BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 69(4), 2021, Article number: e137629
DOI: 10.24425/bpasts.2021.137629

Data irregularities in discretisation of test sets used for evaluation of classification systems: A case study on authorship attribution

Urszula STAŃCZYK[1]* and Beata ZIELOSKO[2]

[1] Silesian University of Technology, ul. Akademicka 2A, 44-100 Gliwice, Poland
[2] University of Silesia in Katowice, ul. Będzińska 39, 41-200 Sosnowiec, Poland

Abstract. When patterns to be recognised are described by features of continuous type, discretisation becomes either an optional or necessary step in the initial data pre-processing stage. Characteristics of data, distribution of data points in the input space, can significantly influence the process of transformation from real-valued into nominal attributes, and the resulting performance of classification systems employing them. If data include several separate sets, their discretisation becomes more complex, as varying numbers of intervals and different ranges can be constructed for the same variables. The paper presents research on irregularities in data distribution, observed in the context of discretisation processes. Selected discretisation methods were used and their effect on the performance of decision algorithms, induced in classical rough set approach, was investigated. The studied input space was defined by measurable style-markers, which, exploited as characteristic features, facilitate treating a task of stylometric authorship attribution as classification.

Key words: discretisation; data irregularities; evaluation and test sets; rough sets; authorship attribution; stylometry.

1. INTRODUCTION

Initial pre-processing of data can take much more time than actual data mining, and can greatly affect obtained results [1]. Depending on the nature of input datasets, and techniques used for knowledge exploration and discovery, some steps in the data preparation can be considered as optional or indispensable [2]. One such step is discretisation, responsible for the reduction of information. It transforms continuous attributes describing concepts into their categorical forms [3]. The process is influenced by characteristics of data, and occurring irregularities can result in obtaining such representations that cause unsatisfactory performance of classification systems [4].

When input data is divided into several disjoint parts, such as corresponding to train, evaluation, and test sets, discretisation becomes even more complicated, as properties of variables studied in the local context of these sets are practically never the same [5]. With the governing idea of discovering knowledge only from the learning samples, they can serve as a base for construction of a discretisation model – the lists of ranges representing categories of values for attributes. This model can be employed to interpret values present in evaluation and test sets, which leads to their translation into discrete domain [6]. However, when data points are scattered, with highly disjoint groupings, in this processing it is possible that for some of attribute values in a test set there is no good match in any of the intervals defined for it.

On the other hand, all sets can be discretised independently of each other. With this kind of transformation, the discretisation model based on test data is yet another factor compared to the one obtained for train data. It is safe to assume that independent processing will cause different cut-points between constructed intervals, but also, depending on selected discretisation approaches, varying numbers of bins can be obtained. Since cardinalities of learning and test sets most often vary, not always the same ranges of parameters for discretisation can be used for all sets. Thus the obtained discretisation models can be used to observe irregularities between train and evaluation and test sets.

The paper presents research in which input data were discretised by several selected approaches, and with varying parameters. From each version of discrete train sets decision rules were induced. Then, with weighted voting as a conflict resolution strategy [7], decision algorithms were employed in classification of all variants of learning, and evaluation and test sets. The latter were discretised in two ways: independently, and with transformation based on discrete train sets. For all these tested combinations of sets, the resulting performance of inducers was studied.

Decision rules are often preferred as classification systems, as they offer a transparent representation of patterns discovered in data, by conditions included in rules [8]. In the research they were inferred in rough set approach, implemented in Rough Set Exploration System (RSES) [9]. Rough set theory is well suited to data mining tasks with uncertain and incomplete knowledge [10]. In a rough set perspective, the universe is seen through granules corresponding to equivalence classes of objects that cannot be discerned, based on values of considered

*e-mail: urszula.stanczyk@polsl.pl

Manuscript submitted 2020-02-20, revised 2021-03-14, initially accepted for publication 2021-05-21, published in August 2021

attributes. The classic approach requires categorical characteristic features [11].

The input space, used in experiments, was defined to solve a task of authorship attribution in the field of computer-assisted stylometric analysis of texts [12]. In the domain of stylometry, recognition of authorship is based on stylistic fingerprints: quantitative descriptors that reflect linguistic preferences and habits of writers, often involving the calculation of frequencies of occurrence for selected words or characters [13]. It makes stylometric features continuous-valued [14]. Based on texts, a classifier assigns authors as class labels to samples, and to measure its performance application of test sets gives much more reliable predictions than typically used standard cross-validation [15]. This is a reason why stylometric data is a well suited example for illustration of the described research works on data irregularities and discretisation.

The objectives of the research presented, and main contributions of this paper are:

- exploration of a new aspect of data irregularity problem. Very often in the literature, this problem is studied in terms of the number of samples representing a given decision class [16, 17], for example, class imbalance, class distribution, or missing/absent values issues. In the paper, the problem of data irregularities is considered in the framework of discretisation process, which can be performed in different ways;
- investigation into the influence of data irregularities on categories and data models constructed for continuous attributes within discretisation through various methods and approaches, in particular in the case of disjoint datasets, such as train, evaluation and test sets, for their independent transformations, and for translation of test sets based on ranges formed for train sets;
- examination of inducers performance for a wide range of variants for all datasets, treating comparisons of different discrete data models as a part of classification processes.

The content of the paper is organised as follows. Section 2 includes comments on various irregularities that can be observed in data. Section 3 presents the problem of recognising authorship through stylometric features. Section 4 provides the background for selected discretisation approaches. The fundamental notions of rough set theory are described in Section 5. Experimental set-up and test results are detailed in Section 6. Section 7 contains conclusions and directions for future research.

2. DATA IRREGULARITIES

Data irregularities can be observed in many real-life applications [18], as well as data mining tasks, which try to deal with this issue [19]. In this framework, one of the most popular research topics is the study of the influence of certain data irregularities on obtained classification results. From this point of view, data irregularities can be considered as the following problems [4, 20]:

- class imbalance, where the classes in a dataset are not equally represented. It is a common form of the distribution-based data irregularity, with one or more classes under-represented in a set, while some other class or classes are over-represented. It leads to distinction of minority and majority classes;
- small disjuncts, when there are small (that is under-represented) subconcepts within classes;
- class distribution skew, when different classes possess significant disparate class distributions;
- missing features, in the cases of somehow lost or simply unrecorded values of some attributes;
- absent features, where certain variables can be undefined or non-existing for certain data instances rather than having an unobserved or unrecorded value.

To minimise the bias caused by data irregularities on classifier performance, several approaches are employed that attempt to make up for the listed problems. For example, under- and over-sampling for imbalanced classes either decreases or increases numbers of instances in order to reach balance [21]. Missing or absent feature values can be replaced by averages or most common values [22]. Irregularities can also be studied in the context of data processing methods, such as discretisation. The research described was dedicated to the examination of classification accuracies for decision algorithms induced from, and then tested on various discrete versions of the same data.

3. PROPERTIES OF STYLOMETRIC DATA

Stylometry advocates uniqueness of writing styles, visible in linguistic characteristics, observed for all authors [23]. Whatever particular topic they write about [24], through their individual preferences and habits, the authors leave their stylistic fingerprints in *how* they write, which leads to authorial profiles [13]. Not only can a stylistic profile be generally described, but also its approximation can be expressed by measurable characteristic features, specific to writers, to the point of reliable recognition of authorship [25]. The degree of precision of definitions and descriptions, obtained for stylistic profiles, determines techniques and algorithms that can be used for stylometric data mining [26].

