Information Sciences 622 (2023) 357–373

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Effect of inconsistency rate of granulated datasets on classification performance: An experimental approach ☆

ChienHsing Wu

Department of Information Management, National University of Kaohsiung, 700, Kaohsiung University Rd., Nanzih District, 81148 Kaohsiung, Taiwan

ARTICLE INFO

Article history:
Received 1 January 2022
Received in revised form 9 November 2022
Accepted 27 November 2022
Available online 5 December 2022

Keywords:
Knowledge discovery
Granulation
Data inconsistency
Prediction accuracy

ABSTRACT

An experiment was conducted to investigate the effect of the inconsistency rate (IR) of granulated datasets on classification performance. Unsupervised (equal-width interval, EWI) and supervised (minimum description length, MDL) techniques were used to granulate 36 datasets. An algorithm was developed to divide the original granulated datasets into consistent and inconsistent subsets. Five classifiers including one simple tree-based and four ensemble-based on datasets before granulation (BG), after granulation but before removal of inconsistent granulated datasets (AGBR), and after removal of inconsistent granulated datasets (AR) were used, followed by testing and comparisons of predication accuracy (PA). The experimental results showed the following: (1) 24 out of 36 via EWI and 28 out of 36 via MDL datasets contain inconsistent datasets. (2) PA of AR is more likely higher than of BG and AGBR datasets with both EWI and MDL by all classifiers. (3) Mean PA improvement ranges from 5.74% to 10.01% with EWI and from 8.74% to 13.73% with MDL. (4) The correlation coefficient between IR and PA improvement ranges from 0.7413 to 0.7901 with EWI and 0.7870 to 0.9683 with MDL. These results demonstrate the value of uncovering the effect of IR on classification performance in the domain of machine learning.

1. Background

The process of machine learning involves various stages, such as innovative technology ideas, data collection, data cleaning, data pre-processing, development of learning mechanisms, testing, adjustment, and refinement [1,9,25,26,32,34]. The suitability of techniques and reliability of results from a stage may therefore influence the following steps and the overall performance.

Previous studies have extensively discussed the contribution of each stage to overall classification performance. For example, a review of innovative components, tools, and technologies identified future research directions related to the processing of large amounts of data [25]. The issue of label noise has also been reviewed and discussed to benefit classification feasibility and performance [9,16–18,30,34]. For example, risk minimization approaches were reported showing classification accuracy improvement by considering the surrogate loss function and using weighted surrogate loss to reduce risk minimization under noisy labels [17]. The Markov chain sampling algorithm was introduced to enhance classification effectiveness by computing mislabeling probability for each instance [34]. The effects of different degrees of noise were com-

Peer review under responsibility of For special issue article please include a footnote that this paper belongs to the special issue "special issue name" edited by "editor name".

E-mail address: chwu@nuk.edu.tw

https://doi.org/10.1016/j.ins.2022.11.135
0020-0255/

pared among supervised classifiers, revealing that noise class in datasets leads to unreliable outcomes [18]. Compression rate of the compression model considering adversarial-aware was used to examine model robustness in comparison with compact models for convolutional neural networks models [30]. The findings highlight that regularization methods, such as dropout, make models more robust to input perturbations. This implies that the effect of noisy on classification performance should be evaluated for both model robustness and classification accuracy.

In line with limitations imposed by data types on classification models, data pre-processing is a crucial step. It works in parallel with analysis and learning mechanisms to convert or aggregate data, and consists of methods such as granulation of continuous data, supervised and unsupervised conversion strategies [25], and experienced and expert-oriented conversion methods [9]. Previous studies have also discussed the merits of pre-processing in various domains. For example, Data quality measures based on different granular approaches were introduced to estimate levels of difficulty in extracting knowledge from datasets [26]. The neighborhood rough set theory used in feature selection produces a better classification performance than four baseline algorithms [4]. Application of Chi-square granulation algorithm was proposed as a supervised approach to group classification methods and resulted in better predictions [21]. A self-adjustment method to reduce the negative fluctuations in the discretization interval was also proposed [24] with superior prediction accuracy. Based on dataset characteristics, the selection of outlier detection and normalization scheme helps obtain the most likely best performance [12].

Particularly, an adaptive granulation algorithm for community detection with improved label propagation has indicated the potential value of supergranulation strategies [6]. Data reliability has also received increasing attention, with 4 V features (volume, velocity, variety, variability) being argued to increase the value of data [1,24,25]. In this context, data cleaning can be particularly sensitive to missing data, noisy classes, and imbalanced classes [4,8,18,30]. To overcome this issue, missing data are either replaced with generated data using the similarity approach [14], or excluded when data volume is large enough or the class distribution remains similar. However, how noisy classes influences learning performance still needs to be disclosed in more appropriateness.

Amongst existing procedures is the granulation of continuous data. Granulation quality is known to influence the final classification performance [20,27,29,31]. Continuous data granulation may be dynamic or static, global or local, and supervised or unsupervised [5,20,25]. The dynamic approach considers the number of attributes and their independence, determines the number of bins (or granules), and then granulates the data [20]. The static approach regards all attributes as independent, and therefore the number of bins for each attribute is also independent [7]. The supervised approach determines the number of intervals and processing rules settings before granulation based on known classes, whereas the unsupervised approach does not impose restrictions and is fully data-oriented [5,21]. These two approaches have been widely used in classification-oriented knowledge discovery, mostly because they provide for granulation efficiency while maintaining data independence and reliability.

Granulation approaches have been widely applied with increasing efficiency in diverse contexts [6,19,21,24,26,27]. However, granulation often produces inconsistency when cases (or samples) in a granulated dataset have the same decisional attribute values but different labels (or classes, categories) [17,18,31]. The output of continuous data granulation often produces subsets with the same decision attribute values but different (or inconsistent) classification categories. These subsets from a granulated dataset may result in three drawbacks. The first is the difficulty of returning reliable knowledge in the form of rules, which limits practical uses, the second is the disturbance in reliably performing training ability for classifiers, and the third is the insufficiency in producing reliable testing results for a program of machine learning.

Although various data pre-processing methods, such as principle of uncertainty, polynomial features, standard scaler and MinMax scaler have been reviewed and discussed [25], the influence of inconsistent granulated data on predictive attributes of performance still needs to be uncovered. For example, one of the subsets of a granulated dataset may have one decision attribute value but two predictive values, as in "*if* citizen age is over 60, smokes moderately, lives in downtown, wears a mask moderately, *then* the risk of Covid-19 infection may be either mild (denoted by subset-1) or severe (denoted by subset-2)". In such cases, it may be prone to producing classification failure and be difficult to generate reliable learning outcomes.

In the above example, assume there were 30 mild cases (subset-1) and 300 severe cases (subset-2) under the same decision attribute values. The relative frequency of subset-1 is 30/(30 + 300) = 0.0909, whereas subset-2 represents 300/(30 + 300) = 0.9091 of cases and is the dominating subset. Theoretically, under a 70–30 % testing scenario, there will be a 9.09 % probability of failed classification, and a 30/99 = 30.30 % probability if the testing dataset (99 cases) is stochastically sampled entirely from subset-1. To deal with this issue, an evidence-oriented scenario may be applied by selecting cases from the largest class to represent the final datasets being used.

Although granulation may produce inconsistent data [31]), its effect on learning performance and whether this differs at various inconsistency levels is still open to debate. Accordingly, the research conducted an experimental analysis with two main objectives. The first was to propose a mechanism to define inconsistent and consistent subsets from an original granulated dataset [31]. The second was to examine the effects of inconsistent granulated data with various inconsistency acceptance levels on learning performance. Experiments were conducted using 36 datasets (28 continuous and 8 mixed) retrieved from public data repositories [28]. Both unsupervised [5,31] and supervised [7,11] granulation techniques were used, and five classifiers containing one tree-based classifier (Iterative dichotomiser 3, ID3) [23] and four ensemble-based learning models, namely, Support vector machine (SVM) [13,22], a bagging-based classifier (Random forest, RF) [3], a boosting-based classifier (AdaBoost) [10], and an updatable multiclass classifier with stochastic gradient decent (UMC-SGD), were utilized [2]. The predication accuracy results of the predication models including the original datasets (before granulation) and

the granulated datasets before removal and after removal of inconsistent datasets were compared. The findings demonstrate how inconsistent granulated data influence learning performance and thereby contribute an alternative approach to improving the performance of machine learning.

2. Related concepts

2.1. Knowledge discovery in databases

The merits of classification-oriented learning models are determined by a series of consecutive operations relating to the quality of knowledge resources (owner of knowledge in the form of numerical data, text, video, sound, pictures, or animation), quality of pre-processing, quality of classification mechanism, and output evaluation criteria. The results from each phase may affect the result in the next stages. Data pre-processing is particularly relevant for the overall success of classification. For example, missing or uncertain data, and datasets with unbalanced classes [4,8] may be problematic, and thus transforming data attributes (e.g., granulation of continuous data) [21,25,27], and integrating attributes are often applied to solve those issues.

The inconsistency diagnosis model from unsupervised granulation techniques has shown that a certain proportion of inconsistent data is left after granulation [31]. However, the effect of the inconsistent granulated data on classification performance remains unknown. Combined filters, wrappers, and embedded functions were proposed to link supervised and unsupervised granulation techniques to classification models [27], and reported support vector machine and minimal description length (MDL) techniques produced the best classifications. Nevertheless, neither the issue of inconsistent granulated data nor the influence on classification performance have been properly addressed.

