On Sequential Selection of Attributes to be Discretized for Authorship Attribution

Grzegorz Baron
Silesian Univesity of Technology
Institute of Computer Science
44-100 Gliwice, Poland, Akademicka 16/317
Email: grzegorz.baron@polsl.pl

***Abstract*—Different data mining techniques are employed in stylometry domain for performing authorship attribution tasks. Sometimes to improve the decision system the discretization of input data can be applied. In many cases such approach allows to obtain better classification results. On the other hand, there were situations in which discretization decreased overall performance of the system. Therefore, the question arose what would be the result if only some selected attributes were discretized. The paper presents the results of the research performed for forward sequential selection of attributes to be discretized. The influence of such approach on the performance of the decision system, based on Naive Bayes classifier in authorship attribution domain, is presented. Some basic discretization methods and different approaches to discretization of the test datasets are taken into consideration.**

***Index Terms*—discretization, authorship attribution, sequential selection, Naive Bayes.**

I. Introduction

Authorship attribution is one of the tasks of stylometry that deals with authors recognition of anonymous texts, detecting plagiarism, selecting possible texts authors from the given set. Different data mining techniques can be employed, such as artificial neural networks, decision trees, support vector machines, rough set theory, principal component analysis [1].

Discretization is the method allowing to convert continuous data into discrete form. It can be taken into account as an obligatory or optional element of data processing chain in a decision system. The first case takes place when subsequent modules can operate only on nominal data. In the second case the application of discretization for input data can be the subject of consideration in order to obtain some benefits such as the improvement of the decision system by simplification and generalization of input data.

The previous research [2] addressed issues related to the analysis of classification improvement by discretization of the input data. Typically, all attributes of input sets were discretized using the given method and the obtained classification results were investigated. In many cases, discretization of input data led to improvement of the system performance. But there were situations in which poor results were obtained. Deeper investigation of outcomes allowed to state the question if the discretization of selected attributes would be more profitable, and what would be the relation between performance of a decision system and selection of attributes being discretized. Immediately the next question arose, namely what method of attributes selection should be applied. The aforementioned problems constituted the objective of presented research.

978-1-5090-5795-5/17/$31.00

Very often decision systems are built basing on supervised learning techniques [1]. In the presented research test sets approach to system evaluation was used [3]. That generates additional issues related to discretization of test sets. Three approaches were developed during the author's previous research [4], named "Independent", "Test on Learn", and "Glued". The "Independent" approach rely on separated discretization of learning and test sets. That can lead to inconsistency between training and test data, which hampers the subsequent processing of data. The next two ways of test sets discretization deliver consistent data with the same ranges of bins assigned to learning and test datasets. For "Glued" approach both sets are assembled together, discretized, and separated back to obtain learning and test data. For "Test on Learn" method discretization of test datasets is performed using bin ranges obtained for learning data.

The paper presents the results of research on searching the dependency between selection of attributes to be discretized and performance of a decision system. The Naive Bayes classifier was chosen as inducer because of its proven quality and good performance. As attributes selection method the forward sequential selection was picked [5].

The results obtained during the research were very promising. The analysis of experimental outcomes revealed clear and relatively strong dependence between the selection of attributes to be discretized and the performance of the decision system. Some preliminary observations of rules describing the experimentally obtained order of attributes were made.

The paper is organized as follows. Section II presents theoretical overview of the used methods, Sect. III explains the experimental setup, including description of datasets and employed techniques. Discussion of the results is given in Sect. IV whereas conclusions are formulated in Sect. V.

II. Introduction to Naive Bayes Classifier and discretization techniques

The following points give theoretical background of the presented research including Naive Bayes classifier and discretization methods.

A. Naive Bayes Classifier

Naive Bayes classifier is very useful and popular. Very often it is selected as reference model in different application domains [6]–[9]. The main assumption concerning that classifier is the independence of attributes. While it is not true in most real-world tasks, paradoxically the classifier often performs very well. Thanks to that assumption the learning process for each attribute can be conducted separately, which accelerates training for large sets of attributes.

Naive Bayes classifier utilizes Bayes' rule of conditional probability:

$$p(c_j \mid d) = \frac{p(d \mid c_j)p(c_j)}{p(d)}, \tag{1}$$

where: $p(c_j \mid d)$ – a'posteriori probability of instance d being in class c_j, $p(d \mid c_j)$ – probability of generating instance d given class c_j, $p(c_j)$ – a'priori probability of class c_j occurrence, $p(d)$ – probability of instance d occurring.

