Identification of the Minimal Set of Attributes That Maximizes the Information towards the Author of a Political Discourse: The Case of the Candidates in the Mexican Presidential Elections

Antonio Neme[1,2], Sergio Hernández[3], and Vicente Carrión[4]

[1] Complex Systems Group, Universidad Autónoma de la Ciudad de México
San Lorenzo 290, México, D.F. México
[2] Institute for Molecular Medicine, Finland
neme@nolineal.org.mx
[3] Postgraduation Program in Complex Systems, Universidad Autónoma de la Ciudad de México
[4] CINVESTAV IDS, México D.F.

Abstract. Authorship attribution has attracted the attention of the natural language processing and machine learning communities in the past few years. Here we are interested in finding a general measure of the style followed in the texts from the three main candidates in the Mexican presidential elections of 2012. We analyzed dozens of texts (discourses) from the three authors. We applied tools from the time series processing field and machine learning community in order to identify the overall attributes that define the writing style of the three authors. Several attributes and time series were extracted from each text. A novel methodology, based in mutual information, was applied on those time series and attributes to explore the relevance of each attribute to linearly separate the texts accordingly to their authorship. We show that less than 20 variables are enough to identify, by means of a linear recognizer, the authorship of a text from within one of the three considered authors.

Keywords: authorship attribution, mutual information, genetic algorithms.

1 Introduction

Authorship attribution (AA) and stylistics have attracted the attention of different practitioners from areas as diverse as computer science, philosophy, the arts, mathematics, and engineering. AA refers to the task of identifying the author of a text from a group of possible candidate authors [1], whereas stylistics refer to the identification of attributes that may lead to unequivocally identify an author [4]. Both tasks have witnessed an outstanding advance in recent years. However, there are still a lot of open questions.

One of the open questions is the identification of the minimum set of attributes that can lead to the identification of the author. Several attributes have been

I. Batyrshin and M. González Mendoza (Eds.): MICAI 2012, Part II, LNAI 7630, pp. 81–90, 2013.

proposed, for example, the use of certain words and the lack of use of other [1]. In general, the concept of *bag of words* is frequently mentioned and, although relevant results have emerged, there are even more questions to be answered [2]. Writers use language following different ways to express their ideas. This variation in language allows the authorship attribution possible [3].

Several algorithms are able to identify the author of a given text, but, however, most of them lack of explanatory properties. For example, some kernel methods present good performance, but the model is unable to show what attributes are really relevant.

In this contribution, we present results regarding the identification of the minimal set of attributes that define a specific feature space. We are interested in such a feature space in which the mutual information between the coordinates of the points (texts) and the label (author of the text) is maximal. We present results regarding three authors, that happen to be the main candidates for the Mexican presidential elections of 2012. We selected the texts from several political discourses of those authors for two main reasons. The first one is that the subject in all texts for the three authors is very similar and is mainly in the economic public services, and taxes themes. This allows an easier isolation of the stylistics as it is not affected by a large variety of themes. The second reason is that of there are several texts available from candidates. Finally, it is relevant to know at least some aspect of stylistics from political leaders.

The rest of this contribution is presented as follows. In section 2 we describe the attributes that will lead to definition of the stylistics of the authors. In section 3 we present our proposal to identify the minimum set of attributes that define a space such that the coordinates of texts in that space give the maximum information about the author (class). In section 4 several results are described, and in section 5 some conclusions are discussed.

2 Attributes and Stylistics

Several attributes have been proposed as marks in order to discern the stylistics of an author [1]. Also, many features have been proposed to be relevant for the AA task, as vocabulary size or the use of certain words or structures [2]. Here we will focus our attention in attributes about the way authors make use of words. We refer to words as the vocabulary but also to punctuation marks.

In this contribution, we consider two kinds of attributes that are extracted from each text. The first one consists of probabilities of appearance of certain words. The second group consists of the mutual information of several time series constructed from texts. Fig. 1 show those attributes and an identification name.