Empirically, sentiment analysis of community opinions was reviewed to reveal research gaps and suggest research directions [32]. To extract sidewalk traffic characteristics (density, circulation, speed), combined spatial and temporal data discretization was proposed [19], resulting in higher model stability. Both interval-typed time mode (dynamic) and continuous-typed time mode (consecutive) as pre-processing methods were adopted to develop Bayesian network model analyzing chronic obstructive pulmonary disease [15], with overall experimental results showing that the dynamic mode was superior to the continuous one. The user-oriented clustering method to reconstruct user-item network and increase structure density [33] was also successfully applied to three datasets.

A few conclusions are derived from the above. First, granulation of continuous data is necessary for many classification-oriented learning models. Second, the utilized supervised or unsupervised data pre-processing techniques describe granulation approaches in general, but often fail to discuss their effect of granulation output on class noise, as well as learning performance. In the data pre-processing stage, low granulation quality due to large volumes of inconsistent granulated data are likely to influence classification success. Finally, inconsistent granulated data have not been discussed in the required level of detail. In summary, inconsistent granulated data should be addressed appropriately to ensure that the pre-processed datasets are adequate and their effect on final classification performance still needs to be properly explored.

2.2. Inconsistency of granulated data

The unsupervised granulation method is simple and convenient because it is based on the original data. It has two typical algorithms: equal width interval (EWI) and equal frequency interval They involve dividing the observed values of a continuous attribute into k equally sized granules (EWI) or equally volume granules, where k is the number of granules defined by users. When EWI and equal frequency interval are used, nearly 40 % of the granulation of continuous datasets contains inconsistent data, and almost 22 % of the datasets had more than 20 % inconsistent data which may influence classification performance [31]. Chi-square discretization algorithms to group classification methods have been successfully applied [21]. Moreover, MDL has been used to compress data using inductive inference [11] based on the principle that the shortest description of data results in the best model. However, whether supervised granulation techniques generate inconsistent data and their consequent effect on classification performance still needs to be disclosed.

The level of inconsistency due to granulation of continuous data depends on the number of attributes, the value space of decision attributes, the value space of predictive attribute, and the dataset size. To deepen the concept, Table 1 represents a granulated dataset with 19 cases, seven decision attributes (Att-A, Att-B, Att-C, etc.) and a predictive attribute. It is divided into five subsets S1, S2, S3, S4, and S5., of which S1 is consistent subset and S2, S3, S4, and S5 are inconsistent ones. For example, S1 contains five cases with same attribute values V-A1, V-B2, V-C2, V-D4, V-E1, V-F3, and V-G5, and their class value is CLS2. S2 contains two cases with the same attribute values of V-A1, V-B2, V-C2, V-D4, V-E1, V-F4, and V-G4, but different class values (CLS2 and CLS4), representing inconsistent granulated data.

When inconsistent data is used in learning classification, there will be a reliability problem when deriving decision rules. For example, two rules are produced from the three cases in S3, with reliabilities CLS1 = 0.33 (1/3), and CLS4 = 0.67 (2/3). The classification results, number of supports for each generated rule, and inconsistency are presented in Table 2. The final granulated dataset being considered in the classification stage depends on the inconsistency filtering criteria, which may reveal various effects on both classification performance evaluation and interpretation of the discovered outcomes.

Table 1
Sample of inconsistent granulated datasets.

Subset	No.	Att-A	Att-B	Att-C	Att-D	Att-E	Att-F	Att-G	Class
S1	1	V-A1	V-B2	V-C2	V-D4	V-E1	V-F3	V-G5	CLS2
	2	V-A1	V-B2	V-C2	V-D4	V-E1	V-F3	V-G5	CLS2
	3	V-A1	V-B2	V-C2	V-D4	V-E1	V-F3	V-G5	CLS2
	4	V-A1	V-B2	V-C2	V-D4	V-E1	V-F3	V-G5	CLS2
	5	V-A1	V-B2	V-C2	V-D4	V-E1	V-F3	V-G5	CLS2
S2	6	V-A1	V-B2	V-C2	V-D4	V-E1	V-F4	V-G4	CLS2
	7	V-A1	V-B2	V-C2	V-D4	V-E1	V-F4	V-G4	CLS4
S3	8	V-A3	V-B1	V-C3	V-D5	V-E4	V-F1	V-G4	CLS1
	9	V-A3	V-B1	V-C3	V-D5	V-E4	V-F1	V-G4	CLS4
	10	V-A3	V-B1	V-C3	V-D5	V-E4	V-F1	V-G4	CLS4
S4	11	V-A2	V-B3	V-C1	V-D2	V-E2	V-F2	V-G1	CLS2
	12	V-A2	V-B3	V-C1	V-D2	V-E2	V-F2	V-G1	CLS3
	13	V-A2	V-B3	V-C1	V-D2	V-E2	V-F2	V-G1	CLS3
	14	V-A2	V-B3	V-C1	V-D2	V-E2	V-F2	V-G1	CLS3
S5	15	V-A5	V-B4	V-C3	V-D2	V-E3	V-F2	V-G2	CLS4
	16	V-A5	V-B4	V-C3	V-D2	V-E3	V-F2	V-G2	CLS5
	17	V-A5	V-B4	V-C3	V-D2	V-E3	V-F2	V-G2	CLS5
	18	V-A5	V-B4	V-C3	V-D2	V-E3	V-F2	V-G2	CLS5
	19	V-A5	V-B4	V-C3	V-D2	V-E3	V-F2	V-G2	CLS5

Table 2
Classification results, supports, and inconsistency.

Subset	Att-A	Att-B	Att-C	Att-D	Att-E	Att-F	Att-G	Class	Support	Inconsistency rate
S1	V-A1 → V-B2 → V-C2 → V-D4 → V-E1 → V-F3 → VG5							CLS2	5	0/5 =0.00 %
S2	V-A1 → V-B2 → V-C2 → V-D4 → V-E1 → V-F4 → VG4							CLS2	1	1/2
								CLS4	1	=50.00 %
S3	V-A3 → V-B1 → V-C3 → V-D5 → V-E4 → V-F1 → VG4							CLS1	1	1/3
								CLS4	2	=33.33 %
S4	V-A2 → V-B3 → V-C1 → V-D2 → V-E2 → V-F2 → VG1							CLS2	1	1/4
								CLS3	3	=25.00 %
S5	V-A5 → V-B4 → V-C3 → V-D2 → V-E3 → V-F2 → VG2							CLS4	1	1/5
								CLS5	4	=20.00 %

The final granulated dataset from Table 2 differs in inconsistency filtering criteria. For example, when excluding all subsets with inconsistent data (only accepting 100 % consistent data), only S1 (five samples) remains. In this case, inconsistency rate IR= (19–5)/19 = 73.68 %. When excluding subsets with level of inconsistency greater than 30 %, S2 and S3 are removed, meaning that five cases in S1, three in S4, and four in S5 remain (12 samples in final), resulting in IR=(19–12)/19 = 36.84 %.

2.3. Classification-oriented learning models

Various supervised classification models have been widely proposed and used [13,22,23,25], including simple tree-based approaches such as ID3 and C4.5 [23] and ensemble-based approaches such as SVM [13,22], RF [3], and AdaBoost [10]. Based on information theory, ID3 and C4.5 determine the attribute separation ability to generate a decision tree. Based on a linear form, SVM generates hyperplanes that separate samples with a maximum margin to determine the class that a sample is grouped. Extended from the bagging approach, RF utilizes bagging and feature randomness by decreasing the variance to create an uncorrelated forest of decision trees. AdaBoost uses an iterative approach to learn from the mistakes of weak classifiers and turn them into strong ones. The updatable multiclass classifier using stochastic gradient descent (SGD) is another choice to reach global optimality, particularly when efficiency is considered.

To be general, the experimental design considers supervised (EWI) and unsupervised (MDL) granulation techniques on dataset variety including the number of datasets used, the number of dataset attributes, the types of data (continuous or mixed), the number of classes, and the granulation techniques (supervised and unsupervised). For comparison, the selected classifiers covering various approaches are used for datasets at the stages of before granulation, after granulation but before removal, and after the removal of inconsistent granulated datasets.

3. Method

3.1. Design framework and design features

The methodology is based on a three-phase approach consisting of data collection and pre-processing (granulation), data separation (consistent or inconsistent), and training, testing, and comparison (Fig. 1).

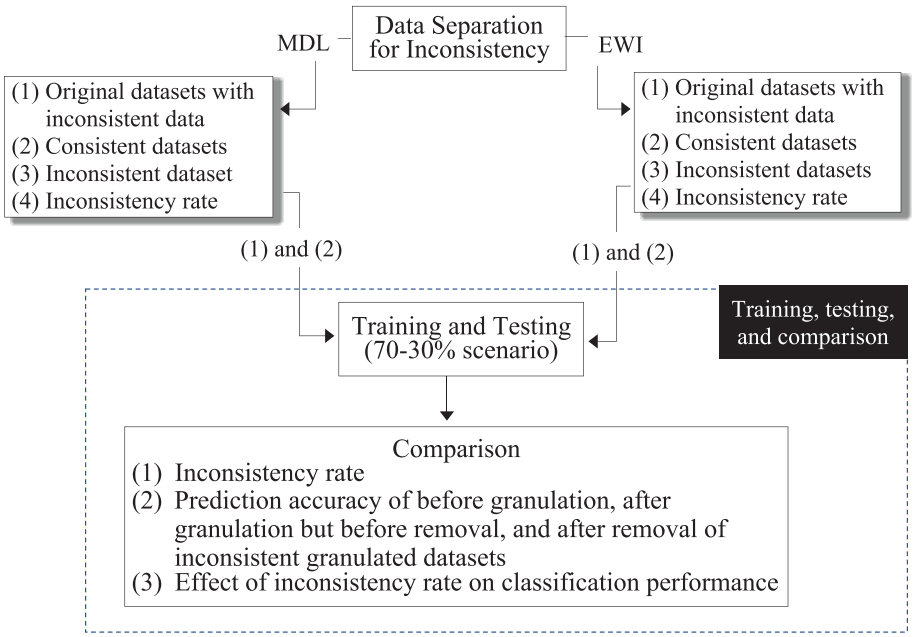

Fig. 1. Design framework.