Value of $p(d \mid c_j)$ is a product of probabilities for all elementary instances d_i:

$$p(d \mid c_j) = p(d_1 \mid c_j)p(d_2 \mid c_j)\dots p(d_m \mid c_j). \tag{2}$$

The MAP (maximum a'posteriori) rule is employed to obtain the result of classification process, as follows:

$$NBC(d_1, \dots, d_n) = \\ = \underset{c}{\operatorname{argmax}}\, p(C = c) \prod_{i=1}^{n} p(D_i = d_i \mid C = c). \tag{3}$$

It is assumed that numeric attributes are normally distributed, so the Gaussian normal distribution can be utilized for calculating the probability values:

$$p(D = d \mid C = c) = \frac{1}{\sqrt{2\pi}\sigma_c} e^{-\frac{(d-\upsilon_c)^2}{2{\sigma_c}^2}}, \tag{4}$$

where υ is the mean of the attribute given the class, and standard deviation is σ. It is the simplest approach but for the specific purposes other distributions may be more suitable [10].

B. Discretization

Discretization is the process of partitioning values of continuous variables into categories. The goal of discretization is to find a set of cut points to divide the range of data into some number of intervals. There are many criteria of discretization methods categorization [11]–[13]. In the presented research, the division into supervised and unsupervised groups is essential. Supervised methods utilize class information during discretization process whereas unsupervised do not. Basic discretization process can be composed of four steps:

1) sorting the continuous range of data to be discretized,
2) evaluating points for splitting (or intervals for merging),
3) applying splitting or merging process established on specified rules,
4) stopping the process after reaching some postulated criteria (especially for iterative, incremental processes).

1) Unsupervised discretization: Basic unsupervised methods are equal width and equal frequency discretization. The equal width algorithm evaluates the minimum and maximum values of the discretized attribute and then divides the range into the previously defined number of equal width discrete intervals. In WEKA package for equal width discretization there is also the option which allows to optimize the number of bins using leave-one-out estimation of estimated entropy. The algorithm performed consists of the following steps:

1) $\forall a \in A$ the distribution table $d(a, b)$ is calculated, where A is the set of attributes, $b = 1 \dots B$ and B is the maximum required number of bins; $d(a, b)$ reflects the number of instances of attribute a placed in each bin given the b,
2) calculate entropy for all attributes $\forall a \in A$:

$$b_{opt}(a) = \underset{b}{\operatorname{argmin}}\, E(a) = \\ = -\sum_{k=1}^{b} d(a, k) \log \frac{d(a, k) - 1}{w(a, k)} \tag{5}$$

   where $w(a, k)$ is width of the bin for given attribute a and number of bins k,
3) compute cut points for $b_{opt}(a)$

The equal frequency algorithm evaluates the minimum and maximum values of the discretized attribute, sorts all values in ascending order, and splits the range into a defined number of intervals so that every interval contains the same number of values. There is also another approach in which the desired number of instances per interval is defined (weight of instances per interval) and the resultant number of bins depends on input data.

2) Supervised discretization: For these methods class information entropy for verification of the quality of candidate cut point is used. Starting from one big interval, the recursive partitioning is performed until a stopping criterion is satisfied. To test it, two algorithms employing the minimum description length (MDL) principle were used in the presented research: Fayyad and Irani's [14] and Kononenko's [15].

a) Fayyad and Irani's MDL: Let us assume that in the set S of N examples there is k classes $C_1, \dots, C_k$. Class entropy of S is defined as:

$$\mathrm{Ent}(S) = -\sum_{i=1}^{k} P(C_i, S) \log(P(C_i, S)) \tag{6}$$

where $P(C_i, S)$ is the proportion of class C_i examples in S.

Considering the binary discretization of continuous variable A, if we want to select an optimal cut point T_A, it is necessary to test all possible cut points T and select one for which the class information entropy $E(A, T_A; S)$ of newly generated partition is minimal. Such entropy for the single cut point T which splits S into two subsets S_1 and S_2 can be calculated as follows:

$$E(A, T; S) = \frac{|S_1|}{|S|}\mathrm{Ent}(S_1) + \frac{|S_2|}{|S|}\mathrm{Ent}(S_2) \tag{7}$$

The process is performed recursively. Fayyad and Irani formulated criterion which allows to evaluate when to stop the discretization process. It is based on Minimum Description Length (MDL) idea.