Author characterisation and comparison make attribution possible, and these three are the main stylometric tasks. For authorship attribution there can be executed calculations referring to statistics [12], for example, based on language models and probabilities of occurrence of transitions between characters, letters [27]. Also methods from the computational intelligence domain are employed [28, 29]. In this case the task of authorship attribution becomes a problem of supervised learning, where an inducer is trained on a part of a corpus of attributed texts, and then the discovered stylistic patterns are compared with the another part of the corpus, in order to label previously unknown samples into classes corresponding to recognised authors.

Corpus construction is an important part of stylometric data mining process [30]. Due to the high variety of linguistic features that make a base for any language, style-markers that can be employed are also numerous, and they are often categorised as belonging to one of the four main types:

- lexical: giving frequencies of occurrence, distributions, and averages for characters, words, specific phrases, sequences of words (word n-grams);
- syntactic: reflecting sentence formulation by employed punctuation marks;
- structural: describing the organisation of a text in some units (such as paragraphs or chapters), also specific formatting;
- content-specific: detecting important words in the context, of special significance or meaning.

Descriptors need to be calculated over several text samples, the more the better representation obtained for a stylistic profile. For the values of features to be comparable, lengths of these samples need to be as close as possible, and the samples should be of sufficient length [31], as for short forms of writing different stylometric rules apply [32]. With these considerations in mind, and taking into account that not that many writers author a sufficiently high number of long texts, it is a widely applied practice to divide longer works into several smaller parts. Not only does it increase the number of available samples, it also allows for closer observation of stylistic variations exhibited in long texts, written over longer periods of time. Otherwise, such intricacies would be invisible in general calculations and conclusions.

Operation on examples, obtained from text samples being a part of some larger whole, comes at a certain risk. As they reflect some variations of a style, in which this longer work was written, such samples show closer similarity to each other than to examples based on entirely different texts [33]. If all works of a writer were collected and divided into smaller parts, then the input space would include unevenly distributed data points, grouped according to certain texts they were part of.

Standard cross-validation approach, popularly used for evaluation of classifiers, assumes testing on randomly selected samples not used for training, but in a space where no specific similarity among samples is detected. The properties of stylometric data cause this approach to be unreliable, as it is not only possible but highly probable that for testing there would be used samples from the same novel that was used in training, thus closer in style and easier to correctly recognise. The resulting performance would be falsely increased, as the improved predictions would be caused by leakage of information from training into testing. It is the reason why evaluation and test sets, constructed for evaluation of performance, need to use samples based on separate texts, never used for knowledge mining. For stylometric data, reliable cross-validation would mean swapping not just samples, but original long texts between training and test sets, but such processing brings very high computational costs [30].

4. OBJECTIVES AND ALGORITHMS OF DISCRETISATION

One of the stages in knowledge discovery is data preprocessing. It can include discretisation, which transforms numerical attributes into categorical ones with a finite number of intervals [3]. Discretisation can be considered a part of data reduction methods as it maps continuous space of attributes values into a reduced subset of discrete values. From this point of view, discretisation simplifies the data and removes possible noise, so the data are easier to use and interpret. Discretisation usually causes some loss of information, therefore it should always be used with caution, and adjusted to data and their properties, as existing irregularities influence the outcome.

Generally, discretisation can be considered a process consisting of four steps:

- sorting all values of a discretised attribute;
- determining cut-points for splitting or merging intervals;
- executing splitting/merging according to an algorithm criterion;
- evaluating the stopping condition of the discretisation algorithm, and assigning values of the discretised attribute from the input set to one of the evaluated intervals.

There are many discretisation methods and algorithms, which can be grouped based on various criteria. One of the most popular is division into *supervised* and *unsupervised*. For supervised methods, information about classes is taken into account while searching for intervals among ranges of attribute values. Some heuristic measures, e.g. entropy [34], can be used to determine the best cut-points. In the case of unsupervised methods, information about class labels is omitted during discretisation.

Another division of the algorithms is into *local* and *global*. Local discretisers are defined separately for distinctive parts of an attribute domain. Global methods consider the whole attribute domain to define the initial set of candidate cut-points.

Static discretisation methods are independent of the learning algorithm and are performed before the learning task. *Dynamic* methods are based on the information exchange between discretiser and learner units, and can be considered as a component of the learning algorithm. Most of discretising algorithms are static, and dynamic discretisers are considered as a part of data mining algorithms.

Discretisation can either be *top-down* or *bottom-up*. Bottom-up approaches initially consider a number of intervals determined by the set of cut-points, and then these intervals are merged until a certain stopping criterion is achieved. In top-down methods at the beginning one big interval containing all known values of an attribute is considered, and then partitioning of this interval into smaller and smaller subintervals is performed, until a certain stopping criterion is reached.

In *univariate* methods, discretiser is working with a single attribute at a time. In *multivariate* approaches, the work of a discretiser is based on interactions among attributes, and simultaneously values of all attributes are studied to define the initial set of cut-points.

4.1. Supervised discretisation algorithms

Fayyad and Irani [35], and Kononenko [36] methods were two representatives of supervised discretisation approaches applied in the research works reported in this paper. Both belong to static, global, top-down, univariate approaches. These two methods base evaluation of candidate cut-points on class entropy of considered intervals, and Minimum Description Length (MDL) [37] principle as a stopping criterion.

4.1.1. Fayyad and Irani MDL

Let set S contain N instances and k decision classes $C_1,\ldots,C_k$. Class entropy $Ent(S)$ of S is defined as follows:

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

where $P(C_i,S)$ is the proportion of class C_i instances included in the set S.

Taking into account binary discretisation of a continuous attribute a, the selection of cut-point is made by testing all possible candidates T. A cut-point T splits the set S into two subsets, S_1 and S_2, where $S_1 \subset S$ contains instances with attribute values $\leq T$ and $S_2 = S \setminus S_1$. Entropy for T is 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 cut-point T_{opt}, class information entropy $Ent(a,T_{opt};S)$ is minimal. The process of evaluating cut-points starts from a single interval of a discretised attribute, and is repeated again and again until the stopping criterion is met.

For Fayyad and Irani approach, the stopping criterion is connected with information gain, considered as the difference between the entropy for the whole range and the entropy after selecting cut-point T. The discretisation process is applied recursively until the following inequality (3) is satisfied:

$$\begin{aligned} Gain(a,T;S) = Ent(S) - E(a,T;S) > \\ \frac{\log_2(N-1)+\log_2(3^k-2)}{N} + \\ -\frac{[k\cdot Ent(S) - k_1\cdot Ent(S_1) - k_2\cdot Ent(S_2)]}{N}, \end{aligned} \tag{3}$$

where k_1 and k_2 give the numbers of classes, distinguished correspondingly in the sets S_1 and S_2 after the split.