A total of 36 datasets were granulated using unsupervised (EWI) and supervised (Minimum Description Length, MDL) techniques. An algorithm divided the granulated dataset into consistent and inconsistent subsets according to five levels of consistency (L1: 60 %, L2: 70 %, L3: 80 %, L4: 90 %, and L5: 100 %). For example, the L1 criterion required the proportion of consistent cases to be over or equal to 60 % to be considered consistent, with the remaining 40 % of cases being removed from the original granulated dataset. In the training and testing stages, original ungranulated datasets, original granulated datasets with inconsistent cases, and consistent datasets extracted with different levels of consistency were trained and tested. The 70 %–30 % scenario was applied, where five selected classifiers were used to derive prediction models from 70 % of samples randomly selected from a dataset. The remaining 30 % samples were used to test the trained models. In the comparison phase, the PAs of the original datasets (before granulation), datasets after granulation but before the removal of inconsistent datasets, and after the removal of inconsistent datasets were compared to determine the effect of IR on classification performance. The design features are presented in Table 3.

3.2. Datasets and granulation

The experiment collected 28 continuous and 8 mixed datasets from public data repositories. Table 4 presents information on dataset code, size without missing data, number of attributes (discrete and continuous), and number of classes (categorical and continuous). Data size is from 150 to 45,223, number of attributes is from four to 36, number of categorical classes is from 2 to 29, and five datasets contain continuous classes.

The techniques used to granulate continuous data for decision and predictive attributes (class attributes) are EWI and MDL. EWI was applied to both continuous attributes and classes for all 36 datasets. To determine the number of granules (denoted by k), we adopted the histogram development technique in which the smallest k is used where $2^k > n$, the dataset size. For the five continuous datasets (D03, D09, D14, D16, D26), EWI was used to granulated continuous class attributes and MDL was used to granulate continuous decision attributes.

3.3. Data separation

A data separation algorithm (DSA) was developed in two steps (by Step-1 and Step-2) to derive the consistent and inconsistent datasets according to selected levels of consistency. Step-1 extracts an entirely consistent dataset (ECD) from the original dataset, while step-2 returns consistent data from the remaining dataset back into the ECD according to the selected level of consistency acceptance. The final ECD contains singular and multiple cases with same attribute values and same class value. The DSA is presented below.

Begin

Let D be the granulated dataset;

Let D_{InCONS} be an empty data table with the same attributes as D;

Table 3
Design features.

Features	Description
Objectives	(1) Separate inconsistent data from original granulated dataset. Compare prediction accuracy of consistent and original dataset to estimate effect of IR on classification performance.
Datasets	(1) Two bags of 36 datasets from public data repository. 28 continuous, 8 mixed, and five with continuous data type in class.
Pre-processing	(1) Remove missing data. Granulate datasets using EWI (unsupervised) for one bag and MDL (supervised) for another. Granulate continuous data type of class using EWI for five datasets to apply MDL.
Separation algorithm	(1) Define five consistence acceptance levels at 60 % (L1), 70 % (L2), 80 % (L3), 90 % (L4), and 100 % (L5). Employ data separation algorithm using five consistency acceptance levels to divide an original granulated dataset with inconsistent dataset into consistent and inconsistent subset. Datasets having consistent cases are used for further analysis.
Classification model	(1) Employ ID3 (simple tree-based) and four ensemble-based classifiers: SVM, random forest (bagging-based), AdaBoost (boosting-based), and multiclass classifier with stochastic gradient decent as the training models.
Training and testing	(1) Adopt 70 %-30 % testing scenario. Use datasets before granulation, original granulated datasets, and their consistent granulated datasets after five-level removal of inconsistency. Obtain the best predication accuracy among five levels using ID3 and four ensemble-based classifiers
Comparison of effect	(1) Compare prediction accuracy of before-granulated datasets, and original granulated and consistent datasets by the five levels. Estimate effect of IR on classification performance.

Table 4
Datasets used.

Code	Name	Size	Number of attributes (discrete + continuous)	Number of classes
D01	Audit-risk	772	17	2
D02	Abolone	4177	8 (1 + 7)	29
D03	Airfoil	1503	5	Continuous
D04	Audlt	45,223	13 (7 + 6)	41
D05	Avila	20,867	10	12
D06	BankNotes	1372	4	2
D07	Bc569	569	30	2
D08	Bc683	683	9	2
D09	BikeShare	17,379	13	Continuous
D10	Bupa	345	6	2
D11	ContrMeCho	1473	9 (7 + 2)	3
D12	Credit	653	15 (9 + 6)	2
D13	DryBean	13,611	16	7
D14	Energy-effiC	768	8	Continuous
D15	EyeState	14,980	14	2
D16	ForestFire	517	12 (2 + 10)	Continuous
D17	Glass	214	9	7
D18	HeartFailure	299	12 (5 + 7)	2
D19	HTRU2	17,898	8	2
D20	ILPD	579	10 (1 + 9)	2
D21	Iris	150	4	3
D22	Letter	20,000	16	26
D23	Liver	589	12 (1 + 11)	4
D24	Page	5473	10	5
D25	Pend	3498	16	10
D26	QSAR	546	8	Continuous
D27	Satle	2000	36	6
D28	Segm	2310	19	7
D29	Shut	14,500	9	7
D30	Vehi	846	18	4
D31	Vowe	990	10	11
D32	Wave	5000	21	3
D33	Wilt	4839	5	2
D34	Wine	178	13	3
D35	WineRedQly	1599	11	6
D36	WineWhtQly	4898	11	7

//D_{InCONS} stores inconsistent data
Let D = Sorting D by all attributes in descending order;
Let D = Grouping tuples from D with same attribute value but different class value;

```
Let D_CONS = singular tuple and multiple tuples with same attribute values and same class value;
                //D_CONS stores consistent data
Let n be the number of groups;
  If n > 0
    Read consistency acceptance level DoC;
                //DoC is the defined level of consistency acceptance
    Do while n ≠ 0
      Let Gn be the size of the n^th group;
                //Gn is the number of samples in n^th group
      Let Mj be the size of majority by classes in the n^th group;
                  //Mj is the number of tuple in the major data subset by class
      If (Mj/Gn) ≥ DoC;
        Append majority tuples of the n^th group to D_CONS;
                //Add all majority tuples to consistent dataset D_CONS
        Append remaining tuples of n^th group to D_InCONS;
                //Add remaining tuples to inconsistent dataset D_InCONS
        Endif;
        n = n-1;
      Enddo;
    Endif; //No inconsistent data
End
```

3.4. Training and testing

Training and testing are applied to datasets with EWI and MDL, containing datasets before granulation (BG), after granulation before removal of inconsistent dataset (AGBR), and after removal (AR). A granulated dataset is randomly divided into two subsets. One subset contains 70 % of cases used in training, and the other contains the remaining 30 % cases for testing. The ID3 (tree-based) and four ensemble-based classifiers (SVM, RF, AdaBoost with bagging, and UMC-SGD) are used to derive the classification outcomes. The testing criterion is PA regarding 30 % of the dataset based on the outputs from 70 % of the dataset using the five classification models selected. The comparison among classifiers on BG, AGBR, and AR is based on the best PA from five inconsistency acceptance levels. The best outcome from various SVM kernels (normalized poly, poly, Puk, and radial basis function) and five inconsistency acceptance levels for each dataset is the final PA by SVM.

4. Results

4.1. Inconsistency rate

The experimental results shown in Table 5 presents IR from EWI and MDL for five consistency acceptance levels (L1 to L5). There are 24 datasets that contain inconsistent data from EWI and 28 from MDL for all levels, including 66.67 % and 83.33 % inconsistent datasets, respectively. Of the 24 inconsistent EWI datasets, nine produced IR over 10 % for L1, L2 and L3, 11 for L4, and 13 for L5, and 37.5 % inconsistent datasets for L1, L2, and L3, 45.83 % for L4, and 54.16 % for L5. Of the 28 inconsistent MDL datasets, 12 produced IR over 10 % for L1 and L2, 14 for L3, 16 for L4, and 20 for L5, with 40 % inconsistent datasets for L1 and L2, 46.67 % for L3, 53.33 % for L4, and 66.67 % for L5.

IR likely increases with consistence acceptance level, which is expected since higher acceptance levels become increasingly more exclusive (Figs. 2 and 3). IR from EWI and MDL across the five acceptance levels vary in dataset size, number of attributes, and number of classes. This implies that inconsistent data from EWI and MDL cannot be predicted in terms of dataset size, number of attributes, and number of classes, or their combinations. This echoes the argument that classification performance depends on the characteristics of a dataset, and not solely on granulation approaches, learning algorithms, and testing scenarios [26,31]. Finally, IR is extremely high in the D15 dataset from EWI (over 99 %) and in the D02, D03, and D26 datasets from MDL (over 87 %), both based on L1. These datasets may contain insufficient data for the stages of training and testing. For example, 14,947 out of 14,980 samples in D15 from EWI are removed, resulting in only 43 samples available in the dataset. Therefore, datasets with IR over 80 % based on L1 were excluded from the stages of prediction accuracy analysis and comparison. Accordingly, 23 datasets from EWI and 27 datasets from DML were used for the next stage of training and testing.

The results confirm that inconsistent data result both from unsupervised and supervised granulation techniques. Moreover, a higher number of attributes is expected to reduce IR, as they reduce the size of inconsistent subset to avoid the probability of same decision attribute values, and therefore the probability of inconsistent data.