The algorithm uses the information gain value:

$$Gain(A, T; S) = \text{Ent}(S) - E(A, T; S) \tag{8}$$

or substituting equation 7:

$$Gain(A, T; S) = \text{Ent}(S) - \frac{|S_1|}{N}\text{Ent}(S_1) - \frac{|S_2|}{N}\text{Ent}(S_2) \tag{9}$$

Fayyad and Irani's rule recommends to test the following inequality:

$$Gain(A, T; S) > \frac{\log_2(N-1)}{N} + \frac{\Delta(A, T; S)}{N} \tag{10}$$

where

$$\Delta(A, T; S) = \log_2(3^k - 2) - \\ - [k\text{Ent}(S) - k_1\text{Ent}(S_1) - k_2\text{Ent}(S_2)] \tag{11}$$

Until inequality 10 is satisfied, the recursive discretization process should be continued.

b) Kononenko MDL: Let N denotes the number of training instances, N_{C_x} the number of training instances from class C_x, N_{A_x} the number of instances with x-th value of given attribute and $N_{C_x A_y}$ the number of instances from class C_x with y-th value of given attribute and N_T number of possible cut points. Kononenko proved that tested split can be accepted if inequality

$$\log \binom{N}{N_{C_1}, \ldots, N_{C_k}} + \log \binom{N+k-1}{k-1} > \\ > \sum_j \log \binom{N_{A_j}}{N_{C_1 A_j}, \ldots, N_{C_k A_j}} + \\ + \sum_j \binom{N_{A_j}+k-1}{k-1} + \log N_T \tag{12}$$

is satisfied. If not, the discretization process should be stopped.

C. Test Sets Discretization

Because the application of test sets during classifier evaluation stage is the more reliable approach [3], some issues related to the discretization of test sets must be taken into consideration. The most intuitive way of the discretization of test data is applying the same algorithm and parameters like for learning set. Such approach is called "Independent". But depending on the method, some problems may appear. Especially for supervised methods, where cut-points and bin ranges are determined basing on the nature of data, their values may be calculated in quite a different way for learning and test data. It may lead to such big inconsistency between training and test data that evaluation of classifier may fail.

Therefore, two additional approaches to test sets discretization were developed, namely "Test on Learn" and "Glued" [4]. For both, the result number of bins and cut-points are the same in training and test datasets. For the "Test on Learn" approach, the ranges of bins are calculated for learning data and then the same values are applied for test sets. For the "Glued" method, test and training data are concatenated, the discretization process is performed and finally the discretized set is divided into training and test ones.

The main disadvantage of presented approaches is that the main assumption of total independence between learning and test data is not satisfied because of possible mutual influence of both sets during the discretization.

III. Experimental Setup

The decision system developed for presented research contains the following modules:

1) input – prepares the data (preprocessing, splitting into training and test sets),
2) sequential selector,
   - discretizator – performs discretization iteratively for subsequent attributes (one at a time) using various methods, including different approaches to test sets discretization (only one for each experiment),
   - classifier and evaluator – selects the best result obtained for current iteration,
   - attributes manager – manages the ordinal list of attributes selected in subsequent sequential selection instances.

A. Input Datasets

The bag-of-words approach was utilized during the preparation of input datasets. The corpus of texts collected for that purpose was built basing on the works of selected male and female English authors. The source texts were split in order to obtain almost equal blocks of words. To get class-balanced datasets the number of blocks for each author was similar [16].

Next, the characteristic features sets were built. Because the purpose of presented research was to analyze how selective discretization of attributes can improve the quality of the decision system, the selection of attributes constituting their basic set was not in the main focus of the work. The two-letter function words coming from the list of first hundred most used words in English were chosen. The personal pronouns were excluded. The basic list of attributes is as follows: as, at, by, if, in, no, of, on, or, so, to, up.

Finally, the frequency of occurrence of each selected attribute in each block of text were calculated, and formatted into datasets containing two classes representing two authors of source texts, in respect to gender.

According to the previous research [3], the most suggested method of the decision system evaluation is the usage of test sets, taking the application domain into consideration. Therefore, additionally, datasets were split to obtain training and test sets based on disjunctive works of the given authors.