Texts are represented as sequences of symbols, so they are transformed to series of integers. From the vocabulary for each text, each word is assigned an integer in order of appearance. The first word to appear in the text will be assigned to 0, the second non-repeated word 1, and so on. For example, the sentence $S =$ *In the city as well as in each neighborhood...* is transformed to the sequence $T = \{0, 1, 2, 3, 4, 3, 0, 5, 6, ...\}$. The word *in* is assigned to code 0 as it

Attribute	Description	No. var	Attribute	Description	No. var
V	Vocabulary size	1	mdThe	Minimum distance between consecutive appearances of *the*	1
T	Text length in words	1	MdThe	Maximum distance between consecutive appearances of *the*	1
V/T	Ratio V/T	1	pMCWx	probability of the most common word (except articles, prepositions and ",")	1
H	Entropy	1	adMCWx	Average distance between most common word (except articles, prepositions and ",")	1
MPL	Maximum paragraph length (sentences per paragraph)	1	mdMCWx	Minimum distance between most common word (except articles, prepositions and ",")	1
APL	Average paragraph length	1	MdMCWx	Maximum distance between most common word (except articles, prepositions and ",")	1
mPL	Minimum paragraph length	1	PkMCWx	Probability distribution of the 30 most common words (except articles, prepositions and ",")	30
PDPL	Probability distribution of paragraph length (up to 30 sentences per paragraph)	30	pComma	Probability of the comma	1
MSL	Maximum sentence length (words per sentence)	1	adComma	average distance between consecutive appearances of the comma	1
ASL	Average sentence length	1	mdComma	Minimum distance between consecutive appearances of the comma	1
mSL	minimum sentence length	1	MdComma	Maximum distance between consecutive appearances of the comma	1
PDSL	Probability distribution of sentence length (up to 200 words per sentence)	200	MIFS	Mutual information function for time series S (40 displacements)	40
pMFSL	Probability of the most frequent sentence length	1	MIFPL	Mutual information function for time series paragraph length	40
PkMCW	Probability distribution of the 30 most common words	30	MIFSL	Mutual information function for time series sentence length (40 displacements)	40
pMCW	probability of the most common word (except , and *the*)	1	MIFMCW	Mutual information function for time series distance between MCW (40 displacements)	40
adMCW	Average distance between consecutive appearances of most common word	1	MIFMCWx	Mutual information function for time series distance between MCWx (40 displacements)	40
mdMCW	minimum distance between consecutive appearances of most common word	1	MIFComma	Mutual information function for time series distance between comma (40 displacements)	40
MdMCW	maximum distance between consecutive appearances of most common word	1	MIFThe	Mutual information function for time series distance between *the* (40 displacements)	40
pThe	Probabilityof the word *the*	1	MIFBin	Mutual information function for time series B (40 displacements)	40
adThe	Average distance between consecutive appearances of *the*	1			40

Fig. 1. The included attributes. Some are scalars, some are probability distributions, and others are mutual information functions (see next section).

is the first word. The second appearance of *in* is also assigned code 0. In this contribution, there is no difference between upper and lower cases.

Time series from texts are relevant in the vision about stylistics we follow. Several time series were constructed from each text, and they are generated measuring the distance (counting the number of words) between consecutive appearances of certain words. For example, a certain time series that measures the distance (number of words) between consecutive appearances of the comma may reads as $\{3, 9, 40, 11\}$, which means that the number of words between the second and first appearance is 3, the distance between the second and third appearances is 9, and so on. Fig. 2 shows an example of the construction of the time series for the comma. Time series for the following instances were constructed:

- the comma
- sentence length (number of words between them)
- number of sentences per paragraph
- the most common word excluding the comma and the word *the*
- the most common word excluding articles and prepositions
- the word *the*

However, time series *per se* only give some visual details, and more processing on them is necessary.

Texts may present different lengths, so a normalizing scheme is needed in order to compare time series. At the same time, time series are not analyzed directly. In general several tools from the time series and signal processing fields can be applied in order to extract subtle and non-evident patterns [5]. Several attributes can be extracted from time series, such as the power spectrum, the Lyapunov exponent, and many others [6]. In this contribution, we applied mutual information function (MIF).

MIF is an information measure. Once we know the state of a system, How much information does that give about the state of which a second system?

