Supervised and Unsupervised Discretisation in Stylometric Domain: A Case Study

Urszula Stańczyk[1], Beata Zielosko[2], and Grzegorz Baron[1]

[1] Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
[2] Institute of Computer Science, University of Silesia in Katowice, Będzińska 39, 41-200 Sosnowiec, Poland
urszula.stanczyk@polsl.pl, beata.zielosko@us.edu.pl, grzegorz.baron@polsl.pl

Abstract. Stylometric characteristic features enable description and recognition of authorship by referring to a style of writing. As they are of numeric type, to explore and mine data either techniques capable of (or adjusted to) working for this type of variables are needed, or some discretisation procedure is considered as a necessary part of pre-processing. The paper presents results of investigations on selected supervised and unsupervised discretisation approaches applied in the domain of stylometry, with a task of authorship attribution treated as classification performed by chosen inducers.

Keywords: Supervised Discretisation, Unsupervised Discretisation, Attribute, Interval, Stylometry, Classification

1 Introduction

Discretisation is one of the data reduction methods where numerical attributes are converted into discrete or nominal ones with a finite number of intervals [8]. A mapping from the wide spectrum of continuous values into a subset of discrete values constitutes the reduction process. When properly applied, such approach can filter out the information noise. In many cases a learning process is more efficient and effective for discrete data, and this type of data is easier to store, explain, and understand. On the other hand, discretisation always causes some loss of information, therefore must be applied carefully and this loss minimised.

In many domains available characteristic features, describing objects and concepts, are of numerical nature. In case of data mining with such variables, we need to decide whether to confine considerations only to the techniques and methodologies that are capable of using and operating on continuous valued attributes without their transformation, to modify algorithms to allow for numerical elements, or to perform discretisation, accepting the resulting loss of information and its consequences [12].

When discretisation is to be executed, still the question remains which approach to this process should be employed, as there many methods, divided into categories depending on elements they focus upon. One of possible distinctions is of supervised versus unsupervised procedures. In supervised discretisation information about class labels is taken into account during the search for good cut-points between ranges of attribute values. Such approaches are often considered as superior to unsupervised ones, where class information is disregarded, and the numbers of constructed intervals are specified as an input parameter. In particular equal width binning is frequently criticised, not always deservedly.

The paper presents research results for two discretisation methods: supervised Fayyad and Irani's [9], which refers to entropy while establishing boundaries between intervals; and unsupervised that divides the whole range of attribute values into arbitrarily set numbers of bins with equal width. The two approaches were employed for two tasks of binary authorship attribution [13]. Stylometric analysis of texts makes use of quantitative descriptors of writing styles [1], thus the available features are naturally continuous. To solve these kind of problems typically there are employed either statistics-oriented calculations or machine learning techniques. In the research executed for data discretised with various methods there were investigated results of three well known classifiers, often used in comparisons, Naive Bayes, J4.8, and k-Nearest Neighbours.

The paper is organised as follows. Sec. 2 describes characteristics of stylometric data, Sec. 3 gives background information on discretisation approaches and their categories. Sec. 4 provides some details of the experiments executed, and obtained results are commented in Sec. 5. Conclusions are listed in Sec. 6.

2 Nature of Data in Stylometric Domain

Upon its invention, the stylometric notion that something so subtle as a writing style can be defined by quantitative (as opposed to qualitative) descriptors was revolutionary [7]. The fundamental concept of authorial invariant, a group of numerical features that allow to define, describe, and recognise authorship with a sufficient level of reliability, leads to treating the task of authorship attribution as classification [11]. As a result, data mining approaches can be used to solve such task, either basing on statistics or machine learning.

To recognise an author basing on their writing style requires detecting such traits or habits that are independent on a specific subject, which means that it is not enough to look for certain keywords or key phrases as in typical matching searches. Instead, available text samples of known authorship need to be mined in order to discover qualities differentiating these examples of style from those of other authors. What is more, to avoid the possibility of being deceived by imitated characteristic phrases, the more reliable way lies with referring to linguistic elements used rather subconsciously, such as function words, and patterns of sentence formulation. These considerations give numerical attributes referring to lexical and syntactic markers. The former specify frequencies of occurrence of such elements as certain words or phrases, while the latter reflect the structure of sentences indicated by employed punctuation marks [6].

