Attribute Relevance and Discretisation in Knowledge Discovery: A Study in Stylometric Domain

Urszula Stańczyk[1][0000−0002−5071−7187], Beata Zielosko[2][0000−0003−3788−1094], and Grzegorz Baron[1][0000−0001−8613−631X]

[1] Department of Graphics, Computer Vision and Digital Systems, Silesian University of Technology, Akademicka 2A, 44-100 Gliwice, Poland
{urszula.stanczyk,grzegorz.baron}@polsl.pl
[2] Institute of Computer Science, University of Silesia in Katowice, Będzińska 39, 41-200 Sosnowiec, Poland
beata.zielosko@us.edu.pl

Abstract. The paper demonstrates the research methodology focused on observations of relations between attribute relevance, displayed by rankings, and discretisation. Instead of transforming all continuous attributes before data exploration, the variables were gradually processed, and the impact of such a change on the performance of a classifier was studied. Considerable experiments carried out on stylometric data illustrate that selective discretisation could be more advantageous to predictive accuracy than some uniform transformation of all features.

Keywords: Discretisation · Attribute ranking · Stylometry.

1 Introduction

Feature selection and discretisation are two processes that can have a significant impact on the operations taking place inside the knowledge discovery phase and its outcome. The discretisation of attributes produces changes in their representation and discards some information from available data [5]. As a consequence, it enables the use of inducers that operate only on categorical variables. On the other hand, exploration of transformed data can lead to overlooking some properties or relations existing in continuous features. Thus, discretisation should be treated with caution—it cannot always be assumed to be advantageous [1].

In standard discretisation approaches [5], all variables are processed in some uniform way, and the entire input domain is changed from continuous to discrete. In the methodology presented in this paper and examined through extensive experiments, the discretisation was performed in stages. Starting with the set of all attributes with continuous domains, the variables were next chosen one by one for discretisation, with the selection directed by the observed relevance of features. The importance of attributes was estimated with the help of rankings.

In the proposed methodology, some popular supervised and unsupervised discretisation algorithms were applied to the data, while following the selected

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_27

rankings. The impact of transformation on the performance of the chosen classifiers was studied for a binary authorship attribution task [12], for two datasets with balanced classes, and stylometric features in the lexical category. The experimental results show that the presented non-standard discretisation procedure led to many cases of improved predictions for subsets of features with transformed domains, making selective discretisation worth closer investigation and demonstrating the merits of the procedure for ranking-driven discretisation.

The content of the paper was organised as follows. Section 2 indicates related areas and works. Section 3 gives comments on the discretisation procedure controlled by a ranking of attributes. Section 4 details the experimental setup and presents the results. Section 5 includes concluding remarks.

2 Characteristics of Input Space and Data Mining

In the research presented, important roles were played by characteristics of the input space, algorithms used to estimate attribute importance, possible transformations of their representation, and methods employed for data exploration.

A ranking allows to estimate the importance of attributes and order them by some adopted criterion. In the research, three rankers for features were used: Support Vector Machine (SVM), Wrapper Subset Evaluation with Naive Bayes classifier (WrappB), and Correlation Attribute Evaluation (Corr).

SVM can be an effective tool for the ranking construction process [6]. The attributes are evaluated using an SVM classifier with information on how well each feature contributes to the separation of classes. WrapB uses the induction algorithm along with a statistical re-sampling technique such as k-fold cross-validation to evaluate feature subsets. NB classifier assumes that, within each class, the probability distributions for the attributes are independent of each other, so its performance in domains with redundant features can be improved by removing such variables [9]. Corr takes into account the usefulness of individual features for predicting class label along with the level of inter-correlation among them. It uses Pearson's correlation between a given feature and the class [8].

Discretisation can be considered as a process aiming at a reduction of the number of values of a given continuous variable, by dividing its range into intervals [5], which can be executed in many ways. Unlike supervised discretisation, unsupervised algorithms ignore instance labels during the transformation of attributes. In the proposed methodology, four methods were used: supervised Fayyad and Irani (dsF) [4], and Kononenko (dsK) [10] algorithms, and unsupervised equal width (duw) and equal frequency (duf) binning.

Classification is one of the main tasks in the process of knowledge discovery and pattern recognition. In this work, two state-of-the-art classifiers were used, namely, Bayesian Network (BNet) [11] and Random Forest (RF) [3].