4.1.2. Kononenko MDL

In the case of Kononenko method, the discretisation process is recursively executed until the inequality (4) is true:

$$\begin{aligned} \log\binom{N}{N_{C_1}\ldots N_{C_k}} + \log\binom{N+k-1}{k-1} > \log N_T + \\ \sum_j \log\binom{N_{a_j}}{N_{C_1a_j}\ldots N_{C_ka_j}} + \sum_j \binom{N_{a_j}+k-1}{k-1}, \end{aligned} \tag{4}$$

where

- N – the number of training instances,
- N_{C_i} – the number of training instances from the class C_i,
- N_{a_x} – the number of instances with x-th value of the given attribute a,
- $N_{C_ia_y}$ – the number of instances from class C_i with y-th value of the given attribute a,
- N_T – the number of possible cut-points.

4.2. Unsupervised discretisation algorithms

Equal width and equal frequency binning are representatives of unsupervised discretisation methods. In both the input parameter k determines the number of bins constructed for each attribute. Each bin is associated with a distinct discrete value.

Equal width binning algorithm sorts the values of a continuous attribute, designates the minimum and maximum and divides this range into k equal width discrete intervals. In the case of equal frequency binning, an equal number of continuous values are placed in each bin. So, the minimum and maximum values of the discretised attribute are determined, sorted and the range of values is divided into k intervals where each bin contains the same number of sorted values [38].

The two methods are simple and sensitive to data irregularities. When values of a continuous attribute are not distributed evenly, some information can be lost after the discretisation process. For the equal frequency approach, many occurrences of a continuous value could cause that such value is assigned into different bins. Therefore, during the selection of cut-points, it is important that duplicated values are assigned only to one and the same constructed bin.

5. ROUGH SET THEORY

Rough set theory (RST) was proposed by Z. Pawlak in 1982 as a mathematical tool for work with inconsistent and imprecise data [11]. Perception of knowledge through its granular structure is a feature of RST. An elementary set contains all indiscernible objects (i.e. characterised by the same values of attributes). It forms a granule of knowledge about the universe. It means that in the rough set theory, granules of indiscernible objects, instead of particular objects, are considered.

A union of elementary sets is referred to as a crisp (precise) set, otherwise the set is rough (imprecise). Each rough set has objects which cannot be properly classified by employing available knowledge, and they are called boundary-line cases. So, in RST, rough concepts cannot be characterised in the framework of knowledge available about their elements. Hence, there are used approximations of the rough concept.

Any imprecise concept is replaced by a pair of precise concepts called the lower and the upper approximation of the rough concept. The lower approximation consists of objects which surely belong to the concept and the upper approximation contains objects which possibly belong to it. The difference between the upper and the lower approximation constitutes the boundary region of the rough concept. If the boundary region of a set is nonempty, it means that our knowledge about the set is insufficient to define the set precisely.

In the rough set theory, the main structure for data representation is an *information system*, and its special case – *decision table*. Information system is a pair of the form $S = (U,A)$, where U is a nonempty, finite set of objects, and $A = \{a_1,\ldots,a_m\}$ is a nonempty, finite set of attributes, i.e., $a_i : U \to V_a$, where V_a is the set of values of attribute a_i, called the domain of a_i. A decision table is a pair of the form $S = (U,A\bigcup\{d\})$, with a distinguished attribute $d \notin A$. The attributes belonging to A are called *condition attributes* while d is called a *decision*.

Sets of decision rules are a popular form of knowledge representation, used in many areas connected with data mining. In the paper, decision rules are formulas presented in the form:

$$(a_{i_1} = v_1) \wedge \ldots \wedge (a_{i_k} = v_k) \to d = v_d, \tag{5}$$

where $1 \leq i_1 < \ldots < i_k \leq m, v_i \in V_{a_i}, 1 \leq v_d \leq |V_d|$.

Many algorithms for the construction of decision rules exist [8, 39]. In the research presented, an exhaustive algorithm, implemented in RSES system [9] was used. This algorithm allows to construct all decision rules with a minimal number of descriptors (pairs attribute=value).

6. EXECUTED EXPERIMENTS

The experiments started with the determination of writers to be compared and recognised in the author attribution tasks, and division of their available works into three disjoint groups. The novels were next partitioned into smaller text parts of a comparable size, over which selected stylometric markers were calculated. This resulted in the construction of respectively learning (training), and evaluation and test sets.

After this initial preparation of data, all sets were discretised through various approaches, with varying discretisation parameters, with independent processing of all sets, and with obtaining discrete test sets while using models build on discrete learning data in their corresponding versions. The processing resulted in obtaining several variants of all sets.

Next, from all versions of discrete training sets decision rules were inferred in rough set approach, by an exhaustive algorithm. The obtained sets of rules were then used to classify samples, applying standard voting in the case of conflicts, that is weighting the vote of each rule by its support. The decision algorithms were applied to all sets within a dataset – all variants of a learning set, and all versions of evaluation and train sets, constructed within the same discretisation approach. In all cases the resulting performance was studied.

The details of the experiments are given below.

6.1. Preparations of input datasets

Two pairs of writers were chosen for analysis, Edith Wharton and Mary Johnston (Wharton-Johnston, or W-J dataset), and Jack London and James Curwood (London-Curwood, or L-C dataset). This particular pairing of selected writers was not coincidental. As stylistic profiles for female and male writers show certain distinguishing characteristics [40], their comparison (for example Wharton with Curwood) would mean introducing additional factor into the equation that has non-negligible potential of influencing obtained predictions.

As the governing idea is for all text samples to be attributed to their authors, regardless of the topic, similarity or dissimilarity of the genre, these elements were disregarded in corpus construction [24]. For all four writers 10 novels were randomly selected and then grouped, 4 for training data, and two times 3 novels for the two test sets. All works were further divided into text parts, and per author for learning 25 such samples per novel were taken, and for both test sets 15 per novel. It resulted in 200 samples in a training set (100 per author), and 90 per each evaluation and test set (45 per writer), making all sets balanced with respect to recognised classes and classification binary [41].

Then, using the list of the most popular words in English language and a set of punctuation marks, the frequencies of occurrence of 100 lexical and syntactic style-markers were calculated over all text samples [42]. To the results obtained for both training sets (for W-J and L-C data), in the next step several ranking algorithms for features were applied [43]. The attributes that obtained low rank were discarded, and only these always highest-ranking left, which returned the set of 24 characteristic features (22 lexical and 2 syntactic at the end), with the values for all attributes in the range $< 0,1)$, corresponding to occurrence frequencies for:

after almost any around before but by during how never on same such that then there though until what whether who within ; ,

6.2. Discretisation of input datasets

For the discretisation of input data four methods were chosen, two examples from the supervised category, and two representatives of the unsupervised group. All these algorithms are implemented in WEKA workbench [44], which was employed in the execution of the experiments.

Supervised discretisation with both Kononenko, and Fayyad and Irani approaches is non-parametric and calculations involved have a highly local context, so they were selected for the purpose of gathering information on data irregularities observed for attributes. Independent application of these methods to all sets of samples resulted in obtaining significantly different characteristics for the same features in various sets [5], as shown in Table 1 and Table 2 respectively for two datasets.