Described stylometric characteristic features are of numerical nature so they require techniques capable of efficient processing such values, or discretisation needs to be employed before data can be mined. The choice of a particular approach is not straightforward, as no method can be given a guarantee of returning the best classification results for obtained nominal input sets for all cases.

3 Discretisation Approaches

Formally discretisation is a process which transforms values of a continuous attribute a into m discrete intervals $D = \{[d_0, d_1], \ldots, (d_{m-1}, d_m]\}$. d_0 and d_m are respectively the minimum and maximum values, and for $i = 0, \ldots, m-1$, $d_i < d_{i+1}$. $P = \{d_1, \ldots, d_{m-1}\}$ is the set of split or cut-points of a attribute [10].

Most of discretising algorithms are static, which means that discretisation is performed prior to the learning process, as opposed to dynamic approaches where the discretisation model is built by a learner and it is based on the information exchange between discretiser and learner units. When discretisation is performed separately for each attribute, it is called univariate. If the discretiser takes into consideration values of all attributes, and examines relations and interactions between them, the discretisation is multivariate. Unsupervised procedures omit class information. Supervised discretisation algorithms base on various heuristic measures and analyse the relations between attributes and classes.

Equal width binning is one of basic unsupervised discretisation methods. It determines the minimum and maximum values of a given attribute and then divides the whole range into k intervals of equal size.

Fayyad and Irani's [9] supervised discretisation is based on the notion of entropy. Assuming existence of k classes $C_1, \ldots, C_k$ in the set S of N instances, class entropy of S is defined as:

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

where $P(C_i, S)$ denotes the proportion of class C_i instances in S.

If for all instances an attribute assumes different values, to find the optimal cut-point T_A the exhaustive search throughout all possible $N-1$ cut-points T is necessary. If a cut-point T splits S set into two subsets, S_1 and S_2, where $S1 \subset S$ contains instances with attribute values $\leq T$ and $S_2 = S \setminus S_1$, then the entropy for this cut-point can be calculated as follows:

$$Ent(A, T; S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2). \tag{2}$$

For the optimal T_A class information entropy $E(A, T_A; S)$ is minimal. The process of inserting cut-points should be performed recursively for subsequent intervals obtained during the antecedent partitioning. Fayyad and Irani formulated the stopping criterion for this process based on Minimum Description Length

(MDL) principle. According to this rule, the recursive discretisation should be performed as long as the following inequality is satisfied:

$$Gain(A,T;S) = Ent(S) - E(A,T;S) > \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N}, \quad (3)$$

where for k_i being the number of class labels contained in S_i

$$\Delta(A,T;S) = \log_2(3^k - 2) - [k \cdot Ent(S) - k_1 \cdot Ent(S_1) - k_2 \cdot Ent(S_2)] . \quad (4)$$

Supervised discretisation algorithms are considered as the best approach, which theoretically should deliver an optimal set of intervals and cut-points for each attribute. Unsupervised procedures, in particular equal width, which takes into account only widths of constructed intervals while disregarding all other important factors, tend to have rather a bad opinion. However, uneven partition of input space and cases of single interval attributes result in such loss of information that suggest and motivate more thorough study on applications of both, and of their combinations worth investigating.

4 Experiments

In the first step of experiments the input sets were constructed, and the performance of selected classifiers in continuous domain for attributes checked to be used as a reference point. Then the input sets were discretised with supervised Fayyad & Irani's method and with unsupervised equal width binning for varying numbers of intervals. Next, there was executed discretisation combining supervised with unsupervised strategies. To all attributes found as single interval variables in Fayyad's processing of learning sets, additionally unsupervised discretisation was applied for the range of bin numbers. For all discrete sets the classification power of considered inducers was tested and results compared.