Bayesian Network is a probabilistic model based on Bayes' theory [11]. It is considered as a representation of joint probability distribution over a set of random variables and presented in the form of a directed acyclic graph, where each node corresponds to a random variable and the edges represent probabilistic

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_27

dependence. Random Forest belongs to the ensemble data mining techniques used for classification [3]. It is a combination of decision trees as predictors such that each tree depends on the values of a random vector, sampled independently from a dataset, and with the same distribution for all trees in the forest. During classification, each tree votes, and the most popular class label is returned.

3 Procedure for Ranking Driven Discretisation

To be unbiased, the methodology for ranking-driven discretisation of attributes required some assumptions and limitations. These elements were as follows.

- Input data and features. Attributes are expected to be continuous, with comparable ranges of values, and should be chosen only on the basis of domain knowledge. A classification task is binary and with balanced classes.
- Ranking mechanisms. The methods to be used must treat all variables as relevant by always assigning a non-zero rank and work in continuous domain.
- Discretisation algorithms. A discretiser needs to work independently on a learner, process variables separately, and ignore any interdependencies among them. Both supervised and unsupervised approaches can be employed.
- Classifiers. The learner needs to be able to operate in both continuous and discrete domains, and capable of discovering knowledge from both forms.
- Starting point. The exploration of data starts in continuous domain. Based on the knowledge discovered in real-valued variables in the train sets, performance is next evaluated by labelling samples in the test sets.
- Steps and direction of processing. Each step involves discretisation of a single attribute, indicated by its ranking position, following either up or down.
- Stopping point. The procedure can be stopped once the entire datasets become discrete. It is also possible to end transformations sooner, when some degradation of performance is detected.

4 Experiments

The experiments started with the construction of the input datasets. Then, the attribute rankings were obtained and used for gradual discretisation of sets. For the selected classifiers, their performance was estimated and investigated.

4.1 Preparation of Input Stylometric Datasets

Two pairs of authors were taken for stylometric analysis, Edith Wharton and Mary Johnston (female writer dataset, F-writers), and Henry James and Thomas Hardy (male writer dataset, M-writers). Long texts of novels were divided into smaller chunks of text of comparable lengths. Over these shorter texts, the frequency of occurrence was calculated for 12 selected function words [15]. They belong to the lexical category of stylometric descriptors widely used in authorship attribution tasks. The words were as follows: as, at, by, if, in, no, of, on, or, so, to, up. When mentioned in the text, the attributes were given in italics.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_27

The preparation resulted in real-valued features, to be employed by some approach to data mining [12]. Due to the specifics of the sample construction process, the input space was stratified. Taking this into account [2], a dataset consisted of a train set and two test sets for performance evaluation. All sets were balanced, including the same number of samples for both classes.

4.2 Rankings of Characteristic Features

Three ranking mechanisms were applied to the data, all implemented in the WEKA [7] workbench, namely WrapB, Corr, and SVM. The obtained orderings of the variables were provided in Table 1, where the highest ranking position was shown on the left and the lowest on the right.

Table 1. Rankings of attributes for the F-writer and M-writer datasets.

F-writer dataset														M-writer dataset											
Ranking position														Ranking position											
1	2	3	4	5	6	7	8	9	10	11	12	Ranker	1	2	3	4	5	6	7	8	9	10	11	12
to	on	of	no	at	if	so	up	in	or	by	as	WrapB	by	if	to	in	so	no	at	of	as	on	up	or
on	to	of	as	by	no	or	so	in	if	at	up	Corr	by	if	in	or	at	of	to	so	no	on	up	as
on	to	of	as	so	by	if	at	up	no	or	in	SVM	by	or	if	at	in	so	no	to	as	on	up	of

The two datasets shared the same features, but their placement in the rankings was different. For a dataset, some similarities could be observed between the rankings. For M-writers, *by* was always the highest ranking, for F-writers, *to* and *on* took the top two positions, and *of* always came third. Lower-ranking positions were more varied. All rankings were followed in ascending order.

4.3 Employed Discretisation Algorithms

In the experiments, all sets were discretised independently [13]. Unsupervised methods (duf and duw) were employed with the number of bins from 2 to 10, so returned 9 variants of data each. Two supervised algorithms (dsF and dsK), gave single variants of the data. These methods rely on the MDL principle and the calculation of entropy, which led to some variables for which one interval was found as representation in a discrete domain. For F-writers for both supervised discretisation algorithms, there were 6 such features. For M-writers and dsK, also 6 variables had single bins, and for dsF this set was expanded to 7 elements.

