,}=1 AND {(}=1 AND {!}=1) OR
({.}=0 AND {!}=1) OR
({,}=0 AND {.}=0) OR
({,}=0 AND {(}=0 AND {!}=1)

While keeping discretised attributes the decision algorithm can be perceived as a definition of a logic function and thus expressed in either CNF or DNF which are easily implementable by logic elements such as gates, commutators or popular programmable logic devices resulting in a hardware solution that is cheap and works fast.

On the other hand, this form of the algorithm can be modified by taking into account the fact that discretisation procedure applied to occurrence frequencies of punctuation marks that constitute conditional attributes is only needed during the construction phase of the algorithm. Once the decision algorithm is obtained, its conditional clauses can work on continuous data, which means that previously used medians of frequencies can be incorporated within the algorithm and testing examples in fact do not have to be discrete

Hence the Decision Algorithm is composed of the conditional sentences which are created from inequalities

checking frequencies of attributes indicated by relative value reducts.

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 $F_{\{(\}} \geq \mathrm{MF}_{\{(\}}$ AND $F_{\{!\}} \geq MF_{\{!\}}$) OR
($F_{\{,\}} < MF_{\{,\}}$ AND $F_{\{(\}} < MF_{\{(\}}$ AND $F_{\{!\}} \geq MF_{\{!\}}$)

The Decision Algorithm in its software implementation was then subjected to testing for verification of classification accuracy.

V. Results and Discussion

For validation purposes of the obtained rough set-based classifier in experiments there was used the same number of testing samples as training ones, that is 36. The obtained results are given in the Table 8 into three categories of total verdict per sample: as correct classification, incorrect classification and undecided, in relation to the total number of testing examples.

It should be understood that when some sample returned several partial classification verdicts from different constituent conditional clauses of the decision algorithm for the final verdict for this sample the decision was based on majority of verdicts when possible and in the case of tie verdict it is classified as undecided.

Table 8. Classification results for the individual samples.

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

The overall classification accuracy for all training samples considered separately is satisfactory 30/36%=83.3%, yet it can be presented as classification of whole novels to be attributed as specified by the Table 9.

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 just 4 of them, there are 24=16possible points in the discrete input space. Table 10 consists of rows described by coordinates of such points present either during the training (columns denoted by "Tr") or testing (columns denoted by "Ts") phase. Some points were present during just one phase and some appeared in both.

Table 9. Classification results for the whole novels.

Author	Text	Classification
Prus	"Emancypantki"	100%
	"Placówka"	77.8%
Sienkiewicz	"Rodzina Połanieckich"	66.7%
	"Quo vadis"	88.9%

Training data is present only in 12 (out of 16 possible) rows of the table, which means that coverage of the input space is 12/16%=75% and for 25% there are no representatives within the training set.

On the other hand, testing data is contained only in 10 rows, which means the coverage of the input space is 10/16%=62.5% which is even less than in case of training.

Table 10. Coverage of the discrete input space by training and testing data

Values of attributes				Prus		Sienkiewicz		Result of classification
,	.	(	!	Tr	Ts	Tr	Ts	
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	c
1	1	0	0					
1	1	0	1	1				
1	1	01	0	2	1			c
1	1	1	1		1			cn
Sum				18	18	18	18	

While comparing coordinates of points in the input discrete space for learning and testing samples it becomes evident that 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.

This observation brings immediate conclusion as to how the classification properties of rough set-based classifier can be enhanced. The higher coverage of the input space by the

training samples the higher chance of correct classification.

VI. Conclusions

The results of authorship attribution studies obtained with rough set-based methodology presented in this paper were satisfactory thus confirming that syntactic textual descriptors in form of punctuation marks expressing the text structure can be successfully used as writer invariants.

Yet possibly recognition and classification accuracy can be further enhanced by applying different discretisation approaches (for example quartiles instead of medians), another choice of descriptors to work as writer invariants, i.e. incorporating the usage of lexical features such as function words, and widening the set of training data to obtain full coverage of the discrete input space, and also to contain texts not only from novels but short stories as well, since they are likely to have less uniform distributions of selected features which can help to tune-in the rough set-based classifier.

Another direction of future research is also indicated by more detailed considerations of the choice of relative reducts and relative value reducts and how this selection reflects upon classification procedure.