4.1 Input Datasets

When authorship attribution is treated as a classification task, recognised authors correspond to distinguished classes, and characteristic features to stylometric descriptors. In the considered case the classification was binary and classes represented in the same degree to prevent problems resulting from imbalanced data [16]. Two pairs of authors were chosen for comparison, female and male, as writing styles are strongly influenced by genders of their authors [15].

For female writer dataset E. Wharton and M. Johnston were selected, and for male writers J.F. Curwood and J. London. Their novels were grouped into three categories, to give base for separate learning and two test sets of samples. Long texts were divided into several smaller parts of comparable lengths. For all these text samples (100 per author in learning, and 45 per author in test sets) next there were calculated frequencies of occurrence for function words most popularly used in English language, and punctuation marks, together giving the preliminary set of a hundred features. Next, to these sets there were applied several attribute ranking algorithms [18], and from resulting rankings there was selected an intersection of these variables that were assigned non-zero ranks. Such processing gave 24 characteristic features for further considerations.

Evaluation of classifiers performance was executed by test sets instead of cross-validation, as the latter approach tends to give falsely higher classification accuracy [2], because of some similarities within groups of examples, due to the construction of the input sets as described above. As a consequence, the separate two test sets were discretised independently on training samples [3]. In case of discretised sets used together with cross-validation the results are also often overoptimistic as compared to evaluation by independent sets.

4.2 Discretisation Applied

In the first part of experiments all input sets were discretised with supervised Fayyad and Irani's method. Due to specifics of this approach, the continuous space is partitioned in varying degrees, with several intervals for some attributes, while for others single bins are found. Single interval variables in discrete domain can be treated as bringing zero information about described concept [17]. However, these characteristics are valid for a considered set whereas in the context of any other set for the same attributes there can be established different numbers of bins. The numbers of intervals throughout all sets ranged from 1 to 4.

In the second part of experiments the input sets were discretised with unsupervised equal width binning approach, with varying the number of bins from 2 to 10. With this processing all attributes were assigned the same number of intervals, regardless of their individual characteristics. This uniform division is one of the reasons why this method is often criticised.

And in the third stage of experiments, for all attributes which in supervised discretisation of learning sets received single intervals (referred to in the rest of the paper as SDL1B) there was performed additional processing of equal width binning, for the numbers of bins from 2 to 10.

With two independent discretisation approaches of learning and test sets several discrete versions of each set were obtained. For each inducer used and each dataset all combinations of train and test sets were investigated, amounting to the total number of tests dependent on the number N of different values of bin numbers tried in equal width binning, and equal to

$1(\text{supervised})+N(\text{unsupervised})+((N+1)^2-1)(\text{combinations})= N^2+3N+1$, per inducer per train dataset, which was doubled because of two test sets.

4.3 Inducers Employed

In the described research three inducers with default parameters were used, Naive Bayes, J4.8, and k-Nearest Neighbours [5], all implemented in WEKA workbench, and capable of operating on numerical as well as nominal attributes.

Naive Bayes is a relatively simple yet powerful classifier, often outperforming more complex inducers. For this very reasons it is widely used as as reference model for comparisons of research results [4]. It works fast and has low storage requirements for processing. This classifier refers to Bayes' rule of conditional entropy while assuming independence of all considered characteristic features.

J4.8 algorithm is an implementation of Release 8 for C4.5 decision tree learner [14], also assuming independence of attributes. This learning scheme bases on divide and conquer approach for construction of unpruned and pruned decision trees. In this version of the classic algorithm when splits for numerical attributes are considered, an adjustment to the information gain, based on MDL principle, is used with the goal of avoiding overfitting.

k-Nearest Neighbours is a popular instant-based method, which can be tuned by the input parameter of the number of considered neighbours [11]. Each considered sample is compared against its chosen number of nearest neighbours and assigned a class that constitutes the majority amongst them.

To prevent overly optimistic classification results, which may happen in case of cross-validation, the performance of all inducers was estimated basing on independent two test sets. The issue was complicated even more, as test sets were discretised entirely independently on training samples.