Fig. 4 illustrates the relationship between IR from EWI and MDL based on L1 and number of attributes. Although not linear, IR oscillates between 0 and 99.78 % when number of attributes is below 15, but drops to less than 5 % over 15 attributes, for all datasets except D15. This indicates that adding more attributes is one way of reducing IR. However, doing so may also increase the number of decision attributes, making generated rules with less supports from a practical perspective. Solving

Table 5
Inconsistency rate from EWI and MDL for five acceptance levels.

DSC	Size	NA	NC	IR-EWI					IR-MDL				
				L1	L2	L3	L4	L5	L1	L2	L3	L4	L5
D01	772	17	2	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
D02	4177	8	29	73.35 %	76.20 %	76.35 %	76.56 %	76.56 %	87.69 %	89.59 %	90.09 %	90.38 %	90.38 %
D03	1503	5	C	70.06 %	73.39 %	76.98 %	77.84 %	77.84 %	98.60 %	98.60 %	100.00 %	100.00 %	100.00 %
D04	45,223	13	41	2.20 %	2.66 %	3.04 %	4.74 %	6.44 %	4.53 %	5.53 %	6.55 %	10.71 %	21.92 %
D05	20,867	10	12	54.17 %	61.01 %	67.94 %	70.90 %	71.80 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
D06	1372	4	2	0.36 %	0.36 %	1.38 %	1.38 %	1.38 %	9.69 %	9.69 %	9.69 %	13.12 %	33.82 %
D07	569	30	2	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
D08	683	9	2	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.44 %	0.73 %	0.73 %	0.73 %	0.73 %
D09	17,379	13	C	0.47 %	0.47 %	0.47 %	0.47 %	0.47 %	1.38 %	1.67 %	1.72 %	1.78 %	1.78 %
D10	345	6	2	13.33 %	16.81 %	19.42 %	19.42 %	19.42 %	59.71 %	100.00 %	100.00 %	100.00 %	100.00 %
D11	1473	9	3	37.61 %	47.39 %	50.71 %	53.50 %	53.50 %	62.73 %	73.18 %	78.82 %	79.90 %	79.90 %
D12	653	15	2	1.07 %	1.38 %	1.38 %	1.38 %	1.38 %	1.99 %	2.60 %	2.60 %	3.68 %	3.68 %
D13	13,611	16	7	3.82 %	5.31 %	6.33 %	7.43 %	7.57 %	0.18 %	0.24 %	0.24 %	0.24 %	0.24 %
D14	768	8	C	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	43.23 %	48.70 %	56.90 %	66.28 %	90.63 %
D15	14,980	14	2	99.78 %	99.78 %	99.80 %	99.80 %	99.80 %	16.98 %	23.10 %	31.01 %	39.15 %	44.95 %
D16	517	12	C	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	1.74 %	1.74 %	2.32 %	6.58 %	32.50 %
D17	214	9	7	10.28 %	14.02 %	15.42 %	15.42 %	15.42 %	18.22 %	32.24 %	35.05 %	57.48 %	62.15 %
D18	299	12	2	0.67 %	0.67 %	0.67 %	0.67 %	0.67 %	16.39 %	28.09 %	31.10 %	36.12 %	88.29 %
D19	17,898	8	2	1.35 %	1.73 %	1.98 %	2.96 %	22.71 %	1.31 %	1.60 %	1.95 %	2.81 %	40.64 %
D20	579	10	2	13.82 %	17.27 %	19.17 %	19.17 %	19.17 %	43.70 %	57.69 %	60.10 %	65.46 %	71.50 %
D21	150	4	3	0.67 %	0.67 %	2.67 %	2.67 %	2.67 %	3.33 %	7.33 %	9.33 %	9.33 %	16.00 %
D22	20,000	16	26	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.03 %	0.03 %	0.03 %	0.03 %	0.03 %
D23	589	12	4	0.34 %	0.34 %	0.34 %	0.34 %	0.34 %	1.36 %	1.70 %	1.70 %	3.06 %	3.06 %
D24	5473	10	5	5.57 %	7.47 %	8.42 %	12.19 %	38.39 %	1.19 %	1.74 %	1.99 %	2.98 %	3.16 %
D25	3498	16	10	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
D26	546	8	C	32.97 %	39.74 %	43.59 %	43.59 %	43.59 %	98.53 %	98.90 %	98.90 %	98.90 %	98.90 %
D27	2000	36	6	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	1.15 %	1.35 %	1.50 %	1.50 %	2.15 %
D28	2310	19	7	0.35 %	0.35 %	0.35 %	0.35 %	0.35 %	0.43 %	0.74 %	0.74 %	0.74 %	0.74 %
D29	14,500	9	7	1.18 %	2.24 %	2.24 %	3.97 %	28.33 %	0.01 %	0.02 %	0.02 %	0.02 %	0.02 %
D30	846	18	4	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	6.38 %	9.57 %	10.87 %	10.87 %	11.94 %
D31	990	10	11	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	5.25 %	8.38 %	10.71 %	12.93 %	12.93 %
D32	5000	21	3	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
D33	4839	5	2	5.39 %	5.48 %	6.53 %	26.12 %	60.49 %	2.87 %	2.91 %	3.55 %	8.08 %	54.06 %
D34	178	13	3	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
D35	1599	11	6	1.56 %	2.06 %	2.06 %	2.31 %	2.31 %	42.65 %	63.16 %	72.05 %	78.30 %	84.62 %
D36	4898	11	7	2.14 %	3.12 %	3.35 %	3.57 %	3.57 %	32.28 %	43.14 %	49.10 %	54.63 %	55.19 %

NA: number of attributes, NC: number of classes, IR-EWI: inconsistency rate from EWI, IR-MDL: inconsistency rate from MDL

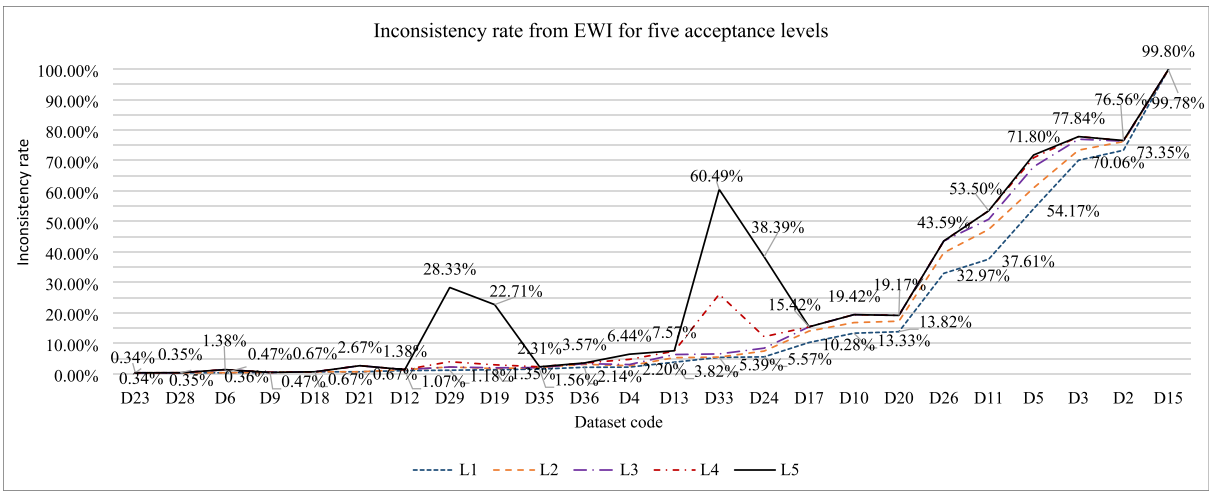

Fig. 2. Inconsistency rate from EWI for five acceptance levels.

this dilemma requires the consideration of various factors, such as domain characteristics, data availability, data collection plan, and application limitations. The chosen solution will depend on practical situations. Nonetheless, examining IR in a granulated dataset before classification mechanisms is likely a necessary condition for a program of machine learning.

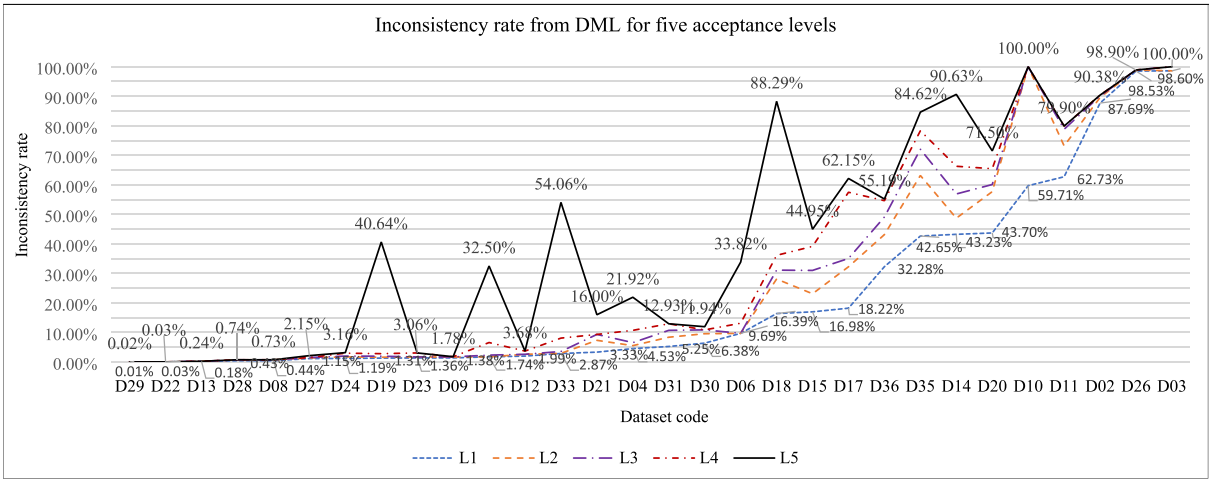

Fig. 3. Inconsistency rate from MDL for five acceptance levels.