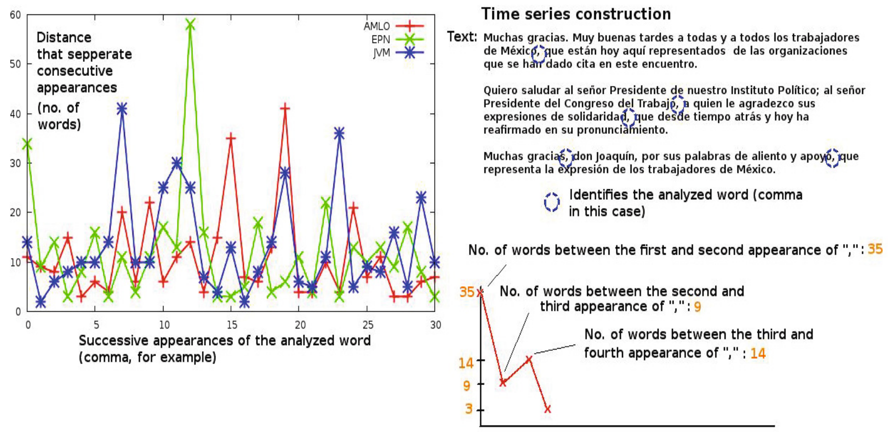

Fig. 2. Time series construction. It is shown the case for distance (number of words) between consecutive appearances of comma. Only the first 31 appearances are shown.

MIF is based in Shannon's information theory. It is based on entropic-related concepts. The entropy H of the training set is defined as:

$$H = -\sum_{i=1}^{\#classes} p_i log(p_i) \tag{1}$$

where the number of classes corresponds to the number of authors and is defined as $\#classes$, and p_i is the probability of randomly chose an input vector whose class is i. The mutual information between two random variables quantifies how much information is gained about the possible state one of them once we know the actual state of the other variable. It is a measure of correlation [8]. Mutual information between two random variables X and Z is expressed as $\Phi(X; Z)$, where X is in this work one of the attributes of the input vectors and Z is the class or label of those vectors. It is defined as:

$$\Phi(X; Z) = \sum_{i}^{ns} \sum_{j}^{\#statesinZ} P(i,j) log \frac{P(i,j)}{P(i)P(j)} \tag{2}$$

The number of states in Z is the number of classes, and ns is the number of states in X. If X is a continuous variable, then it can be discretized into ns different states. $P(i)$ is the marginal probability that a randomly chosen text belongs to a certain state, $P(j)$ is the marginal probability that the text belongs to class (author) j, and $P(i,j)$ is the joint probability that the text is in state i and belongs to author j. In general, for artificial datasets with no noise, all entropy in the label (class) can be removed from the list of attributes $\bar{X}$ that define the high-dimensional feature space. That is, $\Phi(\bar{X}; Z) = H$. Mutual information between the compound system of all attributes or variables and Z ($\Phi(\bar{X}; Z)$) tends to dissipate all entropy in the label. That is, when $ns \to \infty$, $\Phi(\bar{X}; Z) \to H$.

When the correlation between two systems (or random variables) is the mutual information, we are considering high-order momentum able to capture non-linear correlations in data [10,8].

When MIF is applied to a time series, the second system (or random variable) is constructed as a shift applied to the time series. The length of that shift is shown in the x axis. The graph of MIF then responds the question of how much information is achieved once we know the state of a system (the time series) with respect to the next state it will present (k = 1), two steps ahead (k = 2), and so on. Fig. 3