Table 1

Characteristics of attributes for independent supervised discretisation of Wharton-Johnston dataset

Fayyad and Irani algorithm		
Set	Bins	Attributes
Train	1	during though almost within
	2	that by what who there how then any whether after never same such before until around ;
	3	on ,
	4	but
Test 1	1	that but by then during before though whether almost
	2	on what who there how any after never same around such until within ;
	3	,
Test 2	1	that but by what such how then there same whether around during before though almost
	2	on who any after never until within ;
	3	,
Kononenko algorithm		
Set	Bins	Attributes
Train	1	almost
	2	that by what who there how then until such before any never after same during though whether around within ;
	3	on ,
	4	but
Test 1	1	but by then before though almost whether
	2	that on what who there how after any around never same such during until within ;
	3	,
Test 2	1	that but by how then such during before though almost whether around
	2	on what who there any after never same until within ;
	3	,

Table 2
Characteristics of attributes for independent supervised discretisation of London-Curwood dataset

Fayyad and Irani algorithm		
Set	Bins	Attributes
Train	1	never during
	2	that who but by there what how then almost until such after any same before though whether around within ;
	3	on ,
Test 1	1	before but what how then after never same whether who during though within
	2	that on by there any such until almost around ; ,
Test 2	1	that but what who how then after any whether never during same such before though almost
	2	on by there until around within ,
	3	;
Kononenko algorithm		
Set	Bins	Attributes
Train	1	during
	2	that by what who there how then any never same before almost after such though until whether around within ;
	3	on but ,
Test 1	1	but what who then after never same whether within during before though
	2	that on there how any such until almost around ; ,
	3	by
Test 2	1	that what how then after same such almost though during before whether who
	2	on but by there any never until around within
	3	; ,

For all sets there are shown numbers of intervals defined for each attribute in a set, and these numbers ranged from 1 (the cases where for the whole range of continuous values for some attribute there is a single discrete representation) to the maximum of 4. Only a part of attributes had the same numbers of intervals found in all sets. However, even if the numbers of bins were the same, it does not imply that the cut-points selected were identical. On the contrary, they differed as well.

These characteristics of features indicate irregular distributions of data points in the input space influencing recognition of classes, as this information plays an important role in the process of interval construction in supervised discretisation such as with Fayyad and Irani, or Kononenko methods.

Discretisation models obtained for the two training sets were also used to discretise two test sets for both datasets, which resulted in sets denoted as ToL1 (Test 1 on Learn) and ToL2. Such an enforced perspective reflects the general idea of train and test, by which local information, relevant to discretisation, and contained in test sets, is ignored, treated as unknown.

The two employed unsupervised discretisation methods, namely equal width binning and equal frequency binning, required setting the input parameter of the number of bins to be constructed in all sets, and for independent processing this parameter was varied as follows:

- equal width binning for both train and test sets
  - from 2 to 10 with a step of 1;
  - from 10 to 100 with a step of 10;
  - from 100 to 1000 with a step of 100;
  - from 1000 to 10000 with a step of 1000;
- equal frequency binning
  - from 2 to 200 with a step of 1 for train sets;
  - from 2 to 90 with a step of 1 for test sets.

To obtain ToL versions for all test sets, for equal width binning all 36 discretisation models from train data were used. For equal frequency binning all 199 discrete models from train data were imposed on test data.

Together for a dataset there were obtained the following variants of discrete sets:

- train sets – 36 versions from equal width binning, 199 versions from equal frequency binning, 1 version from Fayyad and Irani method, 1 version from Kononenko approach;
- test sets – independent: 36 versions from equal width binning, 89 versions from equal frequency binning, 1 version from Fayyad and Irani method, 1 version from Kononenko approach;
- test sets – based on learn: 36 versions from equal width binning, 199 versions from equal frequency binning, 1 version from Fayyad and Irani method, 1 version from Kononenko approach.

Because of two test sets, the given numbers need to be multiplied by 2. And then the total number also doubled because of two datasets, Wharton-Johnston and London-Curwood.

One more mode for discretisation was considered – performed by combining samples from all sets in a dataset into a single one, for which intervals would be defined, and which would be split back, but after discrete representations for all values are found. Such mode would cause formulation of bins for training data to be based partially on knowledge of test data. As it would be in violation of the governing idea of train and test approach, this way of processing was rejected.

6.3. Induction of decision rules

All versions of learning sets were next subjected to rule induction process using exhaustive algorithm (implemented in RSES system), which typically leads to relatively high numbers of rules. Cardinalities of inferred rule sets ranged from just few thousands to over a hundred thousands.

The characteristics of these rules, such as lengths or supports, also varied. These elements can be used for filtering rules [45], resulting in construction of rule classifiers with enhanced performance [46], however, this aspect was not considered in the research, in order not to cloud the overall picture. Thus for all processing the complete sets of inferred rules were employed, without any hard constraints. The rule sets were used for classification of samples with standard voting as the strategy for conflict resolution.

The attributes, for which through supervised discretisation single intervals were found for representation of all their values, were not excluded from decision tables, however, as brining zero information content, they were never present in the decision rules generated for such variants of discrete data.

6.4. Performance of rule classifiers for supervised discretisation

For supervised discretisation, as two used methods returned single versions of train sets, two rule sets were induced for both datasets. The decision algorithms for London and Curwood data contained in both cases more than twice as many rules as for Wharton and Johnston. When applied for classification of learning samples, they returned the perfect accuracy, which was as expected. The results for other sets are listed in Table 3.

Table 3

Performance [%] of rule classifiers evaluated with test sets obtained from supervised discretisation executed independently for all sets (Test 1 and Test 2), and based on the models constructed from learn data (ToL1 and ToL2)

	Fayyad and Irani method			
Dataset	Test 1	ToL1	Test 2	ToL2
Wharton and Johnston	94.44	86.67	97.78	96.67
London and Curwood	88.89	93.33	88.89	95.56
	Kononenko method			
Dataset	Test 1	ToL1	Test 2	ToL2
Wharton and Johnston	97.78	83.33	96.67	94.44
London and Curwood	93.33	93.33	86.67	94.44

In the case of L-C dataset, for both discretisation approaches, application of models learned for train data to test sets worked rather to advantage of classifiers, as they returned results at the same or improved level with respect to independent discretisation. However, for W-J dataset the trend was opposite, independent discretisation of test sets caused higher classification accuracy. Both groups of results indicated differences in distributions of data points in the input space. If they were the same, changing the mode of executing discretisation would not influence the obtained performance.

6.5. Performance of rule classifiers for unsupervised equal width binning

Because of varying the number of intervals constructed for each attribute, equal width binning returned 36 variants for train sets. From all these versions decision rules were inferred, which led to construction of as many decision algorithms for both datasets. Each of these rule classifiers was applied next for labelling examples contained in:

- 36 versions of train sets;
- 36 versions for two test sets discretised independently;
- 36 versions for two test sets discretised by referring to models constructed for train data;

which led to three groups of results, as commented below.

Re-classification of all samples from the learning sets, displayed in Fig. 1, enabled to study how each decision algorithm recognised the same data, but seen through a perspective of a number of bins, different than the one from which the rules were induced. The diagonals in the charts confirm that the highest performance was obtained when the numbers of intervals matched, but the plots also show other areas of interest, especially in the range of ten intervals constructed in the input datasets.