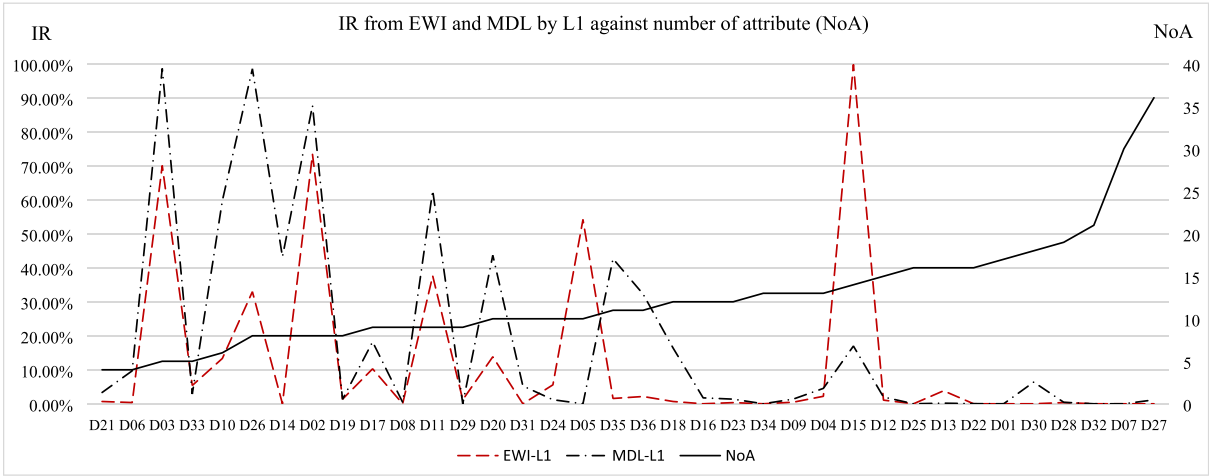

Fig. 4. IR from EWI and MDL by L1 against number of attribute (NoA).

4.2. Performance and comparison with EWI

4.2.1. PA with EWI using five classifiers

The PAs of datasets BG, AGBR, and AR with EWI using the five classifiers vary by dataset (Table 6). For example, on the one hand, the PA of D03 is 55.48 % with ID3, 68.15 % with RF, and 70.37 % with AdaBoost. On the other hand, the PA of D20 is 80.28 % with ID3, 77.86 % with AdaBoost, and 92.99 % with SVM. As for the PA of all datasets with an individual classifier, AR mostly performs better than both BG and AGBR. For example, the PA of AR datasets with ID3 is mostly higher than that of AGBR (22 out of 23 datasets). As for SVM, only one dataset reveals the highest PA on BG and two on AGBR, and only one on BG and two on AGBR for UMC-SGD. However, RF reveals nine on BG and one on AGBR, and AdaBoost shows 10 on BG and one on AGBR. Overall, the performance of datasets after the removal of granulated inconsistent datasets is likely better than that before granulation. As for the inconsistency acceptance level, Level-1 (60 %) likely reveals the best. For example, 12 Level-1 s are in ID3, 16 are in SVM, 13 are in RF, 11 are in AdaBoost, and 10 are in UMC-SGD.

The best classifier of PA on datasets along with the best level of removal of inconsistency is presented in Table 7. The number of datasets with the best performance varies by classifiers. For example, ID3 and RF have five datasets, SVM and AdaBoost have seven, and UMC-SGD has three. It likely reveals that SVM and AdaBoost perform slightly better than the other classifiers. Particularly, AdaBoost demonstrates the best one on three datasets (D05, D09, and D28) with no need of granulation (BG). With 11 out of 23 datasets, Level-1 likely reveals the best removal level of inconsistency that produces the best performance of PA among the five levels.

Table 6
Prediction accuracy with EWI on BG, AGBR, and AR datasets using five classifiers.

	ID3		SVM			RF			AdaBoost			UMC-SGD		
DSC	AGBR	AR (L)	BG	AGBR	AR (L)	BG	AGBR	AR (L)	BG	AGBR	AR (L)	BG	AGBR	AR (L)
D02	27.83 %	**34.85 %(1)**	24.42 %	25.30 %	**32.93 %(1)**	22.35 %	20.27 %	**31.44 %(1)**	23.06 %	22.27 %	**33.83 %(1)**	15.72 %	16.12 %	**24.85 %(1)**
D03	34.76 %	**55.48 %(1)**	31.49 %	30.60 %	**61.48 %(1)**	48.34 %	32.37 %	**68.15 %(1)**	57.22 %	31.04 %	**70.37 %(1)**	18.85 %	21.06 %	**42.50 %(2)**
D04	91.96 %	**93.53 %(1)**	91.72 %	91.60 %	**95.58 %(1)**	91.38 %	91.07 %	**93.47 %(2)**	87.62 %	89.24 %	**91.31 %(2)**	91.45 %	91.48 %	**93.42 %(2)**
D05	64.33 %	**95.77 %(1)**	49.71 %	62.54 %	**92.26 %(1)**	**99.76 %**	64.52 %	96.97 %(2)	**99.89 %**	64.04 %	96.48 %(2)	48.79 %	55.96 %	**82.54 %(4)**
D06	98.18 %	**99.51 %(1)**	98.06 %	98.30 %	**98.77 %(2)**	98.30 %	98.06 %	**99.51 %(1)**	99.27 %	98.79 %	**99.75 %(3)**	98.54 %	98.54 %	**99.51 %(3)**
D09	85.83 %	**86.51 %(5)**	**92.62 %**	84.48 %	85.05 %(1)	**94.74 %**	83.95 %	84.58 %(1)	**97.64 %**	81.70 %	82.35 %(1)	31.15 %	**71.58 %**	71.48 %(1)
D10	57.50 %	**69.11 %(1)**	63.11 %	66.99 %	**75.56 %(1)**	**70.87 %**	65.05 %	68.60 %(2)	**68.93 %**	56.31 %	67.44 %(2)	69.90 %	59.22 %	**75.58 %(2)**
D11	52.40 %	**65.18 %(1)**	47.29 %	52.49 %	**69.93 %(1)**	50.45 %	46.83 %	**66.67 %(1)**	48.64 %	47.74 %	**65.94 %(1)**	43.21 %	49.77 %	**68.97 %(2)**
D12	87.25 %	**87.73 %(2)**	81.63 %	84.18 %	**92.27 %(1)**	82.65 %	78.06 %	**90.21 %(1)**	79.64 %	75.00 %	**84.02 %(1)**	81.12 %	80.10 %	**88.66 %(1)**
D13	91.36 %	**94.07 %(1)**	91.92 %	90.67 %	**93.88 %(3)**	91.92 %	90.23 %	**94.36 %(2)**	91.70 %	89.42 %	**93.46 %(2)**	88.61 %	90.11 %	**93.36 %(3)**
D17	70.31 %	**73.40 %(2)**	51.56 %	73.44 %	**81.48 %(2)**	**78.13 %**	71.88 %	74.07 %(3)	**75.00 %**	68.75 %	72.22 %(3)	51.56 %	68.75 %	**77.78 %(3)**
D18	82.78 %	**88.45 %(1)**	77.78 %	80.00 %	**83.15 %(1)**	**81.11 %**	76.67 %	78.65 %(1)	**83.33 %**	78.89 %	74.16 %(1)	**81.11 %**	77.78 %	74.16 %(1)
D19	97.81 %	**99.18 %(1)**	97.32 %	97.34 %	**99.10 %(4)**	97.88 %	97.43 %	**99.12 %(4)**	97.75 %	97.17 %	**99.03 %(2)**	97.63 %	97.32 %	**99.06 %(4)**
D20	71.72 %	**80.28 %(2)**	72.99 %	**92.99 %**	80.00 %(2)	69.54 %	67.81 %	**80.00 %(3)**	68.97 %	63.79 %	**77.86 %(3)**	72.99 %	63.22 %	**74.31 %(2)**
D21	94.00 %	**97.34 %(2)**	95.56 %	95.56 %	**97.78 %(1)**	**95.56 %**	**95.56 %**	95.45 %(3)	**95.56 %**	**95.56 %**	**95.56 %(1)**	68.89 %	**95.56 %**	93.33 %(1)
D23	**92.66 %**	92.49 %(5)	89.83 %	90.96 %	**93.18 %(1)**	**93.22 %**	88.70 %	92.61 %(1)	93.79 %	87.00 %	**94.32 %(1)**	90.40 %	89.83 %	**93.75 %(1)**
D24	93.84 %	**98.49 %(4)**	93.30 %	94.82 %	**98.60 %(3)**	97.99 %	93.73 %	**98.14 %(1)**	97.69 %	93.48 %	**98.00 %(3)**	93.79 %	94.46 %	**98.67 %(3)**
D26	39.57 %	**44.27 %(1)**	31.71 %	42.07 %	**43.64 %(1)**	41.46 %	33.15 %	**51.82 %(1)**	41.46 %	35.98 %	**49.09 %(1)**	23.17 %	23.17 %	**28.18 %(1)**
D28	93.95 %	**94.99 %(5)**	93.65 %	**95.67 %**	95.66 %(1)	**98.27 %**	96.97 %	96.96 %(1)	**98.70 %**	95.96 %	95.95 %(1)	83.26 %	89.90 %	**90.30 %(1)**
D29	98.75 %	**99.96 %(3)**	95.29 %	98.46 %	**99.91 %(1)**	**99.89 %**	98.46 %	99.86 %(1)	**99.93 %**	98.46 %	99.90 %(5)	84.00 %	98.44 %	**99.91 %(1)**
D33	94.34 %	**100.0 %(1)**	94.00 %	94.28 %	**100.0 %(1)**	98.42 %	94.28 %	**100.0 %(1)**	98.69 %	95.96 %	**100.0 %(1)**	94.00 %	94.28 %	**100.0 %(1)**
D35	60.69 %	**61.68 %(2)**	59.17 %	67.29 %	**69.15 %(3)**	68.75 %	66.04 %	**69.30 %(4)**	**67.50 %**	62.92 %	66.52 %(4)	43.33 %	56.46 %	**59.15 %(2)**
D36	56.93 %	**58.35 %(3)**	51.53 %	61.81 %	**65.65 %(1)**	67.05 %	62.29 %	**66.34 %(4)**	**64.40 %**	58.75 %	63.46 %(2)	50.00 %	52.55 %	**54.18 %(2)**
FBL		12:5:2:1:3	–	–	16:3:3:1:0	–	–	13:4:3:3:0	–	–	11:6:4:1:1			10:7:4:2:0
NHP	1/23	22/23	1/23	2/23	20/23	9/23	1/23	14/23	10/23	1/23	14/23	1/23	2/23	20/23