4.4 Classification Process and Evaluation of Performance

The primary goal of the research was to observe the relations between the importance of features and discretisation, and how the changed representation reflects on classifier performance. As the processing started in the continuous domain and the features were discretised gradually, only at the final step were all variables discrete. With such conditions, the selected classification systems for the

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_27

most part operated on at least partially continuous data. Two chosen classifiers, BNet and RF, implemented in WEKA, were used with default parameters.

To evaluate inducer performance, classification accuracy was chosen [14], as it is suitable for binary classification with balanced classes, both classes of the same importance and the same cost of misclassification. Due to the stratified input space, cross-validation could not be considered reliable [2]. Instead, test sets were used, and the reported performance is the average obtained from them.

For the Bayesian Network classifier, the performance was shown in Fig. 1, and for Random Forest in Fig. 2. In the included charts, the entire processing path was illustrated, starting with zero discrete attributes and ending with 12 discrete features. For unsupervised methods, the average was calculated over all 9 variants of the data, corresponding to different numbers of constructed bins.

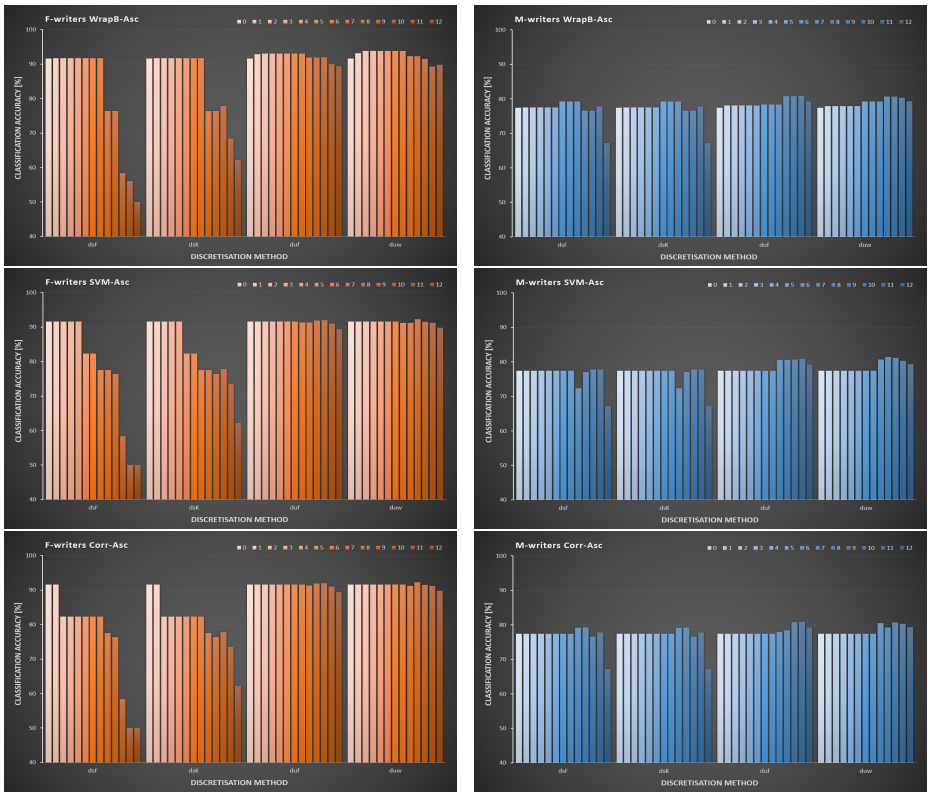

Fig. 1. Performance [%] for the Bayesian Network observed in the discretisation while following the selected rankings. The series specify the number of discretised attributes.

For BNet, both datasets, and all three rankings unsupervised discretisation always brought some improvement for partial transformation of attributes. Supervised discretisation, in particular for more transformed variables, resulted in

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_27

cases of degraded performance, especially for F-writers. For the female writer dataset WrapB resulted in the best performance for all discretisation methods, while for M-writers WrapB and Corr came very close, with SVM slightly behind.