5 Obtained Results

In the preliminary step of tests, the performance of inducers for the input datasets was estimated with the test sets for continuous domain, to be used as a reference point for further comparisons. These results are given in Table 1. Columns AvgT in all tables show the averages from both test sets for a classifier, whereas columns AvgD list the average classification accuracies for a dataset.

Table 1. Classification accuracy for numerical stylometric attributes [%]

	Classifiers and Test Sets									
	Naive Bayes			J4.8			kNN			
Dataset	AvgT	T1	T2	AvgT	T1	T2	AvgT	T1	T2	AvgD
Female	93.33	95.56	91.11	89.45	86.67	92.22	95.56	97.78	93.33	92.78
Male	91.11	88.89	93.33	83.89	84.44	83.33	94.44	93.33	95.56	89.81

Next, the power of inducers was tested for datasets discretised by supervised Fayyad and Irani's approach, as displayed in Table 2. For female writer dataset for Naive Bayes and kNN classifiers the obtained results were worse than in the continuous domain, while for J4.8 they were roughly at the same level. For male writers for Naive Bayes some improvement in recognition was observed, for J4.8 a very significant decrease and unacceptably low performance, and for kNN slightly lowered accuracy. On average the results for both datasets were worse when compared to numerical attributes. It can be explained to some extent by independent discretisation of sets, which resulted in different numbers of intervals found for the same attributes in learning and test sets.

Table 2. Classification accuracy for stylometric attributes discretised by supervised Fayyad and Irani's method [%]

	Classifiers and Test Sets									
	Naive Bayes			J4.8			kNN			
Dataset	AvgT	T1	T2	AvgT	T1	T2	AvgT	T1	T2	AvgD
Female	85.56	90.00	81.11	89.45	90.00	88.89	92.78	95.56	90.00	89.26
Male	93.68	94.44	93.33	50.00	50.00	50.00	93.33	94.44	92.22	79.07

Unsupervised equal width binning was executed with the number of constructed intervals from 2 up to 10 for all features. The results given in Table 3 indicate that for all classifiers and both datasets in many cases the classification was at higher levels than in case of supervised discretisation. It was due to the fact that the inflexible numbers of bins applied to both learning and test sets caused that characteristics of sets became closer to each other than in case of supervised discretisation, which in turn was better adapted to individual sets.

Table 3. Classification accuracy for stylometric attributes discretised by unsupervised equal width binning [%]

Number	Classifiers and Test Sets									
of	Naive Bayes			J4.8			kNN			
bins	AvgT	T1	T2	AvgT	T1	T2	AvgT	T1	T2	AvgD
	Female writer dataset									
2	88.34	87.78	88.89	88.89	85.56	92.22	87.22	87.78	86.67	88.15
3	92.78	96.67	88.89	95.00	94.44	95.56	88.89	93.33	84.44	92.22
4	93.33	92.22	94.44	90.00	88.89	91.11	95.00	96.67	93.33	92.78
5	94.45	96.67	92.22	91.67	87.78	95.56	93.33	93.33	93.33	93.15
6	92.22	96.67	87.78	90.00	88.89	91.11	93.33	96.67	90.00	91.85
7	94.44	94.44	94.44	90.00	87.78	92.22	96.11	96.67	95.56	93.52
8	95.56	94.44	96.67	91.67	90.00	93.33	95.56	97.78	93.33	94.26
9	92.78	94.44	91.11	92.78	92.22	93.33	95.00	97.78	92.22	93.52
10	91.67	94.44	88.89	90.56	88.89	92.22	95.56	96.67	94.44	92.59
	Male writer dataset									
2	88.33	92.22	84.44	83.33	86.67	80.00	84.45	90.00	78.89	85.37
3	83.33	83.33	83.33	82.78	86.67	78.89	89.45	90.00	88.89	85.19
4	86.11	88.89	83.33	83.33	86.67	80.00	88.33	91.11	85.56	85.93
5	86.11	87.78	84.44	90.56	93.33	87.78	92.78	94.44	91.11	89.81
6	84.44	86.67	82.22	86.11	87.78	84.44	91.11	93.33	88.89	87.22
7	85.56	86.67	84.44	86.67	86.67	86.67	91.11	90.00	92.22	87.78
8	86.67	85.56	87.78	82.78	88.89	76.67	90.56	94.44	86.67	86.67
9	88.89	90.00	87.78	86.67	87.78	85.56	91.11	92.22	90.00	88.89
10	87.78	87.78	87.78	85.00	85.56	84.44	91.67	92.22	91.11	88.15