ID3: iterative dichotomiser 3, SVM: Support vector machine, RF: Random forest, AdaBoost: Adaptive boosting, UMC-SGD; Updatable multiclass classifier with stochastic gradient descent, DSC: Dataset code, BG: Before granulation, AGBR: After granulation but before removal of inconsistency, AR(L): After L removal level for inconsistency, FBL: Frequency of best removal level for inconsistency (L1:L2:L3:L4:L5), NHP: Number of datasets with highest performance for a classifier, The bold number is the highest PA.

Table 7
Best prediction accuracy with EWI by five classifiers.

DS#	ID3	SVM	RF	AdaBoost	UMC-SGD
D02	**34.85 %(1)**	32.93 %(1)	31.44 %(1)	33.83 %(1)	24.85 %(1)
D03	55.48 %(1)	61.48 %(1)	68.15 %(1)	**70.37 %(1)**	42.50 %(2)
D04	93.53 %(1)	**95.58 %(1)**	93.47 %(2)	91.31 %(2)	93.42 %(2)
D05	95.77 %(1)	92.26 %(1)	99.76 %(BG)	**99.89 %(BG)**	82.54 %(4)
D06	99.51 %(1)	98.77 %(2)	99.51 %(1)	**99.75 %(3)**	99.51 %(3)
D09	86.51 %(5)	92.62 %(BG)	94.74 %(BG)	**97.64 %(BG)**	71.48 %(AGBR)
D10	69.11 %(1)	75.56 %(1)	70.87 %(BG)	68.93 %(BG)	**75.58 %(2)**
D11	65.18 %(1)	**69.93 %(1)**	66.67 %(1)	65.94 %(1)	68.97 %(2)
D12	87.73 %(2)	**92.27 %(1)**	90.21 %(1)	84.02 %(1)	88.66 %(1)
D13	94.07 %(1)	93.88 %(3)	**94.36 %(2)**	93.46 %(2)	93.36 %(3)
D17	73.40 %(2)	**81.48 %(2)**	74.07 %(BG)	75.00 %(BG)	77.78 %(3)
D18	**88.45 %(1)**	83.15 %(1)	78.65 %(BG)	83.33 %(BG)	74.16 %(BG)
D19	**99.18 %(1)**	99.10 %(4)	99.12 %(4)	99.03 %(2)	99.06 %(4)
D20	80.28 %(2)	**92.99 %(AGBR)**	80.00 %(3)	77.86 %(3)	74.31 %(2)
D21	97.34 %(2)	**97.78 %(1)**	95.56 %(BG)	95.56 %(1)	93.33 %(1)
D23	92.66 %(5)	93.18 %(1)	93.22 %(BG)	**94.32 %(1)**	93.75 %(1)
D24	98.49 %(4)	98.60 %(3)	98.14 %(1)	98.00 %(3)	**98.67 %(3)**
D26	44.27 %(1)	43.64 %(1)	**51.82 %(1)**	49.09 %(1)	28.18 %(1)
D28	94.99 %(5)	95.67 %(AGBR)	98.27 %(BG)	**98.70 %(BG)**	90.30 %(1)
D29	**99.96 %(3)**	99.91 %(1)	99.89 %(BG)	99.93 %(BG)	99.91 %(1)
D33	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(1)**
D35	61.68 %(2)	69.15 %(3)	**69.30 %(4)**	67.50 %(BG)	59.15 %(2)
D36	58.35 %(3)	65.65 %(1)	**66.34 %(4)**	64.40 %(BG)	54.18 %(2)
NDBP	5/23	7/23	5/23	7/23	3/23
NDBG	–	0/7	0/5	3/7	0/4

ID3: iterative dichotomiser 3, SVM: Support vector machine, RF: Random forest, AdaBoost: Adaptive boosting, UMC-SGD; Updatable multiclass classifier updatable with stochastic gradient descent, (L): The removal level for inconsistency, NDBP: Number of datasets with the best performance among five classifiers, NDBG: Number of datasets with before granulation in NDBP.

4.2.2. Effect of IR on PA improvement with EWI

The IR and its corresponding improvement of PA by the five classifiers are presented in Table 8. For a dataset, IR may vary because the best removal level of inconsistency is used and its improvement of PA differs in classifiers. For example, D02 reveals the same IR, but IPA is 7.02 % by ID3, 9.07 % by SVM, 11.17 % by RF, 11.56 % by AdaBoost, and 8.73 % by UMC-SGD. With various IRs, D05 shows improvements of PA by ID3 (31.44 %), SVM (34.07 %), and UMC-SGD (26.58 %) but none by RF and AdaBoost because the PA by RF and AdaBoost on BG is higher than on AR. Granulation or after removal of inconsistency seems not applicable to some datasets. For example, the PA of BG on D09 is higher than that of AGBR or AR for four classifiers (SVM, RF, AdaBoost, and UMC-SGD). Similar situations occur on D18, D21, and D28.

The mean and standard deviation of PA improvement for the five classifiers, ranging from 5.74 % to 10.01 %, are also revealed in the table. The standard deviation increases with the mean. The highest performance of PA improvement is by AdaBoost, whereas the lowest is by ID3. Ensemble-based classifier likely performs better than simple tree-based classifier, although ID3 performs the best PA on three out of 23 datasets (Table 7). With various numbers of datasets, the coefficient of correlation (CoR) between IR and PA improvement ranges from 0.6976 to 0.7901. The highest CoR is by UMC-SGD, whereas the lowest is by SVM. Overall, the CoR demonstrates a moderately high relationship between IR and PA improvement.

4.3. Performance and comparison with MDL

4.3.1. PA with MDL using five classifiers

The prediction accuracy of datasets BG, AGBR, and AR using the five classifiers also varies with MDL by dataset (Table 9). For example, on the one hand, the PA of D04 is 95.49 % with ID3, 95.28 % with SVM, and 95.23 % with UMC-SGD. On the other hand, the PA of D11 is 83.15 % with ID3, 92.73 % with SVM, 91.52 with RF, and 86.67 % with UMC-SGD. Datasets D10 and D16 are not available because they contain a single class value after the removal of inconsistency. As for the PA of all datasets with an individual classifier, AR mostly performs better than BG and AGBR. For example, the PA of AR datasets with ID3 is mostly higher than that of AGBR (24 out of 25 datasets), whereas 23 out of 25 on AR datasets are higher with UMC-SGD. As for SVM, 24 datasets reveal the highest PA on AR than BR and AGBR. RF shows 20, whereas AdaBoost reveals 18 on AR. The overall performance on datasets with MDL by the five classifiers after the removal of granulated inconsistent datasets is likely higher than before granulation. As for the inconsistency acceptance level, Level-1 criteria (40 % removal) likely again reveals the best. For example, nine Level-1 s are in ID3, 10 in SVM, 13 in RF, 15 in AdaBoost, and nine in UMC-SGD.

The best classifier with PA on the datasets along with the best level of removal of inconsistency is presented in Table 10. The number of datasets with the best performance varies by classifier. For example, of the 25 datasets used, seven datasets

Table 8
IR and improvement of PA with EWI by five classifiers.