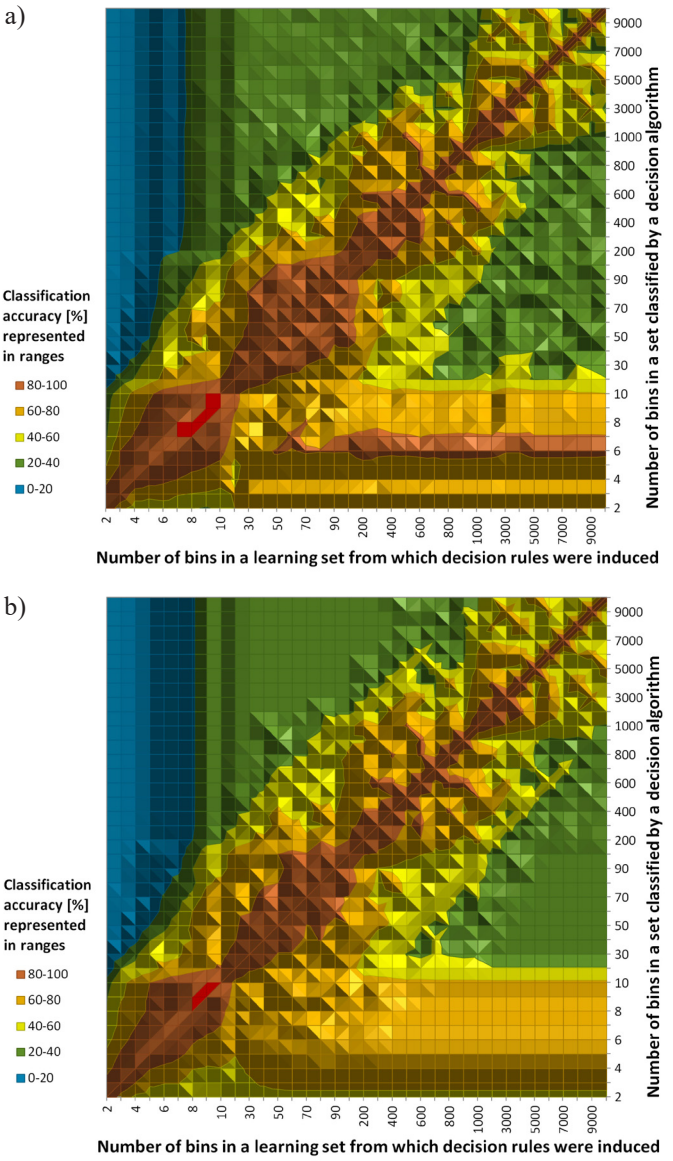

Fig. 1. Performance of rule classifiers for training sets discretised with equal width binning for: a) Wharton-Johnston data, b) London-Curwood data

The performance of rule classifiers evaluated with test sets is given in Fig. 2a for Wharton and Johnston, and in Fig. 2b for London and Curwood. In each row, the chart on the left displays classification accuracy for a test set discretised independently, and the right chart corresponds to the same test set discretised by definitions of intervals formed for the corresponding train set.

Comparison of the effects of two discretisation modes on observed performance yields the conclusion that also for this method the differences resulting from data irregularities were visible. Unlike the previously applied independent supervised approaches, equal width binning can guarantee the possibility of constructing matching numbers of bins for variables between learn and test sets. Yet still groupings of data points in test sets were different in such degree that cut-points, selected regardless of train data, not only better suited local characteristics, but also caused better predictions of rule classifiers.

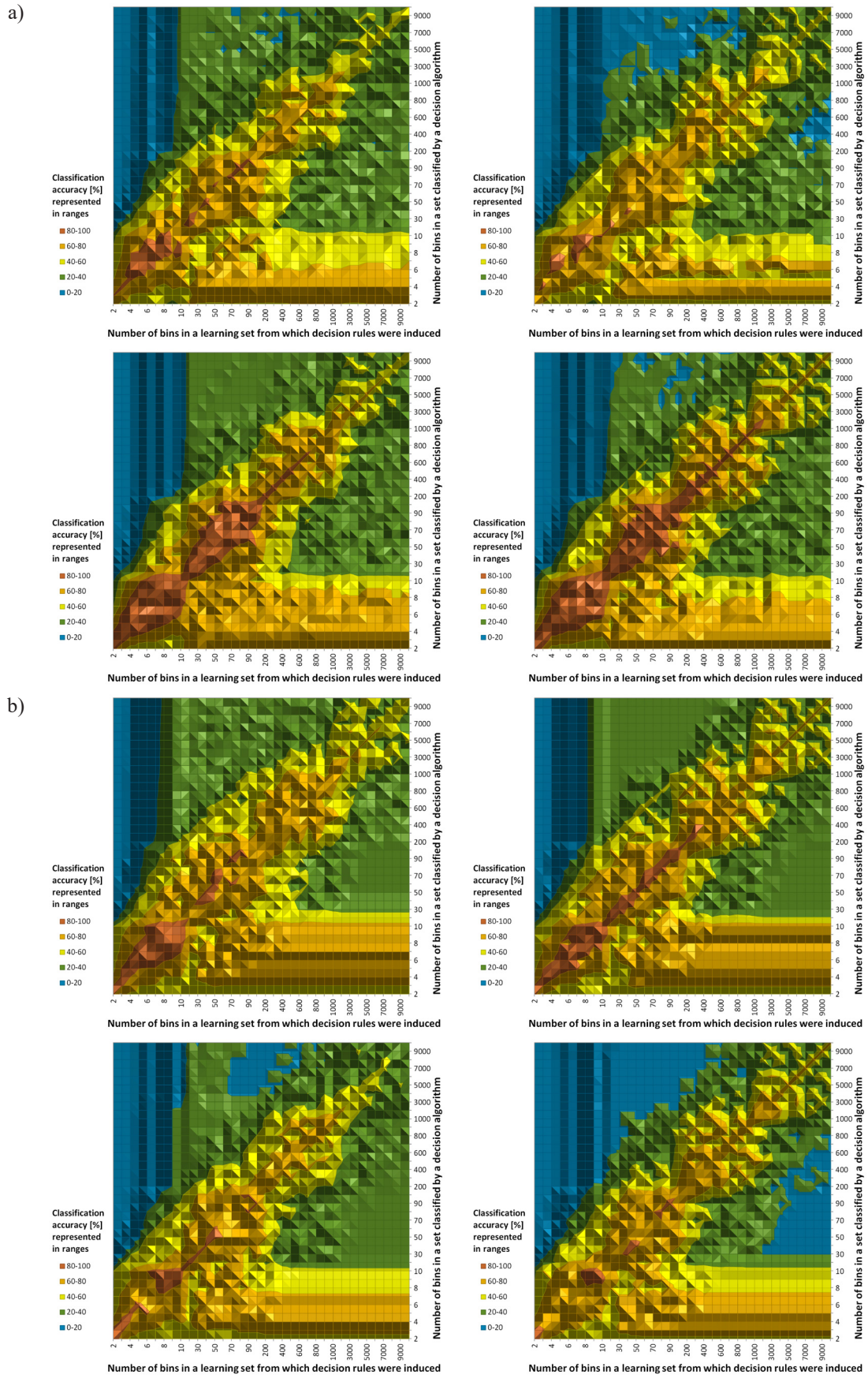

Fig. 2. Performance of rule classifiers evaluated with test sets discretised with equal width binning for: a) Wharton-Johnston dataset, b) London-Curwood dataset. On the left for independent discretisation, on the right for discretisation model obtained from train data

6.6. Performance of rule classifiers for unsupervised equal frequency binning

Equal frequency binning approach to discretisation led to construction of 199 variants for both train sets, from which decision rules were induced. Decision algorithms were used to classify samples included in:

- 199 versions of train sets;
- 89 versions for two test sets discretised independently;
- 199 versions for two test sets discretised by referring to models constructed for train data.