We call the whole set of attributes T

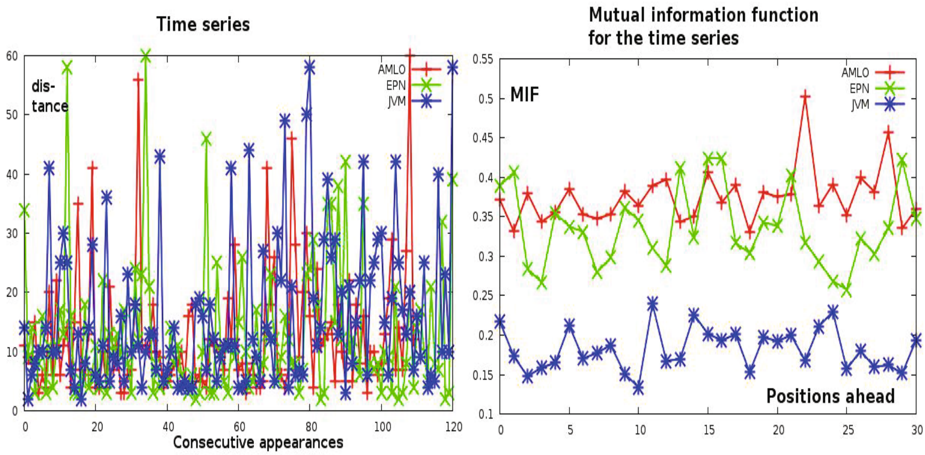

Fig. 3. Mutual information function (MIF) of the distance between commas time series. One time series for each one of the considered authors is shown.

3 The Proposed Model

Each text is transformed to a point in a high-dimensional space. The coordinates of each text are determined by the attributes described in the previous section. Points in this space may be the input data to a classification machine like multilayer perceptrons, identification trees or support vector machines so that a mapping between the coordinates and the label (the author) is found. If the number of attributes, that is, the dimension of the feature space, is high then the procedure followed to find such mapping, known as training process, may be very time consuming. In other cases, as in SVM, the new generated very-high-dimensional space lack explanatory power, that is, it may not be clear what attributes are really relevant in the classification task. On the other hand, classifiers based in trees such as C4.5 [12] offer an explicit explanation about the classification task. C4.5 and related methods, although relevant and useful in many situations, suffer from a major drawback: their greedy strategy may lead them to local optimum.

We are interested in finding a subset $A \in T$ such that $|A| \leq K$ such that the mutual information from A to the classes (author's name) is maximal. Let $\Phi(X,Y)$ be the mutual information between systems X and Y, and Note that this task is not equivalent to that performed by C4.5 related algorithms. We are not interested in classification by means of mutual information. We are trying to find a subspace such that coordinates in that space give as much information about the label or class as possible, and any machine learning algorithm can be fed with vectors in that space A, instead of being fed with vectors from space T whose dimensionality is higher. In that sense, our task is slightly similar to that of testors [13], in which a matrix of differences is systematically explored to identify those features that correctly classify patterns.

Once again, we intend to find an attribute or feature space such that the mutual information between points representing texts in that space and authors is as maximum as possible. In order to do this, the mutual information of a compound system is needed. That is, if there is only one attribute then the MI (mutual information) is calculated straightforward. In the case of two continuous attributes $X^{'}$ and $Y^{'}$ and ns is the number of states in which each attribute is to be discretized (X and Y), a compound system Z is constructed as follows. $Z'_i = X_i \times ns + Y_i$ and $Z = discretize(Z^{'}, ns)$. For more than two attributes, the procedure is applied recursively.

4 Results

The analyzed texts are shown in fig 4, along with the date they were dictated as a public discourse and a short title. All texts were preprocessed to remove transcribed messages from the public and other irrelevant information.

Fig. 5 shows the vocabulary as a function of number of words for the analyzed texts (see fig. 4). It is observed that, although one of the authors present a significant lower vocabulary size, it is still an attribute unable to give a lot of information about the author.

Fig. 6 shows the mutual information between some individual attributes in T and the class (author of the text). The dimension of space T, that is the number of attributes, is $D \sim 500$. That would require a lot of effot to train a multilayer perceptron. For a SVM, that may be an easy task, but we are interested in the identification of attributes that give information about the class. The function that maps from that space to the label space is not explored in this contribution. SVM does not offer such explanation.