Considered limitations for the numbers of intervals tested were based on these numbers established through supervised discretisation, where they varied from

1 to maximum of 4. In case of unsupervised discretisation the final decision with respect to this parameter is not straightforward, as depending on the dataset and inducer we can observe several local maxima in classification accuracy, but for more than 10 bins there was detected a downward trend in performance.

In the third part of experiments to the input datasets there was applied a combination of supervised with unsupervised discretisation in the following manner. From the considered set of available features there were selected those, for which in supervised discretisation of learning sets there were found only single intervals, denoted as SDL1B. For these attributes in all sets, that is both learning and test sets, additionally there was executed unsupervised discretisation with equal width approach, with changing the number of required bins from 2 to 10. Within research all combinations of different versions of pairs of learning and test sets were investigated as shown in Fig. 1 for female writer dataset and in Fig. 2 for male writer dataset, which amounted to 10 by 10=100 tests per inducer and a dataset, doubled because of two test sets.

X axis indicates the numbers of intervals for additionally processed variables in the considered learning sets, and the number of 1 here reflects the automatic supervised processing for such sets, that is resulting in single intervals for these attributes. The series correspond to the numbers of bins in the testing sets for the same attributes, with 1 also denoting purely supervised processing for all variables. Thus a combination of 1 with 1 gives the previously listed classification results for supervised discretisation. In the figures the averaged results for both test sets used were depicted.

For female writer dataset and J4.8 classifier the same repeated pattern of recognition level can be observed for all considered numbers of intervals for single bin variables in the training sets. The maximum performance was obtained when SDL1B variables were transformed into either 4, 5, or 6 interval features in the test sets. For this dataset and Naive Bayes classifier the maximum was found for 3 equal width bins constructed for all attributes which for the learning sets had single intervals. For kNN inducer the highest recognition was obtained when there were at least 5 intervals created for cases of single bins in training sets while for these variables 4 bins were set in test sets. Thus for all three classifiers additional discretising processing brought some gains.

For male writer dataset and all three inducers employed similar overall trends in performance can be observed throughout the whole range of bins with respect to the training sets, while for varying numbers of intervals for the same variables in test some differences can be detected.

For Naive Bayes inducer in the considered range of bin numbers at most only very slight increase of performance could be found, and in the majority of cases further processing did not bring any gain but loss of classification power. The completely opposite trend is visible for J4.8 classifier, but to some extent this is the result of the commented before very low performance for the sets discretised only by Fayyad and Irani's approach. When more than single bins were established for all variables considered, the correct prediction become acceptable, in particular for as many as 9 or 10 bins for attributes in the test sets.