B. Sequential Selection of Attributes for Discretization

The main part of the presented research addressed the analysis of relation between selective discretization of chosen attributes and the overall performance of the decision system.

For each method and approach to test sets discretization the following process was performed:

Init($lAttr$) ▷ list of all attributes in datsets
$lAttrForDiscr \leftarrow empty$ ▷ list of attributes for discr.
$lEvRes \leftarrow empty$ ▷ list of evaluation results
repeat
  for all $currAttr$ **in //**
    ($lAttr$ **not present in** $lAttrForDiscr$) **do**
    $pDiscr \leftarrow currAttr + lAttrForDiscr$
    Discretize($pDiscr, trainSet, testSet$)
    $lEvRes \leftarrow lEvRes +$ **NBClass**($trainSet, testSet$)
  end for
  $bestAttr \leftarrow$ **FindBestAttrib**($lEvRes$)
  $lAttrForDiscr \leftarrow lAttrForDiscr + bestAttr$
until Size($lAttrForDiscr$) = **Size**($lAttr$)

As aforementioned, the Naive Bayes classifier (**NBClass**) was selected for evaluation task. The outer loop is performed until the list of attributes to be tested is empty, or in other words, all attributes were added to the ranked list $lAttrForDiscr$. The position of attribute in this list represents the iteration when it was chosen during the sequential selection process. Another important issue is that for unsupervised methods, like equal width and equal frequency binning, the selection of "the best" attribute in current iteration was performed basing on the averaging of classifier evaluation outcomes obtained in iterative process, where discretization parameter of required number of bins was ranged from 2 to 10. Such range was determined arbitrary, basing on the observation made in previous research [2] which showed, that only for relatively small values of required number of bins the improvement of classification efficiency was observed.

IV. Results and Discussion

To establish the reference point for further discussion the Naive Bayes classifier was applied for datasets without discretization. Figure 1 presents the results of experiments performed using three approaches to discretization of test datasets, respectively "Independent", "Glued", and "Test on Learn". On each diagram horizontal line labeled "reference", indicating the aforementioned reference value, is placed. Diagrams present the efficiency of the system as percentage of correctly classified instances in test sets. Results are averaged basing on outcomes obtained separately for male and female authors. The X axis represents the subsequent iterations of the sequential selection of attributes to be discretized. It must be noticed that each value represents also cardinality of the subset of attributes being discretized in given iteration, selected during the sequential selection process. For each iteration outcomes for equal width, optimized equal width, equal frequency, Fayyad&Irani's MDL, and Kononenko MDL discretization methods are shown.

The first look at the presented diagrams reveals the strong relation between the number of iteration (number of discretized attributes) and classification accuracy. Let us start the discussion from the right group of bars presented on all diagrams in Fig. 1. They represent situation in which all attributes were discretized. It is easy to notice that performance of all discretization methods was very poor. None method exceeded the reference level. Especially for "Independent" way of discretization of test sets, for all methods, except equal width, the accuracy was placed even below the level of 80%, which was set as the lower range of results for all diagrams. Such observation may lead to the false conclusion that discretization of input data is useless in the investigated decision system. It is surprising remark in reference to the previous research where positive influence of discretization on classifier's efficiency was shown. In fact, this preliminary conclusion constituted the motivation to investigate deeper relationship between discretization of selected attributes and performance of the inducer.

The most interesting part of the diagrams starts even from the first iteration. All experiments delivered results above the reference level. It is especially noticeable for supervised Fayyad&Irani's and Kononenko MDL methods. The equal frequency method is less effective but, for "Test on Learn" approach to discretization of test sets, performs comparable to the supervised methods. Both equal width methods delivered the poorest results in the experiments. Taking the number of iteration into consideration their characteristic is rather flat, and efficiency decreases quicker than for other methods, reaching the level below the reference one.

The interesting observation is that the best overall results were obtained averagely for 4 attributes being discretized, and for bigger values the efficiency gradually decreases, or oscillates slightly up to 7 attributes selected for discretization in case of "Glued" approach (Fig. 1b).

An additional comment must be made for "Independent" way of test sets discretization. It is the approach where learning and test datasets are discretized separately, therefore outcome sets may contain different ranges of bins or even numbers of bins. That can lead to serious inconsistency between training and test data that is why evaluation of classifiers is reported so badly. It is clearly visible in Fig. 1a where along with the increasing of the number of attributes being discretized, the efficiency diminishes. On the other hand, it was for "Independent" approach that the best overall results were obtained for supervised discretization methods for 3rd and 4th iteration of sequential selection process.