The náive scheme to construct the space A from T will be to select the K most informative variables. Such strategy is followed, for example, by C4.5 and related algorithms, but that greedy strategy leads to local optima. In fig. 6 can be see an example of the failure of such strategy. If attributes V and T were rejected and H and $ANPL$ (see fig. 1) were selected, the opportunity that the compound system V,T to be selected would not be explored, and, as can

Author	Title	day	month	year	id	Title	day	month	year	id
EPN	Encuentro con estructuras en Tijuana	3	6	2012	1	Celebración del día de las madres	10	5	2012	22
	En la firma del plan de la concertaciónn mexicana	5	6	2012	2	Encuentro con estructuras San Luis Potosí	9	5	2012	23
	En encuentro con mujeres de Tijuana	3	6	2012	3	Encuentros por el futuro de México	9	5	2012	24
	Encuentro con estructuras de Baja California Sur	2	6	2012	4	En el encuentro del club rotario internacional	4	5	2012	25
	En el foro encuentros con el futuro, Merida	31	5	2012	5	En el acto dia Santa Cruz	3	5	2012	26
	En el décimo foro nacional de turismo	30	5	2012	6	Alianza por un proyecto de país	2	5	2012	27
	En el arranque de campaña de Manuel Velasco, Chiapas	29	5	2012	7	Dialogo con profesores UNAM	2	5	2012	28
	En la reunión nacional consejeros Bancomer	29	5	2012	8	Acto conmemorativo Dia del Trabajo	1	5	2012	29
	En la reunión con el consejo nacional agropecuario	28	5	2012	9	En el inicio de campaña de Beatriz Paredes	29	4	2012	30
	En los diálogos por la Paz	28	5	2012	10	Encuentros futuro de México	28	4	2012	31
	En el evento Proecto de Nación	25	5	2012	11	Encuentro con juventud poblana	27	4	2012	32
	En el 4o. Foro Ncional sobre Justicia y Seguridad	24	5	2012	12	Encuentro con productores del campo	26	4	2012	33
	En el evento por un México incluyente	23	5	2012	13	Encuentro con empresarios Tabasco	25	4	2012	34
	XX sesión de la ANUIES	21	5	2012	14	En el foro futuro para todos	24	4	2012	35
	Encuentro con estructuras de Colima	19	5	2012	15	Encuentro con la sociedad, Monterrey	22	4	2012	36
	En 75 convención bancaria	18	5	2012	16	Encuentro con la sociedad, Aguascalientes	21	4	2012	37
	Encuentro con emprearios de la tecnología e información	17	5	2012	17	Encuentro con empresarios turísticos	18	4	2012	38
	Encuentro con la Sociedad Civil Campeche	16	5	2012	18	Encuentro con la sociedad Tuxpam	14	4	2012	39
	En el arranque de campaña de Jesús Ali, Tabasco	14	5	2012	19	Encuentro con la sociedad Poza Rica	14	4	2012	40
	Encuentro sociedad civil Coahuila	12	5	2012	20	Encuentro industria aeroespacial	12	4	2012	41
	Encuentro en Universidad Iberoamericana	11	5	2012	21					
AMLO	Seré un presidente itinerante	16	5	2012	1	Dignidad del pueblo, Torreón	19	8	2005	14
	Comparen trabajos en reduccion de robo con EPN	8	5	2012	2	Asamblea Zócalo	30	7	2006	15
	Algo e comunicaciones / Universal	15	11	2011	3	Auditorio Nacional	26	4	2006	16
	Presentción estrategia de seguridad	11	4	2012	4	Calderón sin apoyo	31	8	2006	17
	Propuesta en materia energética	9	4	2012	5	Cargo presidente legítimo	16	9	2006	18
	Se destinarán 30mmdp para educacion	3	4	2012	6	Discurso diputados desafuero	7	4	2005	19
	Vamos a gobernar juntos	2	4	2012	7	Discurso energía	28	1	2006	20
	Convoca a redes sociales	31	3	2012	8	Mitin apoyo Oaxaca	31	10	2006	21
	Registro ante IFE	22	3	2012	9	Plan cultura	30	5	2006	22
	Protesta candidaco movimiento ciudadano	11	3	2012	10	Presentación nuevo proyecto de nción	20	3	2011	23
	Audita IFE pero no investiga al PRI	1	6	2012	11	Redes ciudadanas	17	7	2005	24
	Discurso Cananea	27	1	2006	12	Discurso Zócalo previo desafuero	7	4	2005	25
	Cierre campaña	28	6	2006	13					
JVM	Visita Ibero	4	6	2012	1	Discuso panistas Pto. Vallarta	26	11	2011	14
	Tehuacń	3	6	2012	2	10 Obras conmemoración indep y revoluc	8	12	2010	15
	Alcaldes panistas	2	6	2012	3	Discurso presup resp	21	11	2010	16
	Reunión turismo	30	5	2012	4	Discuso triunfo candidata Pan	6	2	2012	17
	Evento Cancun	29	5	2012	5	Texto mex just conf	26	12	2010	18
	Evento Nayarit	28	5	2012	6	Texto presidencia avisora	11	2	2012	19
	Desayuno con mujeres Cd Juárez	27	5	2012	7	Postdebate 10 jnio	10	6	2012	20
	Evnto Culiacán	20	5	2012	8	De ciudadana a ciudadana	8	6	2012	21
	Convención nacional bancaria	18	5	2012	9	Evento Coyoacán	4	6	2012	22
	Reunión empresarios CANACO, Mérida	17	5	2012	10	Reunión consejo nacional agropecuario	23	5	2012	23
	Diálogo Javier Corral	4	8	2011	11	Evento Sahuayo Michoacán	19	5	2012	24
	Congreso value interesting	8	11	2011	12					
	Discurso debate precandidatos PAN	1	1	2012	13					