DSC	ID3		SVM		RF		AdaBoost		UMC-SGD	
	IR	IPA	IR	IPA	IR	IPA	IR	IPA	IR	IPA
D02	73.35 %	7.02 %	73.35 %	9.07 %	73.35 %	11.17 %	73.35 %	11.56 %	73.35 %	8.73 %
D03	70.06 %	20.72 %	70.06 %	30.88 %	70.06 %	35.78 %	70.06 %	39.33 %	73.39 %	21.44 %
D04	2.20 %	1.57 %	2.20 %	4.13 %	2.66 %	2.40 %	2.66 %	2.07 %	2.66 %	1.94 %
D05	54.17 %	31.44 %	54.17 %	34.07 %	BG	BG	BG	BG	70.90 %	26.58 %
D06	0.36 %	1.33 %	0.36 %	0.47 %	0.36 %	1.45 %	1.38 %	0.96 %	1.38 %	0.97 %
D09	1.38 %	0.68 %	BG	BG	BG	BG	BG	BG	AGBR	AGBR
D10	13.33 %	11.61 %	13.33 %	17.31 %	BG	BG	BG	BG	16.81 %	16.36 %
D11	37.61 %	12.78 %	37.61 %	17.44 %	27.61 %	19.84 %	47.39 %	18.20 %	47.39 %	19.20 %
D12	1.38 %	0.48 %	1.07 %	11.15 %	1.07 %	12.15 %	1.07 %	9.02 %	1.07 %	8.56 %
D13	3.82 %	2.71 %	6.33 %	3.21 %	5.31 %	4.13 %	5.31 %	4.04 %	6.33 %	3.25 %
D17	14.02 %	3.09 %	14.02 %	8.04 %	BG	BG	BG	BG	15.42 %	9.03 %
D18	0.67 %	5.67 %	0.67 %	6.48 %	BG	BG	BG	BG	BG	BG
D19	1.35 %	1.37 %	2.96 %	1.78 %	2.96 %	1.69 %	1.73 %	1.86 %	2.96 %	1.74 %
D20	17.27 %	8.56 %	AGBR	AGBR	19.17 %	12.19 %	19.17 %	14.07 %	17.27 %	11.09 %
D21	0.67 %	3.34 %	0.67 %	2.22 %	BG	BG	BG	BG	AGBR	AGBR
D23	0.34 %	0.00 %	0.34 %	2.22 %	BG	BG	0.34 %	7.32 %	0.34 %	3.92 %
D24	12.19 %	4.65 %	12.19 %	3.78 %	5.57 %	4.41 %	8.42 %	4.52 %	8.42 %	4.21 %
D26	32.97 %	4.70 %	32.97 %	6.44 %	32.97 %	18.67 %	32.97 %	13.11 %	32.97 %	5.01 %
D28	0.35 %	1.04 %	AGBR	AGBR	BG	BG	BG	BG	0.35 %	0.40 %
D29	2.24 %	1.21 %	1.18 %	1.45 %	BG	BG	BG	BG	1.18 %	1.47 %
D33	5.39 %	5.66 %	5.39 %	5.72 %	5.39 %	5.72 %	5.39 %	4.04 %	5.39 %	5.72 %
D35	2.06 %	0.99 %	2.06 %	10.61 %	2.31 %	3.26 %	BG	BG	2.06 %	2.69
D36	3.35 %	1.42 %	2.14 %	11.40 %	3.57 %	4.05 %	BG	BG	3.12 %	1.63 %
Mean		5.74 %		9.39 %		9.78 %		10.01 %		7.70 %
Std.		7.47 %		9.29 %		9.65 %		10.31 %		7.58 %
CoR.	0.7413		0.6976		0.7605		0.7799		0.7901	
n	23		20		14		13		20	

DSC: Dataset code, IR: Inconsistency rate, IPA: Improvement of prediction accuracy, ID3: iterative dichotomiser 3, SVM: Support vector machine, RF: Random forest, AdaBoost: Adaptive boosting, UMC-SGD; Updatable multiclass classifier updatable with stochastic gradient descent, BG: Before granulation, AGBR: After granulation but before removal of inconsistency, Std: Standard deviation, CoR: Coefficient of correlation, n: the number of datasets used.

are by ID3, whereas six are by UMC-SGD. Both SVM and AdaBoost have 12, whereas RF has 13. Particularly, AdaBoost demonstrates the best one on four datasets (D08, D09, D28, and D29) with no need of granulation (BG). With nine out of 25 datasets, Level-1 likely reveals the best removal level of inconsistency producing the best performance of PA among the five levels.

4.3.2. Effect of IR on PA improvement with MDL

The IR and its corresponding improvement of PA by the five classifiers is presented in Table 11. For a dataset, the IR may vary because the best removal level of inconsistency is used, and its improvement of PA differs in classifiers. For example, on the one hand, D11 reveals the same IR, but IPA is 29.28 % by ID3, 42.73 % by SVM, 43.33 % by RF, 40.61 % by AdaBoost, and 35.99 % by UMC-SGD. With various IRs, D12 shows 1.78 % improvement of PA by ID3, 10.96 % by SVM, 8.98 % by RF, 10.49 % by AdaBoost, and 5.36 % by UMC-SGD. On the other hand, the improvement of PA on D31 is 4.67 % by ID3, 10.43 % by SVM, and 4.60 by UMC-SGD, but no improvement of PA is obtained by RF and AdaBoost because the PA of BG on the dataset is higher than that of AGBR or AR. Granulation or after removal of inconsistency seems unnecessary for some datasets. For example, the PA of BG on D09 is higher than that of AR for three classifiers (SVM, RF, and AdaBoost). Similar situations occur on D22 and D31.

The mean and standard deviation of PA improvement for five classifiers, ranging from 5.74 % to 10.31 % and 10.14 % to 13.40 %, respectively, are also presented in the table. The standard deviation oscillates with the mean. Same as with EWI, the highest performance of PA improvement is by AdaBoost, whereas the lowest is by ID3. Although ID3 performs the best PA on seven out of 25 datasets (Table 10), the ensemble-based classifiers likely perform better than the simple tree-based classifier with MDL. With various number of datasets, the coefficient of correlation (CoR) between IR and PA improvement ranges from 0.7870 to 0.9683. The highest CoR is by ID3, whereas the lowest is by UMC-SGD. Overall, the CoR demonstrates a fairly high relationship between IR and PA improvement with MDL, which is relatively higher than that with EWI.

5. Discussion and limitations

The experimental results on the application of unsupervised (EWI) and supervised (MDL) granulation techniques using the five classifiers confirm that IR values in a granulated dataset influence classification performance independent from levels of inconsistency acceptance. The effect on classification performance with both EWI and MDL depends likely on the degree of IR. This result implies that examining how granulation processing influences IR is an issue to be considered in classification attempts. Particularly, the research finding reveals that a lower PA on granulated datasets likely facilitates the

Table 9
Prediction accuracy with MDL on BG, AGBR, and AR datasets using five classifiers.

	ID3		SVM			RF			AdaBoost			UMC-SGD		
DSC	AGBR	AR(L)	BG	AGBR	AR (L)	BG	AGBR	AR (L)	BG	AGBR	AR (L)	BG	AGBR	AR (L)
D04	92.01 %	**95.49 %(2)**	91.72 %	91.69 %	**95.28 %(3)**	91.38 %	90.79 %	**95.44 %(3)**	87.62 %	89.92 %	**94.04 %(1)**	91.45 %	91.12 %	**95.23 %(3)**
D06	94.76 %	**100.0 %(1)**	98.06 %	95.39 %	**100.0 %(1)**	98.30 %	95.39 %	**100.0 %(1)**	99.27 %	95.39 %	**100.0 %(4)**	98.54 %	95.39 %	**100.00 %(4)**
D08	95.12 %	**97.99 %(5)**	97.07 %	97.07 %	**97.54 %(3)**	**97.07 %**	96.10 %	96.55 %((2)	**98.05 %**	95.61 %	95.10 %(1)	97.07 %	**97.56 %**	96.55 %(2)
D09	93.57 %	**94.36 %(4)**	**92.62 %**	91.01 %	91.81 %(1)	**94.74 %**	92.12 %	93.84 %(2)	**97.64 %**	92.10 %	93.68 %(3)	31.15 %	74.61 %	**76.38 %(2)**
D10	62.31 %	N/A	63.11 %	67.96 %	N/A	70.87 %	67.96 %	N/A	68.93 %	67.96 %	N/A	69.90 %	67.96 %	N/A
D11	53.87 %	**83.15 %(1)**	47.29 %	50.00 %	**92.73 %(1)**	50.54 %	48.19 %	**91.52 %(1)**	48.64 %	50.91 %	**91.52 %(1)**	43.21 %	50.68 %	**86.67 %(1)**
D12	86.23 %	**88.01 %(2)**	81.63 %	84.18 %	**92.59 %(4)**	82.65 %	81.12 %	**90.10 %(1)**	79.08 %	78.57 %	**89.06 %(1)**	81.12 %	84.69 %	**90.05 %(2)**
D13	91.46 %	**91.75 %(4)**	91.92 %	91.84 %	**92.12 %(3)**	**91.92 %**	90.99 %	91.19 %**(1)**	**91.70 %**	90.00 %	90.89 %(1)	88.61 %	**91.05 %**	90.80 %(2)
D14	71.65 %	**100.0 %(1)**	56.96 %	73.48 %	**100.0 %(1)**	72.17 %	73.04 %	**100.0 %(1)**	68.26 %	73.04 %	**100.0 %(1)**	40.87 %	56.52 %	**100.00 %(1)**
D15	78.11 %	**90.24 %92)**	55.61 %	79.95 %	**94.72 %(1)**	92.50 %	80.77 %	**94.93 %(1)**	93.70 %	80.57 %	**94.64 %(1)**	55.63 %	72.72 %	**82.81 %(2)**
D16	97.95 %	N/A	98.71 %	98.71 %	N/A	98.71 %	98.71 %	N/A	98.71 %	98.71 %	N/A	98.71 %	98.71 %	N/A
D17	71.69 %	**88.10 %(3)**	51.56 %	78.13 %	**95.14 %(3)**	91.92 %	68.75 %	**95.24 %(3)**	75.00 %	75.00 %	**95.24 %(3)**	51.56 %	60.94 %	**94.23 %(1)**
D18	85.89 %	**99.52 %(3)**	77.78 %	85.56 %	**100.0 %(5)**	81.11 %	83.33 %	**100.0 %(1)**	83.33 %	85.56 %	**100.0 %(1)**	81.11 %	83.33 %	**100.00 %(1)**
D19	97.98 %	**99.16 %(4)**	97.32 %	97.49 %	**99.45 %(2)**	97.88 %	97.62 %	**99.31 %(5)**	97.75 %	97.30 %	**99.12 %(1)**	97.63 %	97.56 %	**99.18 %(3)**
D20	70.75 %	**99.00 %(3)**	72.99 %	72.99 %	**100.0 %(2)**	69.54 %	70.11 %	**100.0 %(2)**	68.97 %	68.97 %	**100.0 %(3)**	72.99 %	72.99 %	**100.00 %(3)**
D21	95.34 %	**100.0 %(2)**	95.56 %	95.56 %	**100.0 %(1)**	95.56 %	95.56 %	**100.0 %(1)**	95.56 %	97.78 %	**100.0 %(1)**	68.89 %	95.56 %	**100.00 %(1)**
D22	**82.38 %**	81.88 %(4)	81.55 %	89.80 %	**89.78 %(1)**	**95.43 %**	91.83 %	92.20 %(1)	**95.02 %**	88.95 %	90.10 %(1)	39.00 %	75.47 %	**76.56 %(1)**
D23	93.90 %	**94.94 %(4)**	89.83 %	95.48 %	**98.25 %(5)**	93.22 %	91.53 %	**98.83 %(4)**	93.79 %	91.53 %	**98.83 %(4)**	90.40 %	94.35 %	**98.25 %(4)**
D24	97.24 %	**98.42 %(4)**	93.30 %	97.81 %	**98.45 %(3)**	97.99 %	97.81 %	**99.01 %(3)**	97.69 %	97.87 %	**98.45 %(3)**	93.79 %	96.65 %	**97.80 %(4)**
D27	66.22 %	**68.22 %(4)**	60.50 %	64.50 %	**64.30 %(4)**	63.00 %	62.67 %	**64.92 %(1)**	58.67 %	57.33 %	**61.89 %(1)**	48.67 %	54.50 %	**58.21 %(3)**
D28	96.12 %	**96.59 %(1)**	93.65 %	97.55 %	**97.53 %(2)**	98.27 %	96.83 %	**98.40 %(2)**	**98.70 %**	96.97 %	96.95 %(2)	83.26 %	93.07 %	**94.33 %(2)**
D29	99.82 %	**99.90 %(1)**	95.29 %	99.89 %	**99.93 %(2)**	99.89 %	99.81 %	**99.93 %(2)**	**99.93 %**	99.84 %	99.91 %(2)	84.00 %	99.89 %	**99.91 %(2)**
D30	73.78 %	**82.91 %(3)**	74.31 %	70.47 %	**79.20 %(3)**	73.52 %	70.08 %	**79.82 %(5)**	74.70 %	69.69 %	**77.43 %(3)**	64.82 %	67.32 %	**71.30 %(5)**
D31	71.41 %	**76.08 %(2)**	61.95 %	81.48 %	**87.87 %(2)**	**91.58 %**	81.48 %	89.34 %(2)	**88.55 %**	78.45 %	87.87 %(2)	27.61 %	61.95 %	**66.55 %(1)**
D33	97.18 %	**100.0 %(1)**	94.00 %	97.31 %	**100.0 %(1)**	98.42 %	97.25 %	**100.0 %(1)**	98.69 %	97.25 %	**100.0 %(1)**	94.00 %	97.31 %	**100.0 %(1)**
D35	59.58 %	**88.04 %(1)**	59.17 %	58.75 %	**94.18 %(1)**	68.75 %	56.88 %	**93.82 %(1)**	67.50 %	59.38 %	**92.36 %(1)**	43.33 %	52.08 %	**83.27 %(1)**
D36	56.27 %	**67.80 %(1)**	51.53 %	57.18 %	**79.80 %(1)**	67.05 %	55.07 %	**77.99 %(1)**	64.40 %	55.62 %	**76.28 %(1)**	50.00 %	51.33 %	**61.80 %(2)**
FBL		9:5:3:7:1			10:5:6:2:2			13:6:3:1:2			15:3:5:2:0			9:8:4:3:1
NHP	1/25	24/25	1/25	0/25	24/25	5/25	0/25	21/25	7/25	0/25	18/25	0/25	2/25	23/25