The difference in the numbers of versions for the discretised test sets results from different cardinalities of test and train sets. As each test set contained 90 samples, in independent discretisation only that many intervals could be requested. On the other hand, 199 discretisation models were constructed for train data, thus also that many perspectives on test data offered.

Figure 3 includes plots obtained from re-classification of learning samples by generated decision algorithms. In the whole range of tested bin numbers, the cardinalities of inferred rule sets showed almost exactly decreasing tendency, starting with over one hundred thousand of rules for just 2 bins, then falling down gradually to just few thousand for 200 intervals.

For both datasets, with increasing the number of intervals constructed for attributes, the tolerance with respect to forming a few less or a few more bins increased as well. For the numbers of intervals from close to the second half of tested values, rule classifiers recognised perfectly samples from all sets with the same or higher numbers of bins. It was caused by the close resemblance of discretised train sets, and resulting from it similarity of induced rule sets. The probability that all values of all considered attributes are unique is relatively low, and any repeated values are naturally assigned to the same bin, which means that construction of exactly as many bins as there are samples, and that was the upper limit, is highly unlikely.

Independent discretisation of test sets put the upper limit on the number of intervals at the cardinality of these sets, which was less than a half of training sets. The classification results for both test sets are shown together in one plot, on the left in Fig. 4, where the upper half is dedicated to one test set, and the bottom half to the other, in the same order displayed on the right for the discrete versions based on learning data.

The patterns visible for corresponding charts show closer similarities than for the other discretisation methods. Focus on the frequencies of occurrence of values allowed to obtain more alike definitions for intervals between train and test sets. Equal frequency binning enabled projecting input continuous space into discrete with better image of data structure expressed through existing data points. It was not possible to obtain such effect in equal width binning, as this approach requires only the calculation of minimum and maximum values to divide the range into required number of sub-ranges, regardless of positions of data points within these limits.

6.7. Summary of experiments

Advantages of supervised discretisation, as information reduction technique applied to data in the pre-processing stage, are widely acknowledged. In a case of transformations needed for datasets including several separate sets, characterised by different groupings and distributions of data points in the input continuous space, independent discretisation can cause construction of severely distinct models, with not only different cut-points, but different numbers of intervals defined for the same attributes. On the other hand, imposing discrete models discovered in train data on test sets results in ignoring their local characteristics, which can negatively influence the observed performance of inducers.

From the two unsupervised approaches tested, equal width binning in the highest degree allows to represent the uniformity of the input space, with just reduction of scale for details. Yet this methods ignores the absence or existence of most data points while forming definitions of ranges (apart from minimum and maximum), which means that it also cannot preserve their distributions and overall underlying structure. Equal frequency binning, by paying particular attention to groupings of values, comes the closest in descriptions of patterns present in data, and translating them to the discrete space.

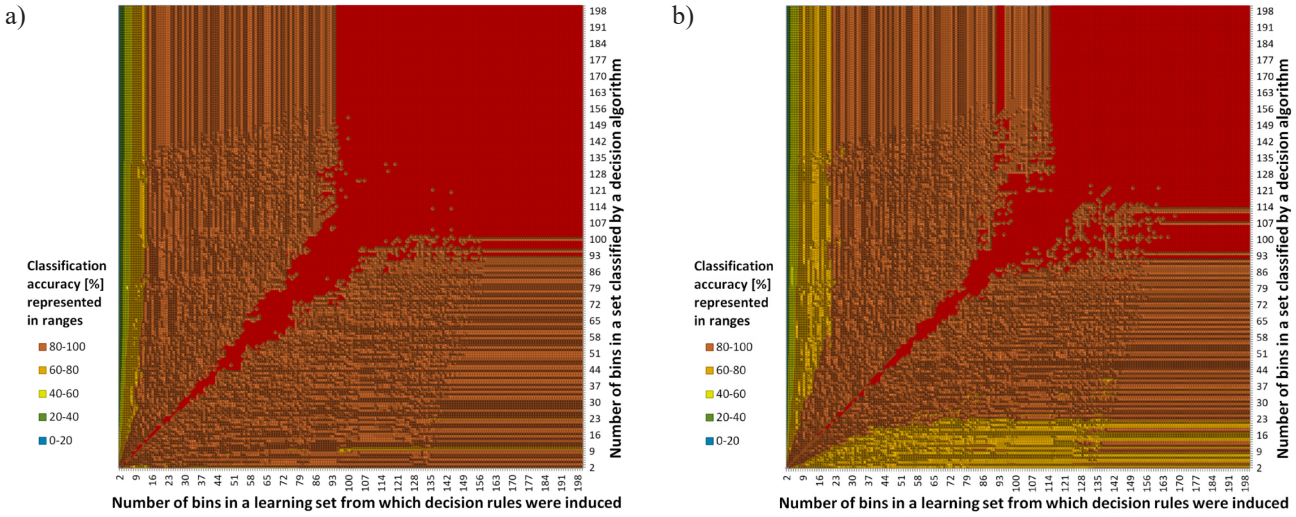

Fig. 3. Performance of rule classifiers for training sets discretised with equal frequency binning for: a) Wharton-Johnston data, b) London-Curwood data