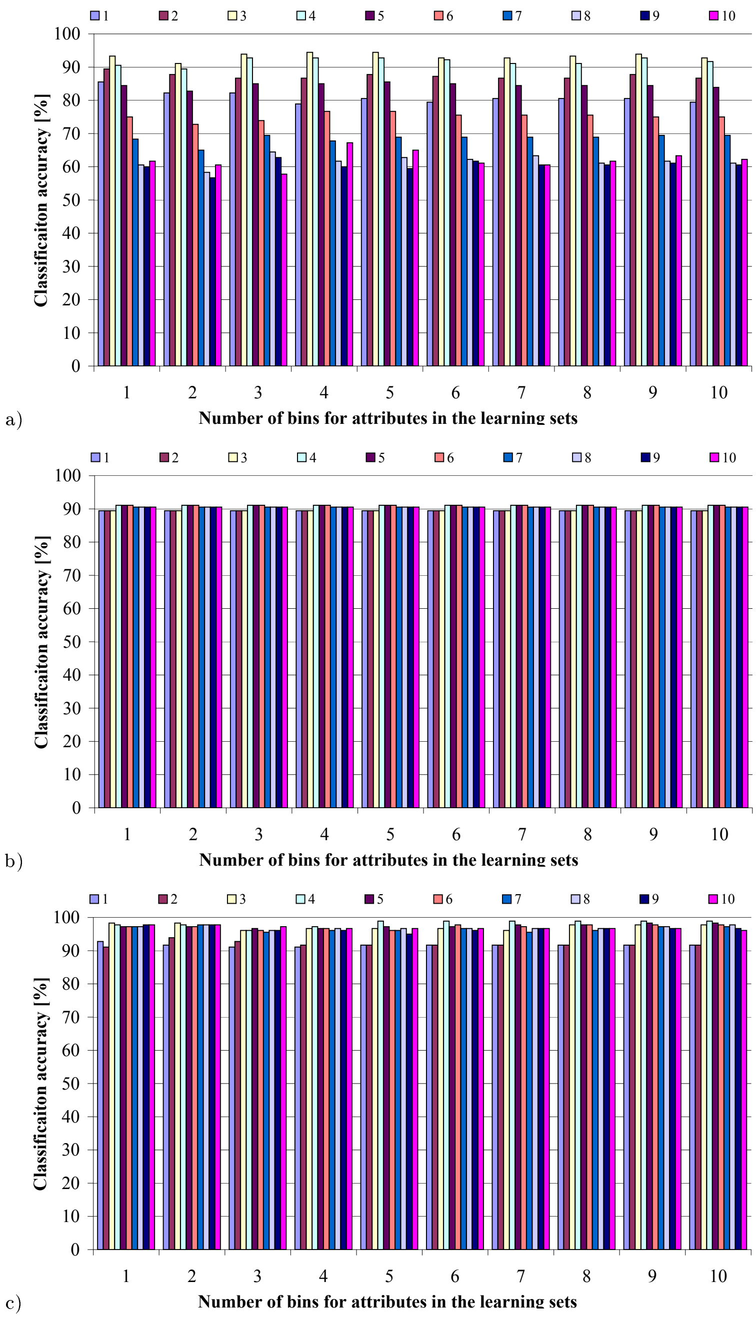

Fig. 1. Averaged classification accuracy for stylometric attributes discretised by combined supervised and unsupervised equal width binning [%] for female writer dataset, for: a) Naive Bayes, b) J4.8, c) k-Nearest Neighbours