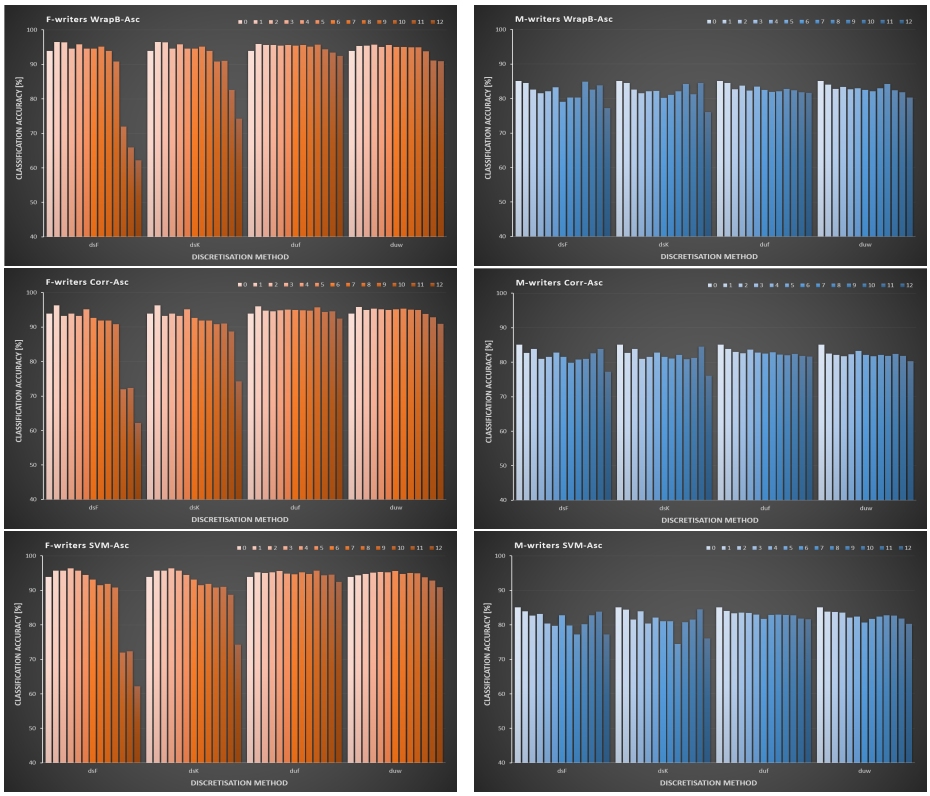

Fig. 2. Performance [%] for the Random Forest observed in the discretisation while following the selected rankings. The series specify the number of discretised attributes.

Because of its mode of operation on data, the RF classifier on the whole fared better in the continuous domain than in the discrete domain when all features were transformed. However, partial discretisation was often advantageous, in particular for F-writers, where, for all rankings and all discretisation methods, improved accuracy was always detected. The maxima observed for rankings were close but again the highest for WrapB, and better for supervised discretisation.

To study the ranges of classification accuracy obtained in gradual discretisation of attributes, the average performance and standard deviation were calculated for the entire transformation process, as shown in Table 2. For reference, the performance of a classifier observed in the continuous domain was also provided. In each row, the highlighted entries correspond to the highest average predictive accuracy for this discrete version for each dataset.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_27

Table 2. Average performance [%] and standard deviation of the Bayesian Network and Random Forest classifiers, for the entire run of the selective discretisation procedure following a ranking (from 1 out of N, to all N discretised attributes).

	F-writers			M-writers		
Discret.	Bayesian Network (Cont.domain: 91.60)			Bayesian Network (Cont.domain: 77.43)		
method	WrapB	Corr	SVM	WrapB	Corr	SVM
dsF	**79.95**±16.32	74.83±13.94	76.74±15.71	**76.94**±03.20	76.83±03.12	76.19±03.19
dsK	**83.61**±10.83	79.44±07.04	81.34±09.12	**76.94**±03.20	76.83±03.12	76.19±03.19
duf	**92.20**±01.27	91.42±00.67	91.40±00.67	**78.96**±01.18	78.27±01.33	78.68±01.59
duw	**92.62**±01.66	91.43±00.59	91.40±00.59	**79.06**±01.10	78.51±01.40	78.75±01.70
	Random Forest (Cont.domain: 91.96)			Random Forest (Cont.domain: 85.07)		
dsF	**87.70**±12.92	87.15±11.40	87.67±11.75	**81.87**±02.29	81.57±01.85	81.16±02.38
dsK	**91.68**±06.68	91.11±05.68	91.63±05.99	**81.89**±02.29	81.63±02.09	81.01±03.03
duf	**95.05**±01.08	94.76±00.87	94.82±00.85	82.67±00.85	82.62±00.70	**82.85**±00.77
duw	94.43±01.64	**94.54**±01.38	94.43±01.33	**82.70**±01.02	82.04±00.68	82.35±01.10

Since the averages were calculated over the whole discretisation process and often the transformation of all attributes caused degraded performance, the averages also often fell below the level reported for the continuous domain. However, for the Bayesian Network and unsupervised methods the values were improved, while for Random Forest that was true only for F-writers and duf method. On the other hand, the highest averages were mostly reported for WrapB ranking, for both datasets and both classifiers.