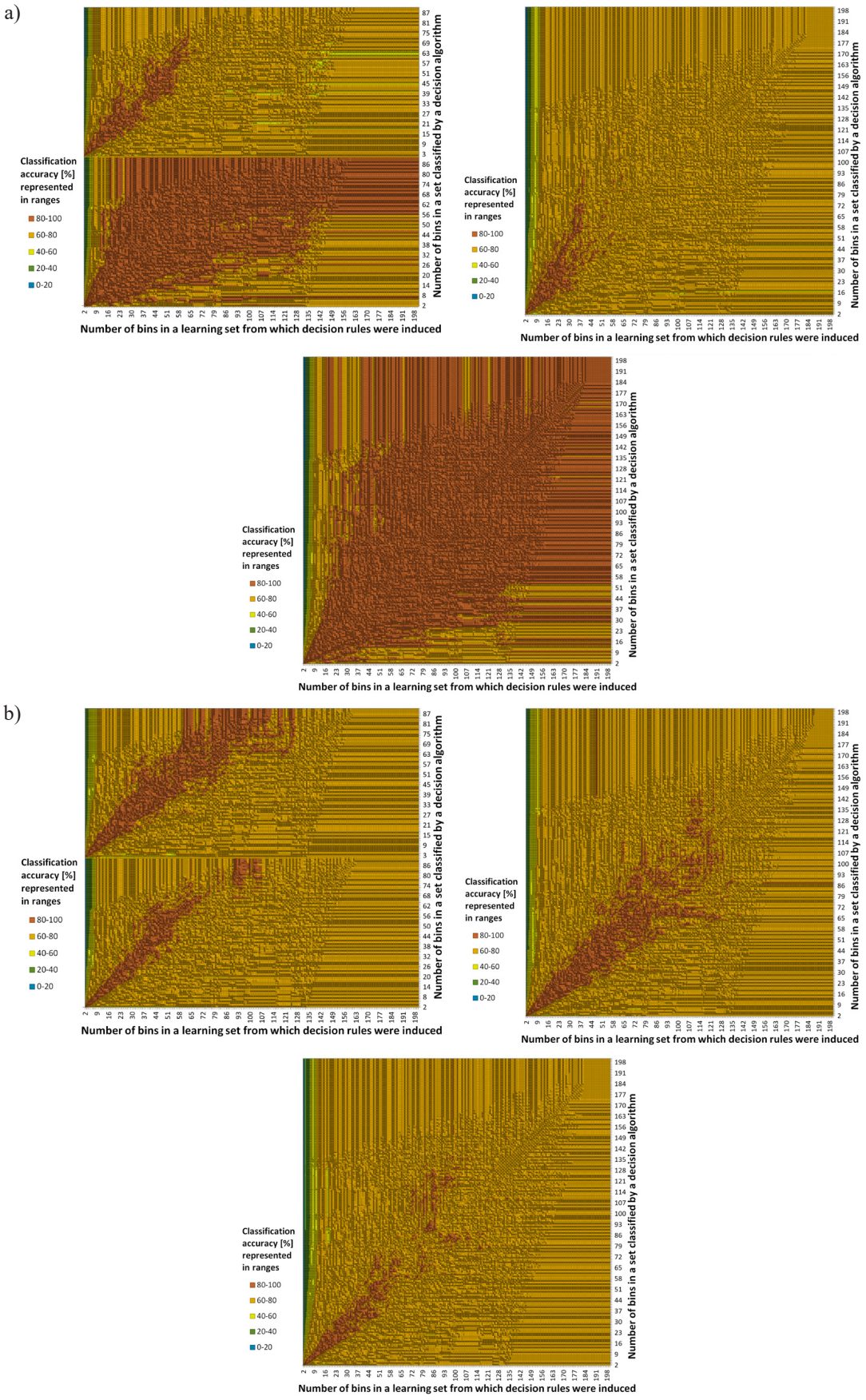

Fig. 4. Performance of rule classifiers evaluated with test sets discretised with equal frequency binning. On the left for independent discretisation, on the right for discretisation model from train data, for: a) Wharton-Johnston data, b) London-Curwood data

The test results showed that independent processing of sets, which is more convenient, leads to obtaining discrete representation allowing for satisfactory performance of classifiers, not only for exactly matching numbers of intervals between learning and test sets, but in other cases as well.

7. CONCLUSIONS

The paper presents research on irregularities in distributions of data points in the input space, observed in the context of discretisation processes, which can be performed as independent processing of learn and test sets, or based on train data. In this framework, supervised and unsupervised discretisation approaches and algorithms were examined. In the executed experiments the investigation was performed on how definitions of intervals, constructed to represent continuous values of attributes in discrete space, influence the performance of rule classifiers induced in rough set approach. The domain of application was defined by a task of authorship attribution, where stylistic profiles were described by stylometric features.

The results from experiments indicate that even in independent processing of sets, as long as the same, or close, numbers of intervals for attributes are defined, satisfactory performance of rule classifiers is obtained, and for lower numbers of bins accuracy tends to be higher. For mismatched numbers of intervals, their higher cardinality in a training set from which rules were induced still allowed to obtain acceptable classification levels for sets with fewer bins, while in the opposite case, with few intervals in the learning set and many in tested, led to much worse results.

The directions for future research are pointed out by other rule induction algorithms, for example optimised with respect to length and support, and other discretisation methods to be applied to data, eg. modification of unsupervised equal frequency binning, which as the required input parameter asks for numbers of instances that should be included in a bin. Such processing would reflect distributions of data points, thus it should simplify detection of irregularities.

ACKNOWLEDGEMENTS

The research described in the paper was performed within the statutory project of the Department of Graphics, Computer Vision and Digital Systems at the Silesian University of Technology, Gliwice (RAU-6, 2021), and at the University of Silesia in Katowice. Texts exploited in experiments are available thanks to Project Gutenberg (www.gutenberg.org).

REFERENCES

[1] G. Franzini, M. Kestemont, G. Rotari, M. Jander, J. Ochab, E. Franzini, J. Byszuk, and J. Rybicki, "Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm," *Front. Digital Humanit.*, vol. 5, p. 4, 2018, doi: 10.3389/fdigh.2018.00004.

[2] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Data level preprocessing methods," in *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018, pp. 79–121, doi: 10.1007/978-3-319-98074-4_5.

[3] S. Garcia, J. Luengo, J. Saez, V. Lopez, and F. Herrera, "A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 734–750, 2013, doi: 10.1109/TKDE.2012.35.

[4] S. Das, S. Datta, and B.B. Chaudhuri, "Handling data irregularities in classification: Foundations, trends, and future challenges," *Pattern Recognit.*, vol. 81, pp. 674–693, 2018, doi: 10.1016/j.patcog.2018.03.008.

[5] U. Stańczyk, "Evaluating importance for numbers of bins in discretised learning and test sets," in Intelligent Decision Technologies 2017: *Proceedings of the 9th KES International Conference on Intelligent Decision Technologies (KES-IDT 2017) – Part II, ser. Smart Innovation, Systems and Technologies*, I. Czarnowski, J.R. Howlett, and C.L. Jain, Eds. Springer International Publishing, 2018, vol. 72, pp. 159–169, doi: 10.1007/978-3-319-59421-7_15.

[6] G. Baron, "On approaches to discretization of datasets used for evaluation of decision systems," in *Intelligent Decision Technologies 2016*, ser. Smart Innovation, Systems and Technologies, I. Czarnowski, A. Caballero, R. Howlett, and L. Jain, Eds. Springer, 2016, vol. 56, pp. 149–159, doi: 10.1007/978-3-319-39627-9_14.

[7] U. Stańczyk and B. Zielosko, "On approaches to discretisation of stylometric data and conflict resolution in decision making," in *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES-2019, Budapest, Hungary, 4–6 September 2019*, ser. Procedia Computer Science, I. J. Rudas, J. Csirik, C. Toro, J. Botzheim, R.J. Howlett, and L.C. Jain, Eds. Elsevier, 2019, vol. 159, pp. 1811– 1820, doi: 10.1016/j.procs.2019.09.353.

[8] J. Bazan, H. Nguyen, S. Nguyen, P. Synak, and J. Wróblewski, "Rough set algorithms in classification problem," in *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*, L. Polkowski, S. Tsumoto, and T. Lin, Eds. Heidelberg: Physica-Verlag HD, 2000, pp. 49–88, doi: 10.1007/978-3-7908-1840-6_3.

[9] J. Bazan and M. Szczuka, "The rough set exploration system," in *Transactions on Rough Sets III*, ser. Lecture Notes in Computer Science, J. F. Peters and A. Skowron, Eds. Berlin, Heidelberg: Springer, 2005, vol. 3400, pp. 37–56, doi: 10.1007/11427834_2.