The above discussion omits the information about the order, and the selection of attributes taken for discretization in subsequent iterations. Tables I–III contain the lists of attributes selected in given iteration for each combination of method and approaches to test sets discretization. Additionally, results are presented separately for male and female authors, because there is no method to average such kind of information. The headers of rows start with prefixes "M_" and "F_" which denotes attributes lists for male and female authors respectively. Their further parts indicate the discretization methods employed, namely: "EWd"– equal width, "EWOd"– optimized equal width, "EFd"– equal frequency, "FId"– Fayyad&Irani's MDL, and "Kd"– Kononenko MDL.

The first observations allow to notice that the presented lists

differ, which might be expected intuitively. It is difficult to select or develop a formal method of measuring the similarity of lists of attributes for different discretization methods. Rather personal observations can be involved to find some relations between attributes.

The conclusion which does not strictly defines the behavior of attributes but formulates informally some observations is that it is possible to point single attribute or their subsets which tend to place in the beginning, the middle, and the end parts of the lists. For example, attributes *in*, *if* were selected mainly during the first or second iteration whereas *on* very often was placed at the end of the list. On the other hand, many outliers can be pointed, which does not support the aforementioned observation. Of course, different patterns of list orders can be identified for male and female data.

The question has arisen if there is any rule describing ordering of the attributes. To investigate this issue the ranking of attributes was prepared for male and female datasets. The *CorrelationAttributeEval* from WEKA was used for that purpose, and Table IV presents the obtained results. Column 1 contains most important attributes for male and female authors respectively, whereas column 12 the less important ones.

The most surprising fact is that some attributes selected during early stages of sequential selection appear at the second half of the ranked list, namely they are less important. It is interesting that discretization of less relevant attributes allows to improve the quality od the decision system. This conclusion is based mainly on investigating the results for supervised discretization methods which behave more regularly than equal width and optimized equal width. Of course, it is not an unequivocal rule and many outliers can be found. Also, differences between datasets for male and female authors can influence the analysis. Generally speaking, female datasets deliver more predictable and mostly better results in all experiments, comparing to the male data. Therefore, for female sets it is easier to observe some relations between relevance of attributes and order determined by sequential selection process. But, unfortunately, formulation of strict rules defining relation between position of attribute in ranking, and its influence in discretized form on quality of the system is impossible. Further research is necessary to identify and describe such relations.

V. Conclusions

The paper presents the research on influence of selective selection of attributes to be discretized on accuracy of the decision system for authorship attribution. The sequential selection method was chosen to perform iterative process of searching among attributes the best candidates to discretization. The Naive Bayes classifier was used for temporary evaluation of the outcomes during the selection process, and final assessment of the decision system performance.

Experimental results showed strong relation between the number of attributes being discretized and the quality of the decision system. To be more specific, starting from the small number of discretized attributes the results exceed the reference level obtained for the same data but without discretization. For selected discretization methods quality measure rose up to 4 attributes being discretized, reaching in the best cases quality about 7% better than reference. For greater number of discretized attributes results were decreasing gradually. In case of discretization of all attributes, the quality of the decision system fell below the reference level. It is the case of the specific datasets and attributes selected for the research because, generally speaking, such approach very often delivered results better than reference [2], [4].