Fig. 4. The texts considered in this contribution. The authors are the three main candidates for Mexican presidential elections in 2012. 41 texts were selected for author EPN, 25 for author AMLO and 24 for author JVM. The selection criteria was that text should be within a certain range of number of words.

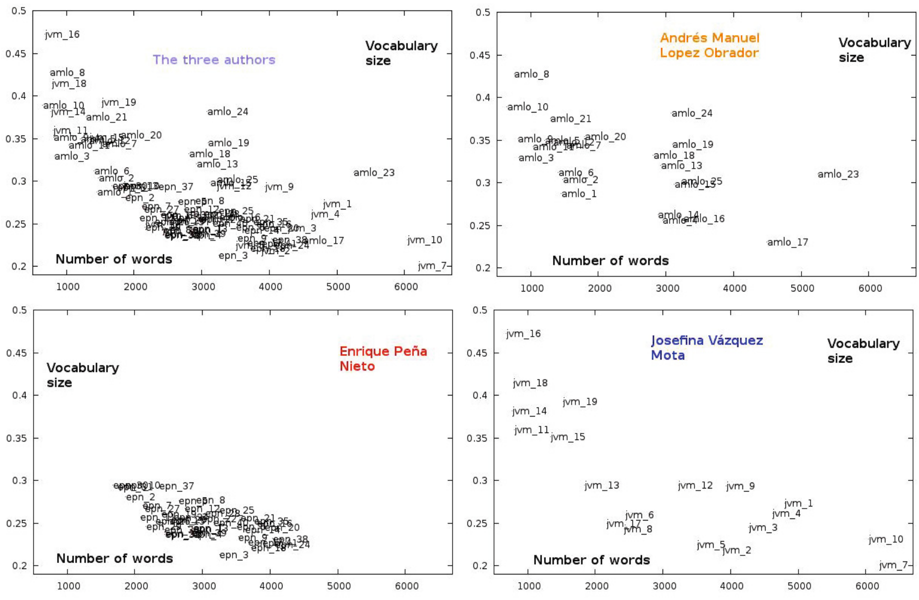

Fig. 5. Vocabulary size as a function of text length (number of words)

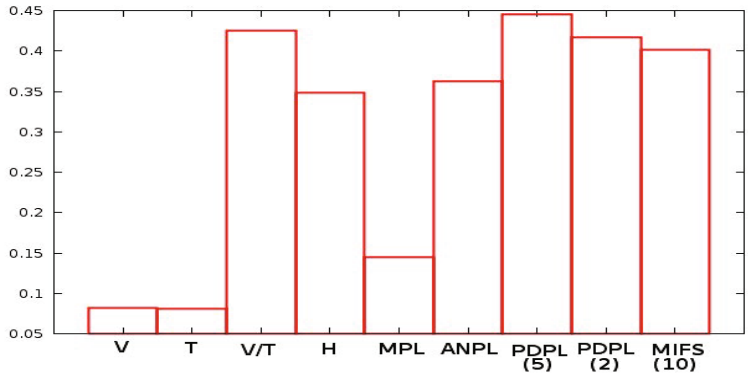

Fig. 6. Mutual information between some of the attributes listed in fig. 1 and the text author (class)

be seen, a compound system (in this case formed as the ratio) was in fact a better option. Thus, we want to explore space T instead of disregarding some of the variables from the beginning. For probability distributions and mutual information functions, the number in the parenthesis represent the component. For example $PDSL(4)$ refers to the probability of sentences of 4 words.