[10] I. Chikalov, V. Lozin, I. Lozina, M. Moshkov, H. Nguyen, A. Skowron, and B. Zielosko, *Three Approaches to Data Analysis – Test Theory, Rough Sets and Logical Analysis of Data*, ser. Intelligent Systems Reference Library. Berlin, Heidelberg: Springer, 2013, vol. 41, doi: 10.1007/978-3-642-28667-4.

[11] Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Inf. Sci.*, vol. 177, no. 1, pp. 3–27, 2007, doi: 10.1016/j.ins.2006.06.003.

[12] J. Rybicki, M. Eder, and D. Hoover, "Computational stylistics and text analysis," in *Doing Digital Humanities: Practice, Training, Research*, 1st ed., C. Crompton, R. Lane, and R. Siemens, Eds. Routledge, 2016, pp. 123–144, doi: 10.4324/9781315707860.

[13] M. Eder, "Style-markers in authorship attribution a crosslanguage study of the authorial fingerprint," *Stud. Pol. Ling.*, vol. 6, no. 1, pp. 99–114, 2011.

[14] H. Craig, "Stylistic analysis and authorship studies," in *A companion to digital humanities*, S. Schreibman, R. Siemens, and J. Unsworth, Eds. Oxford: Blackwell, 2004, doi: 10.1002/9780470999875.ch20.

[15] G. Baron, "Comparison of cross-validation and test sets approaches to evaluation of classifiers in authorship attribution domain," in *Proceedings of the 31st International Symposium on Computer and Inf. Sci.*, ser. Communications in Computer and Information Science, T. Czachórski, E. Gelenbe, K. Grochla, and R. Lent, Eds. Cracow: Springer, 2016, vol. 659, pp. 81–89, doi: 10.1007/978-3-319-47217-1_9.

[16] S.S. Mullick, S. Datta, S.G. Dhekane, and S. Das, "Appropriateness of performance indices for imbalanced data classification: An analysis," *Pattern Recognit*., vol. 102, pp. 107–197, 2020, doi: 10.1016/j.patcog.2020.107197.

[17] J.M. Johnson and T.M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 27, pp. 1–54, 2019, doi: 10.1186/s40537-019-0192-5.

[18] N. Basurto, C. Cambra, and Á. Herrero, "Improving the detection of robot anomalies by handling data irregularities," *Neurocomputing*, 2020, doi: 10.1016/j.neucom.2020.05.101, in press.

[19] G. Shi, C. Feng,W. Xu, L. Liao, and H. Huang, "Penalized multiple distribution selection method for imbalanced data classification," *Knowledge-Based Syst*., vol. 196, p. 105833, 2020, doi: 10.1016/j.knosys.2020.105833.

[20] S. Au, R. Duan, S.G. Hesar, and W. Jiang, "A framework of irregularity enlightenment for data pre-processing in data mining," *Ann. Oper. Res*., vol. 174, no. 1, pp. 47–66, 2010, doi: 10.1007/s10479-008-0494-z.

[21] M. Koziarski, M. Wozniak, and B. Krawczyk, "Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise," *Knowledge-Based Syst*., vol. 204, p. 106223, 2020, doi: 10.1016/j.knosys.2020.106223.

[22] N. Basurto, Á. Arroyo, C. Cambra, and Á. Herrero, "Imputation of missing values affecting the software performance of component-based robots," *Comput. Electr. Eng*., vol. 87, p. 106766, 2020, doi: 10.1016/j.compeleceng.2020.106766.

[23] S. Argamon, K. Burns, and S. Dubnov, Eds., *The structure of style: Algorithmic approaches to understanding manner and meaning*. Berlin: Springer, 2010, doi: 10.1007/978-3-642-12337-5.

[24] S. Sbalchiero and M. Eder, "Topic modeling, long texts and the best number of topics. some problems and solutions," *Qual. Quant*., vol. 54, pp. 1095–1108, 2020, doi: 10.1007/s11135-020-00976-w.

[25] R. Peng and H. Hengartner, "Quantitative analysis of literary styles," Am. Statistician, vol. 56, no. 3, pp. 15–38, 2002, doi: 10.1198/000313002100.

[26] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Am. Soc. Inf. Sci. Technol*., vol. 60, no. 3, pp. 538–556, 2009, doi: 10.1002/asi.21001.

[27] D. Khmelev and F. Tweedie, "Using Markov chains for identification of writers," *Lit. Linguist. Comput*., vol. 16, no. 4, pp. 299–307, 2001, doi: 10.1093/llc/16.3.299.

[28] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *J. Am. Soc. Inf. Sci. Technol*., vol. 60, no. 1, pp. 9–26, 2009, doi: 10.1002/asi.20961.

[29] M. Jockers and D. Witten, "A comparative study of machine learning methods for authorship attribution," *Lit. Linguist. Comput*., vol. 25, no. 2, pp. 215–223, 2010, doi: 10.1093/llc/fqq001.

[30] M. Eder and J. Rybicki, "Do birds of a feather really flock together, or how to choose training samples for authorship attribution," *Lit. Linguist. Comput*., vol. 28, pp. 229–236, 8 2013, doi: 10.1093/llc/fqs036.

[31] M. Eder, "Does size matter? Authorship attribution, small samples, big problem," Digital Scholarsh. Humanit., vol. 30, pp. 167–182, 06 2015, doi: 10.1093/llc/fqt066.

[32] K. Kalaivani and S. Kuppuswami, "Exploring the use of syntactic dependency features for document-level sentiment classification," *Bull. Pol. Acad. Sci. Tech. Sci*., vol. 67, no. 2, pp. 339–347, 2019, doi: 10.24425/bpas.2019.128608.

[33] G. Rotari, M. Jander, and J. Rybicki, "The Grimm brothers: A stylometric network analysis," Digital Scholarsh. Humanit., 02 2020, doi: 10.1093/llc/fqz088.

[34] C. Jankowski, D. Reda, M. Mańkowski, and G. Borowik, "Discretization of data using Boolean transformations and information theory based evaluation criteria," *Bull. Pol. Acad. Sci. Tech. Sci*., vol. 63, no. 4, pp. 923–932, 2015, doi: 10.1515/bpasts-2015-0105.

[35] U. Fayyad and K. Irani, "Multi-interval discretization of continuous valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, vol. 2. Morgan Kaufmann Publishers, 1993, pp. 1022–1027.

[36] I. Kononenko, "On biases in estimating multi-valued attributes," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAI'95*, vol. 2. Morgan Kaufmann Publishers Inc., 1995, pp. 1034–1040.

[37] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978, doi: 10.1016/0005-1098(78)90005-5.

[38] S. Kotsiantis and D. Kanellopoulos, "Discretization techniques: A recent survey," *GESTS Int. Trans. Comput. Sci. Eng*., vol. 32, no. 1, pp. 47–58, 2006.

[39] B. Zielosko, "Application of dynamic programming approach to optimization of association rules relative to coverage and length," *Fundamenta Informaticae*, vol. 148, no. 1-2, pp. 87–105, 2016, doi: 10.3233/FI-2016-1424.

[40] S.G. Weidman and J. O'Sullivan, "The limits of distinctive words: Re-evaluating literature's gender marker debate," *Digital Scholarsh. Humanit*., vol. 33, pp. 374–390, 2018, doi: 10.1093/llc/fqx017.