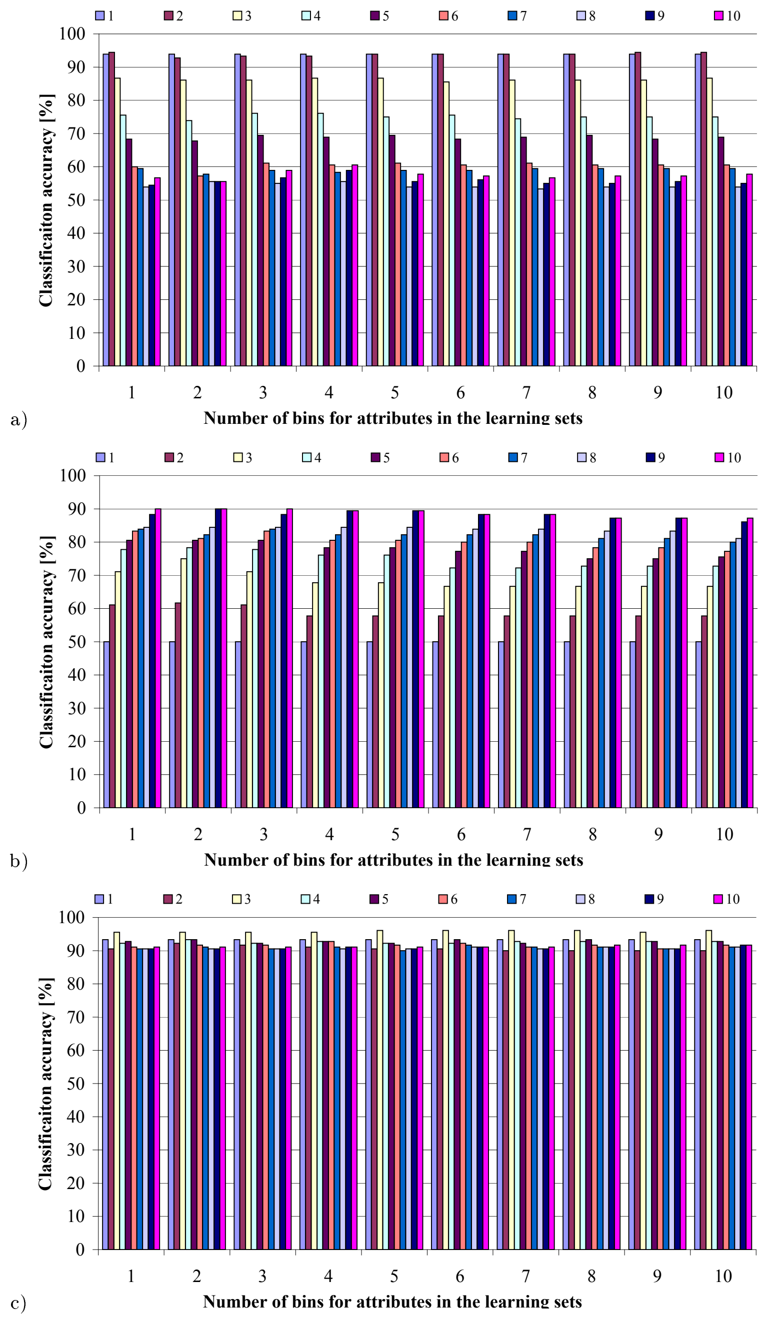

Fig. 2. Averaged classification accuracy for stylometric attributes discretised by combined supervised and unsupervised equal width binning [%] for male writer dataset, for: a) Naive Bayes, b) J4.8, c) k-Nearest Neighbours

In case of kNN classifier the differences in performance were rather slight. With increasing the number of bins in the test sets first there was detected some higher prediction level, then for even more bins the prediction decreased. The maximum was for 3 bins for variables in test sets, and between 5 and 8 intervals for these attributes in the learning sets. Which means that for two out of three tested classifiers the additional processing also brought noticeable gains.

Presented investigations on discretisation approaches applied in stylometric domain clearly show that high expectations with respect to supervised procedures can result in disappointing performance in case of discretisation of independent sets, for which even as simple method as unsupervised equal width binning can give better classification accuracy. The combination of supervised with unsupervised discretisation strategies for single bin attributes can also be used in search for improved performance.

6 Conclusions

The paper presents research on discretising approaches in the domain of stylometric processing of texts. For the considered two tasks of binary authorship attribution with balanced classes the performance of three well known classifiers, Naive Bayes, J4.8, and kNN was observed for natural for this application domain numerical characteristic features, and for several discretised sets. Three approaches to discretisation were tested: supervised Fayyad and Irani's basing on calculated entropies and referring to MDL principle as the stopping criterion, unsupervised with equal width binning, and a combination of the former two. Executed tests show that in case of discretising independent sets the performance for supervised processing can cause decrease in performance with respect to continuous domain. On the other hand unsupervised discretisation just by itself, or used as support for supervised procedures can give the expected satisfactory classification accuracy.

Acknowledgments. In the research, performed at the Silesian University of Technology within the project BK/RAu2/2017, WEKA software was used [19].