The space generated by K attributes from space T is called A. The number of possible spaces A is the number of permutations of K positions available to D different attributes $C(D, K)$. The exhaustive search for the case here presented is prohibitively time consuming for $K > 3$. Thus, a search scheme is needed. We applied an heuristic search method, the genetic algorithm, in order to find at most K attributes from T that generate a space such that $\Phi(A; Class)$ is maximum.

We implemented a genetic algorithm in Python, with elitism and probabilities of mutation of 0.05 and crossover of 0.9. Population size was settled to 100 and the algorithm was allowed to run for 500 epochs.

Note that the algorithm identifies a space A of dimension $D \leq K$. That space is not easily observed once $D > 3$. In order to visualize the distribution of the analyzed texts in that space, a mapping algorithm is needed. We decanted our options towards the self-organizing map (SOM) as it is a powerful visualization tool. SOM is able to present in a low-dimensional space an approximate distribution that resembles the actual distribution of vectors in the high-dimensional input space [14]. It outperforms common mapping tools such as principal component analysis as SOM is able to account for high-order statistics, instead of at most second-order (variance) [15]. In fig. 7 several maps obtained by SOM are presented. It can be observed that, indeed, there are detectable general distribution patterns that may allow to discriminate the author. Texts do not necessarily form clusters: once again, we are interested in an attribute space such that mutual information between the distribution and the author of a text is maximized. Clusters are only one way in which that mutual information can be maximized, but there are many others. Our methodology finds a family of those distributions.