Examination of standard deviation allows to conclude that the highest values, even in two digits, were found when supervised discretisation was applied to the input data, in particular to the female writer dataset. This observation is valid for both studied classifiers and all three rankings. For other discretisation approaches, the values were noticeably smaller, even just fractional.

The statistical characteristics of the process of selective discretisation driven by rankings confirmed earlier observations based on performance trends. In the majority of cases partial discretisation was more advantageous than transformations of entire input domain, showing that attribute relevance incorporated into their processing and discretisation could be beneficial to the knowledge discovery.

5 Conclusions

In the paper a research methodology was demonstrated in which the discretisation process was carried out with gradually expanding the range of transformed variables, selected from the obtained ranking reflecting their relevance. The goal of this processing was to observe the relations between the importance of features and the form of their representation, and how its change can influence the performance of selected classifiers. Extensive experiments were carried out in the stylometric domain, for the binary authorship attribution task. Several rankings and many discretisation methods were investigated and they allowed to discover a remarkable number of cases where partial, instead of complete discretisation of the input data, led to obtaining improved predictive accuracy for the used inducers, making selective discretisation worth deeper study.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_27

Acknowledgements The research works presented in the paper were carried out within the statutory project of the Department of Graphics, Computer Vision and Digital Systems (RAU-6, 2023), at the Silesian University of Technology, Gliwice, Poland, and at the University of Silesia in Katowice, Sosnowiec, Poland.

References

1. Baron, G., Stańczyk, U.: Performance evaluation for ranking-based discretisation. In: Cristani, M., et al. (eds.) Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES-2020, Procedia Computer Science, vol. 176, pp. 3335–3344. Elsevier (2020)
2. Baron, G., Stańczyk, U.: Standard vs. non-standard cross-validation: evaluation of performance in a space with structured distribution of datapoints. In: Wątróbski, J., et al. (eds.) Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES-2021, Procedia Computer Science, vol. 192, pp. 1245–1254. Elsevier (2021)
3. Cutler, A., Cutler, D.R., Stevens, J.R.: Random forests. In: Zhang, C., Ma, Y. (eds.) Ensemble Machine Learning: Methods and Applications, pp. 157–175. Springer, NY, US (2012)
4. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuousvalued attributes for classification learning. In: 13th International Joint Conference on Articial Intelligence. vol. 2, pp. 1022–1027. Morgan Kaufmann Publishers (1993)
5. 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 (April 2013)
6. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46**, 389–422 (2002)
7. Hall, M., et al.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1), 10–18 (2009)
8. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. Ph.D. thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand (1998)
9. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97**(1), 273–324 (1997)
10. Kononenko, I.: On biases in estimating multi-valued attributes. In: 14th International Joint Conference on Articial Intelligence. pp. 1034–1040 (1995)
11. Sardinha, R., Paes, A., Zaverucha, G.: Revising the structure of bayesian network classifiers in the presence of missing data. Information Sciences **439-440**, 108–124 (2018)
12. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology **60**(3), 538–556 (2009)
13. Stańczyk, U., Zielosko, B.: Data irregularities in discretisation of test sets used for evaluation of classification systems: A case study on authorship attribution. Bulletin of the Polish Academy of Sciences: Technical Sciences **69**(4), 1–12 (2021)
14. Stąpor, K., Ksieniewicz, P., García, S., Woźniak, M.: How to design the fair experimental classifier evaluation. Applied Soft Computing **104**, 107219 (2021)
15. Zhao, Y., Zobel, J.: Effective and scalable authorship attribution using function words. In: Lee, G.G., Yamada, A., Meng, H., Myaeng, S.H. (eds.) Information Retrieval Technology. pp. 174–189. Springer, Berlin, Heidelberg (2005)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_27