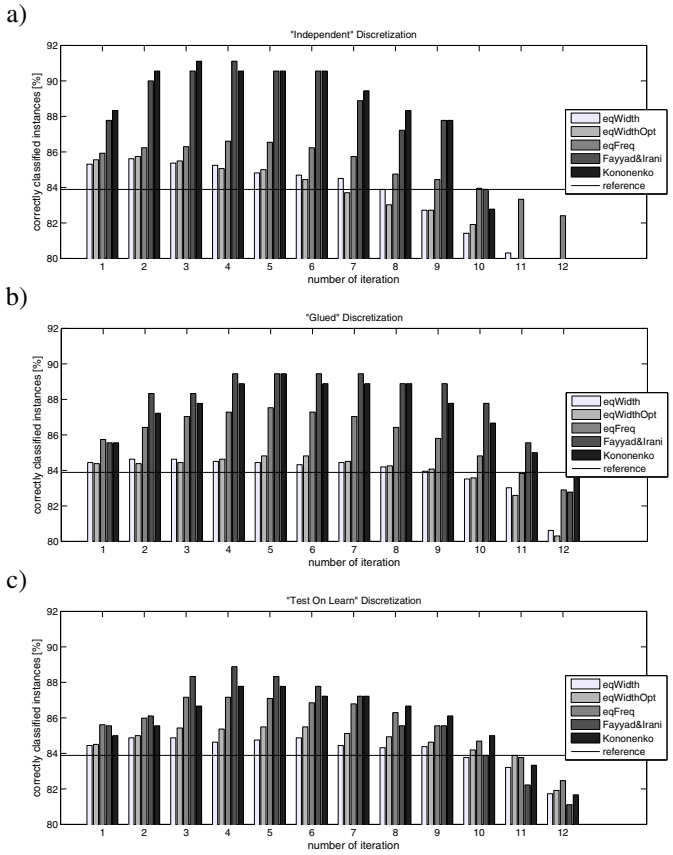

Fig. 1. The results of Naive Bayes classifier evaluation for subsequent selections of attributes to be discretized for "Independent" a), "Glued" b), and "Test on Learn" c) approaches to discretization of test sets.

Attributes, being the best candidates for subsequent iterations, were determined experimentally during the sequential selection process. The issue analyzed during the research was the attempt to find some rules describing the best order of attributes. Some dependencies with ranked lists of attributes obtained by applying chosen ranking function were observed. The worse attributes in ranking were selected in early iterations during the sequential selection process, improving the accuracy of classification. It can be a clue for preliminary selection of attributes for partial discretization. But also many outliers exist, which does not allow to formulate unequivocal rule based on the mentioned observations. Additional research is necessary to investigate this problem deeply.

Summarizing the results, the application of sequential selection for defining subset of attributes to be discretized is

TABLE I
ORDER OF ATTRIBUTES SELECTED FOR DISCRETIZATION FOR "INDEPENDENT" APPROACH TO DISCRETIZATION OF TEST SETS

	The number of iteration											
Discr.	1	2	3	4	5	6	7	8	9	10	11	12
F_EWd	in	so	or	at	up	on	if	as	to	no	of	by
F_EWOd	in	so	at	no	of	if	up	by	as	on	or	to
F_EFd	in	no	of	on	so	by	to	or	up	at	as	if
F_FId	in	no	if	at	to	or	up	by	as	so	of	on
F_Kd	in	no	as	by	or	if	at	up	so	on	of	to
M_EWd	up	in	to	as	no	so	at	of	or	by	on	if
M_EWOd	if	as	to	no	up	in	or	at	so	by	on	of
M_EFd	in	to	as	up	no	at	on	of	by	or	if	so
M_FId	of	in	if	to	up	no	at	by	as	or	so	on
M_Kd	by	as	of	up	to	in	at	or	no	if	on	so

TABLE II
ORDER OF ATTRIBUTES SELECTED FOR DISCRETIZATION FOR "TESTONLEARN" APPROACH TO DISCRETIZATION OF TEST SETS

	The number of iteration											
Discr.	1	2	3	4	5	6	7	8	9	10	11	12
F_EWd	if	or	up	at	in	of	by	so	to	no	as	on
F_EWOd	in	no	of	at	as	so	or	by	up	if	to	on
F_EFd	in	no	of	at	to	or	as	so	by	up	on	if
F_FId	in	of	on	if	at	by	or	to	up	so	no	as
F_Kd	in	of	on	if	at	by	or	to	as	up	so	no
M_EWd	by	if	no	up	at	as	to	so	or	of	in	on
M_EWOd	in	if	or	to	up	no	as	at	by	so	of	on
M_EFd	in	if	as	to	up	no	at	of	or	on	by	so
M_FId	if	as	in	to	on	of	up	or	so	at	no	by
M_Kd	of	in	if	as	on	so	at	up	no	to	by	or