Acknowledgment

The software used in the presented research to obtain frequencies of punctuation marks for texts was implemented by P. Cicho´n under supervision of K. Cyran, in fulfilment of requirements for MSc thesis.

References

[1] G. Bonanno and F. Moschella and S. Rinaudo and P. Pantano and V. Talarico, "Manual and evolutionary equalization in text mining", *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, pp. 262–267, 2007.

[2] R.D. Peng and H. Hengartner, "Quantitative analysis of literary styles", *The American Statistician*, vol. 56, no. 3, pp. 15–38, 2002.

[3] J.F. Jimenez and F.J. Cuevas and J.M. Carpio, "Genetic algorithms applied to clustering problem and data mining", *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, pp. 219–224, 2007.

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

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

[6] P. Jirava and J. Krupka, "Classification model based on rough and fuzzy sets theory", *Proceedings of the 6th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics*, pp. 198–202, 2007.

[7] 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.

[8] A. Lampropoulos and E. Galiotou and I. Manolessou and A. Ralli, "A finite-state approach to the computational morphology of early modern Greek", *Proceedings of the 7th WSEAS International Conference on Applied Computer Science*, pp. 242–245, 2007.

[9] R. Peng, "Statistical Aspects of Literarz Stzle, Bachelor's Thesis", *Yale University*, 1999.

[10] D.V. Khmelev and F.J. Tweedie, "Using Markov chains for identification of writers", *Literary and Linguistic Computing*, vol. 16, no. 4, pp. 299–307, 2001.

[11] D.T. Tran and T.D. Pham, "Markov and fuzzy models for written language verification", *WSEAS Transactions on Systems*, vol. 4, no. 4, pp. 268–272, 2005.

[12] W. Buckland, "Forensic semiotics", *The Semiotic Review of Books*, vol. 10, no. 3, 1999.

[13] M. Doumpos and A. Salappa, "Feature selection algorithms in classification problems: an experimental evaluation", *WSEAS Transactions on Information Science & Applications*, vol. 2, no. 2, pp. 77–82, 2005

[14] 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, pp. G8.1.1–6, 1997

[15] A. Caballero and K. Yen and Y. Fang, "Classification with diffuse or incomplete information", *WSEAS Transactions on Systems and Control*, vol. 3, no. 6, pp. 617–626, 2008.

[16] R.B. Perez, A. Nowe, P. Vrancx, Y. Gomez, and D.Y. Caballero, "Using Ant Colony Optimization and rough set theory to feature selection", *WSEAS Transactions on Information Science & Applications*, vol. 2, no. 5, pp. 512–517, 2005.

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

Stanczyk Urszula received her MSc and PhD degrees in computer science from the Silesian University of Technology, Gliwice, Poland in 1993 and 2003 respectively.

From 1993 till 2000 she was a teaching assistant, from 2000 till 2003 a lecturer, and from 2004 till present an assistant professor at the Institute of Informatics, SUT. From 2004 she has been the Editor-in- Chief of the Activity Report for the Institute of Informatics. Her scientific research interest include digital image processing and recognition, with special emphasis on mathematical morphology methods, computational intelligence and especially rough set theory and artificial neural networks, stylometry and its tasks, elements of theory of logic circuits, their design procedures and optimisation of implementations, as well as arithmetic of digital systems.

Krzysztof A. Cyran was born in Cracow, Poland, in 1968. He received MSc degree in computer science (1992) and PhD degree (with honours) in technical sciences with specialty in computer science (2000) from the Silesian University of Technology SUT, Gliwice, Poland. His PhD dissertation addresses the problem of image recognition with the use of computer generated holograms applied as ring-wedge detectors.

He has been an author and co-author of more than 60 technical papers in journals (several of them indexed by Thomson Scientific) and conference proceedings. Dr. Cyran (in 2003-2004) was a Visiting Scholar in Department of Statistics at Rice University in Houston, US. He is currently the Assistant Professor and the Vice-Head of the Institute of Informatics at Silesian University of Technology, Gliwice, Poland. He is also a member of the Editorial Board of Journal of Biological Systems and a member of the Scientific Program Committees of many international conferences. His current research interests are in image recognition and processing, artificial intelligence, digital circuits, decision support systems, rough sets, computational population genetics and bioinformatics.