References

1. Argamon, S., Burns, K., Dubnov, S. (eds.): The structure of style: Algorithmic approaches to understanding manner and meaning. Springer, Berlin (2010)
2. Baron, G.: Comparison of cross-validation and test sets approaches to evaluation of classifiers in authorship attribution domain. In: Czachórski, T., Gelenbe, E., Grochla, K., Lent, R. (eds.) Proceedings of the 31st International Symposium on Computer and Information Sciences, Communications in Computer and Information Science, vol. 659, pp. 81–89. Springer, Cracow (2016)
3. Baron, G.: On approaches to discretization of datasets used for evaluation of decision systems. In: Czarnowski, I., Caballero, A., Howlett, R., Jain, L. (eds.) Intelligent Decision Technologies 2016, Smart Innovation, Systems and Technologies, vol. 56, pp. 149–159. Springer (2016)

4. Baron, G.: Influence of data discretization on efficiency of Bayesian Classifier for authorship attribution. Procedia Computer Science 35, 1112–1121 (2014), knowledge-Based and Intelligent Information & Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings
5. Baron, G.: Analysis of multiple classifiers performance for discretized data in authorship attribution. In: Czarnowski, I., Howlett, J.R., Jain, C.L. (eds.) Intelligent Decision Technologies 2017: Proceedings of the 9th KES International Conference on Intelligent Decision Technologies (KES-IDT 2017) – Part II, vol. 73, pp. 33–42. Springer International Publishing (2018)
6. Burrows, J.: Textual analysis. In: Schreibman, S., Siemens, R., Unsworth, J. (eds.) A companion to digital humanities. Blackwell, Oxford (2004)
7. Craig, H.: Stylistic analysis and authorship studies. In: Schreibman, S., Siemens, R., Unsworth, J. (eds.) A companion to digital humanities. Blackwell, Oxford (2004)
8. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: Machine Learning Proceedings 1995: Proceedings of the 12th International Conference on Machine Learning. pp. 194–202. Elsevier (1995)
9. Fayyad, U., Irani, K.: Multi-interval discretization of continuous valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence. vol. 2, pp. 1022–1027. Morgan Kaufmann Publishers (1993)
10. García, S., Luengo, J., Sáez, J.A., López, V., Herrera, F.: A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering 25(4), 734–750 (2013)
11. Jockers, M., Witten, D.: A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing 25(2), 215–223 (2010)
12. Kotsiantis, S., Kanellopoulos, D.: Discretization techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering 32(1), 47–58 (2006)
13. Peng, R., Hengartner, H.: Quantitative analysis of literary styles. The American Statistician 56(3), 15–38 (2002)
14. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, USA (1993)
15. Stańczyk, U.: Recognition of author gender for literary texts. In: Czachórski, T., Kozielski, S., Stańczyk, U. (eds.) Man-Machine Interactions 2, AISC, vol. 103, pp. 229–238. Springer-Verlag, Berlin, Germany (2011)
16. Stańczyk, U.: The class imbalance problem in construction of training datasets for authorship attribution. In: Gruca, A., Brachman, A., Kozielski, S., Czachórski, T. (eds.) Man-Machine Interactions 4, AISC, vol. 391, pp. 535–547. Springer-Verlag, Berlin (2016)
17. Stańczyk, U.: Evaluating importance for numbers of bins in discretised learning and test sets. In: Czarnowski, I., Howlett, J.R., Jain, C.L. (eds.) Intelligent Decision Technologies 2017: Proceedings of the 9th KES International Conference on Intelligent Decision Technologies (KES-IDT 2017) – Part II, vol. 73, pp. 159–169. Springer International Publishing (2018)
18. Stańczyk, U., Zielosko, B.: On combining discretisation parameters and attribute ranking for selection of decision rules. In: Polkowski, L., Yao, Y., Artiemjew, P., Ciucci, D., Liu, D., Ślęzak, D., Zielosko, B. (eds.) Rough Sets: International Joint Conference, IJCRS 2017, Olsztyn, Poland, July 3–7, 2017, Proceedings, Part I, Lecture Notes in Computer Science, vol. 10313, pp. 329–349. Springer, Cham (2017)
19. Witten, I., Frank, E., Hall, M.: Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 3rd edn. (2011)