TABLE III
ORDER OF ATTRIBUTES SELECTED FOR DISCRETIZATION FOR "GLUED" APPROACH TO DISCRETIZATION OF TEST SETS

	The number of iteration											
Discr.	1	2	3	4	5	6	7	8	9	10	11	12
F_EWd	if	at	so	up	by	as	no	in	of	or	to	on
F_EWOd	if	at	no	in	of	by	to	so	up	as	or	on
F_EFd	in	no	of	to	as	or	at	so	by	up	if	on
F_FId	in	on	of	as	at	by	up	if	so	or	to	no
F_Kd	in	if	on	of	or	at	by	up	to	as	so	no
M_EWd	by	if	as	so	or	to	up	no	at	in	of	on
M_EWOd	in	if	or	up	no	to	at	by	as	of	so	on
M_EFd	if	in	to	as	at	no	up	of	or	so	by	on
M_FId	if	in	as	of	or	by	to	no	so	at	up	on
M_Kd	if	in	as	so	of	to	on	at	or	no	up	by

advantageous. The recommended way for searching the best solution is employing the sequential selection process, which delivers ordered list of attributes as well as information about number of attributes to be discretized in order to obtain the best decision system.

ACKNOWLEDGMENT

The research described was performed using WEKA workbench [7] at the Silesian University of Technology, Gliwice, Poland, in the framework of the project BK/RAu2/2017.

TABLE IV
RANKING OF ATTRIBUTES

Order of attributes (according to WEKA's *CorrelationAttributeEval*)												
	1	2	3	4	5	6	7	8	9	10	11	12
Female	on	to	of	as	no	by	if	or	in	so	up	at
Male	by	on	if	in	so	as	to	of	no	or	up	at

REFERENCES

[1] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. Amsterdam, The Netherlands: IOS Press, 2007, pp. 3–24.

[2] G. Baron, "Influence of data discretization on efficiency of Bayesian Classifier for authorship attribution," *Procedia Computer Science*, vol. 35, no. 0, pp. 1112 – 1121, 2014.

[3] ——, "Comparison of cross-validation and test sets approaches to evaluation of classifiers in authorship attribution domain," in *Computer and Information Sciences: 31st International Symposium, ISCIS 2016, Kraków, Poland, October 27–28, 2016, Proceedings*, T. Czachórski, E. Gelenbe, K. Grochla, and R. Lent, Eds. Cham: Springer International Publishing, 2016, pp. 81–89.

[4] G. Baron and K. Harężlak, "On approaches to discretization of datasets used for evaluation of decision systems," in *Intelligent Decision Technologies 2016: Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016) – Part II*, I. Czarnowski, M. A. Caballero, J. R. Howlett, and C. L. Jain, Eds. Cham: Springer International Publishing, 2016, pp. 149–159.

[5] U. Stańczyk, "Weighting of features by sequential selection," in *Feature Selection for Data and Pattern Recognition*, U. Stańczyk and L. C. Jain, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 71–90.

[6] H. Zhang, "The Optimality of Naive Bayes." in *FLAIRS Conference*, V. Barr and Z. Markov, Eds. AAAI Press, 2004.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[8] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in *AAAI-98 Workshop On Learning For Text Categorization*. AAAI Press, 1998, pp. 41–48.

[9] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2, pp. 103–130, 1997.

[10] G. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 1995, pp. 338–345.

[11] S. Kotsiantis and D. Kanellopoulos, "Discretization techniques: A recent survey," *International Transactions on Computer Science and Engineering*, vol. 1, no. 32, pp. 47–58, 2006.

[12] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Machine Learning: Proceedings of the 12th International Conference*. Morgan Kaufmann, 1995, pp. 194–202.

[13] R. Dash, R. L. Paramguru, and R. Dash, "Comparative analysis of supervised and unsupervised discretization techniques," *International Journal of Advances in Science and Technology*, vol. 2, no. 3, pp. 29–37, 2011.

[14] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuousvalued attributes for classification learning," in *13th International Joint Conference on Articial Intelligence*, vol. 2. Morgan Kaufmann Publishers, 1993, pp. 1022–1027.

[15] I. Kononenko, "On biases in estimating multi-valued attributes," in *14th International Joint Conference on Articial Intelligence*, 1995, pp. 1034–1040.

[16] U. Stańczyk, "The class imbalance problem in construction of training datasets for authorship attribution," in *Man–Machine Interactions 4: 4th International Conference on Man–Machine Interactions, ICMMI 2015 Kocierz Pass, Poland, October 6–9, 2015*, A. Gruca, A. Brachman, S. Kozielski, and T. Czachórski, Eds. Cham: Springer International Publishing, 2016, pp. 535–547.