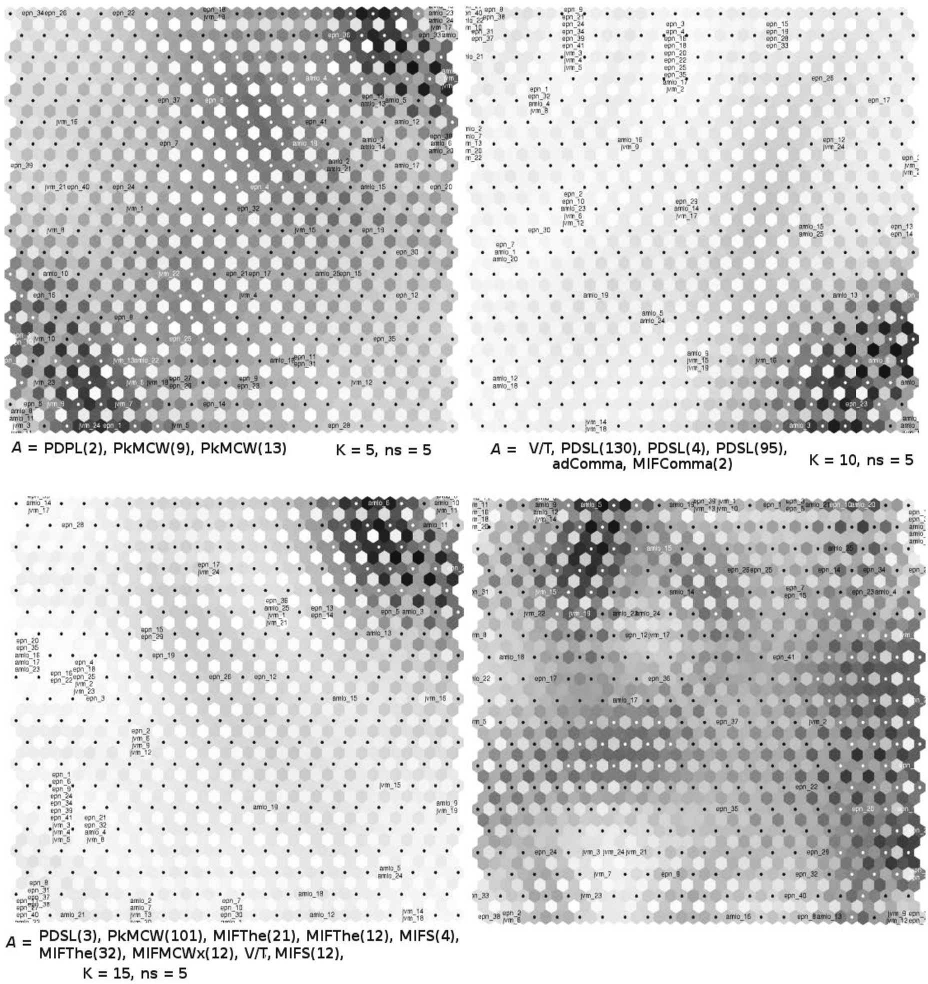

Fig. 7. SOM (low-dimensional approximations) of feature spaces A

In fig. 7, besides the self-organizing map, it is shown the space A. Now, a machine with universal approximation capabilities such as the multilayer perceptron can be applied to space A, instead of being fed by data in space T. The fact that data can be explored in order to identify the combination (subspaces) of attributes that offer the maximum mutual information decreases the training time. Also, it points to the most relevant joint variables, which are in A. It may be possible to add new variables to A as the mutual information will not decrease [11]. However, eliminating attributes from A may have dramatic consequences, as the removed variable may be of great relevance.

5 Conclusions

In the tasks of authorship attribution and computational stylistics, it is of major interest to identify a set of attributes that can offer as much information as possible about the author of the text. Here, we have systematically explored schemes for detecting a subset of a large number of variables that can maximize information about the author. A genetic algorithm that constructs a space of at most K attributes such that it maximized the information about the class or author of the text was implemented.

The methodology here described can be applied to any kinds of texts. Here, we reported results for a special case, regarding political discourses, but in our project (not enunciated for double blind review purposes) we are applying these and other methodologies in order to study computational stylistics and style evolution.

References

1. Juola, P.: Authorship attribution. NOW Press (2008)
2. Stamatatos, E.: A survey of modern authorship attribution methods. J. of the American Soc. for Information Science and Technology 60(3), 538–556 (2010)
3. Neme, A., Cervera, A., Lugo, T.: Authorship attribution as a case of anomaly detection: A neural network model. Int. J of Hybrid Intell. Syst. 8, 225–235 (2011)
4. Manning, C., Schutze, H.: Foundations of statistical natural language processig. MIT Press (2003)
5. Abarbanel, H.: Analysis of observed chaotic data. Springer (1996)
6. Kantz, H., Schreiber, T.: Nonlinear time series analysis, 2nd edn. Cambridge Press
7. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal 27, 379–423, 623–656 (1948)
8. Cellucci, C.J., Albano, A.M., College, B., Rapp, P.E.: Statistical Validation of Mutual Information Calculations. Phy. Rev E. 71(6) (2005), 10.1103/PhysRevE.71.066208
9. Santos, J., Marques de Sá, J., Alexandre, L., Sereno, F.: Optimization of the error entropy minimization algorithm for neural network classification. In: ANNIE V. 14 of Intelligent Engineering Systems Through Art. Neural Net, pp. 81–86. ASME Press, USA (2004)
10. Silva, L., Marques de Sá, J., Alexandre, L.: Neural Network Classification using Shannon's Entropy. In: ESANN 13th European Symp. on Art. Neural Net (2005)
11. Cover, T., Thomas, J.: Elements of information theory, 2nd edn. Wiley (2006)
12. Quinlan, R.: Programs for Machine Learning. Morgan Kaufmann Publishers (1993)
13. Cortes, M.L., Ruiz-Shulcloper, J., Alba-Cabrera, E.: An overview of the evolution of the concept of testor. Pattern Recognition 34, 753–762 (2001)
14. Kohonen, T.: Self-organizing maps, 2nd edn. Springer (2000)
15. The Self-Organizing Maps: Background, Theories, Extensions and Applications. Studies in Computational Intelligence (SCI), vol. 115, pp. 715–762 (2008)