ID3: iterative dichotomiser 3, SVM: Support vector machine, RF: Random forest, AdaBoost: Adaptive boosting, UMC-SGD; Updatable multiclass classifier with stochastic gradient descent, DSC: Dataset code, BG: Before granulation, AGBR: After granulation but before removal of inconsistency, AR(L): After L removal level for inconsistency, FBL: Frequency of best removal level for inconsistency (L1:L2:L3:L4:L5), NHP: Number of datasets with highest performance, The bold number is the highest PA for the classifier, N/A: Not available due to dataset with a single class after removal of inconsistency,

Table 10
Best prediction accuracy with MDL by five classifiers.

DS#	ID3	SVM	RF	AdaBoost	UMC-SGD
D04	**95.49 %(2)**	95.28 %(3)	95.44 %(3)	94.04 %(1)	95.23 %(3)
D06	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(4)**	**100.00 %(4)**
D08	97.99 %(5)	97.54 %(3)	97.07 %(BG)	**98.05 %(BG)**	96.55 %(2)
D09	94.36 %(4)	92.62 %(BG)	94.74 %(BG)	**97.64 %(BG)**	76.38 %(2)
D10	NA	N/A	N/A	N/A	N/A
D11	83.15 %(1)	**92.73 %(1)**	91.52 %(1)	91.52 %(1)	86.67 %(1)
D12	88.01 %(2)	**92.59 %(4)**	90.10 %(1)	89.06 %(1)	90.05 %(2)
D13	91.75 %(4)	**92.12 %(3)**	91.92 %(BG)	91.70 %(BG)	90.80 %(2)
D14	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(1)**	**100.00 %(1)**
D15	90.24 %(2)	94.72 %(1)	**94.93 %(1)**	94.64 %(1)	82.81 %(2)
D16	NA	N/A	N/A	N/A	N/A
D17	88.10 %(3)	95.14 %(3)	**95.24 %(3)**	**95.24 %(3)**	94.23 %(1)
D18	99.52 %(3)	**100.0 %(5)**	**100.0 %(1)**	**100.0 %(1)**	**100.00 %(1)**
D19	99.16 %$4)	**99.45 %(2)**	99.31 %(5)	99.12 %(1)	99.18 %(3)
D20	99.00 %(3)	**100.0 %(2)**	**100.0 %(2)**	**100.0 %(3)**	**100.00 %(3)**
D21	**100.0 %(2)**	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(1)**	**100.00 %(1)**
D22	81.88 %(4)	89.78 %(1)	**95.43 %(BG)**	95.02 %(BG)	76.56 %(1)
D23	94.94 %(4)	98.25 %(5)	**98.83 %(4)**	**98.83 %(4)**	98.25 %(4)
D24	98.42 %(4)	98.45 %(3)	**99.01 %(3)**	98.45 %(3)	97.80 %(4)
D27	**68.22 %(4)**	64.30 %(4)	64.92 %(1)	61.89 %(1)	58.21 %(3)
D28	96.59 %(1)	97.53 %(2)	98.40 %(2)	**98.70 %(BG)**	94.33 %(2)
D29	99.90 %(1)	99.93 %(2)	**99.93 %(2)**	**99.93 %(BG)**	99.91 %(2)
D30	**82.91 %(3)**	79.20 %(3)	79.82 %(5)	77.43 %(3)	71.30 %(5)
D31	76.08 %(2)	87.87 %(2)	**91.58 %(BG)**	88.55 %(BG)	66.55 %(1)
D33	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(1)**	**100.0 %(1)**
D35	88.04 %(1)	**94.18 %(1)**	93.82 %(1)	92.36 %(1)	83.27 %(1)
D36	67.80 %(1)	**79.80 %(1)**	77.99 %(1)	76.28 %(1)	61.80 %(2)
NDBP	7/25	12/25	13/25	12/25	6/25
NDBG	–	0/12	2/13	4/12	0/6

ID3: iterative dichotomiser 3, SVM: Support vector machine, RF: Random forest, AdaBoost: Adaptive boosting, UMC-SGD; Updatable multiclass classifier updatable with stochastic gradient descent, (L): The removal level for inconsistency, NDBP: Number of datasets with the best performance among five classifiers, NDBG: Number of datasets with before granulation in NDBP.

exclusion of inconsistent data and is conducive to PA improvement after the removal of inconsistent datasets. The five selected classifiers mostly obtain positive PA improvement on AR datasets (Tables 8 and 11). This finding is supported by the mean PA improvement ranging from 5.74 % to 10.01 % with EWI and from 8.74 % to 13.73 % with MDL. The best combination based on the experimental results is likely MDL granulation technique, L1 removal level, and AdaBoost classifier.

The experiment selected five classifiers including one simple tree-based and four ensemble-based classifiers with particular settings. These classifiers likely present higher PA on AR datasets than BG with both EWI and MDL. This result implies that these classifiers are likely sensitive to inconsistent datasets in granulated datasets [18,30], independent from inconsistency acceptance levels. For example, the bagging approach (e.g., RF classifier) tries to avoid randomly selecting noisy data being trained to reach a better result. However, when a dataset contains a certain size of noisy data, the possibility of including noisy data in the training remains, which is prone to increasing higher failed classification and producing lower PA. Moreover, the boosting approach (e.g., AdaBoost classifier) increases the weight for misclassified data to develop better patterns and improve classification outcomes. However, again when a dataset contains a certain size of noisy data, the boosting approach faces even more difficulty in reaching a better goal because it focuses mainly on misclassified data that probably contain unacceptably inconsistent data. In this situation, manipulating data weight for dataset with higher inconsistency rate likely produces a less positive effect on classification success. Reducing such possibility by managing noisy data seems advantageous to PA improvement, as addressed in literature [18,30]. The experimental results confirm this attempt based on the evidence that IR is highly related to PA improvement with both EWI and MDL by five classifiers (Tables 8 and 11).

Particularly, the selected AdaBoost in the experiment based on boosting approach with a bagging mechanism reveals the highest mean positive PA improvement. It also presents the highest number of BG datasets with both EWI (10 of 23) and MDL (7 of 25) (Tables 7 and 10). On the one hand, this signifies that reducing bias via boosting and reducing variance via bagging are advantageous to classification success on BG datasets. On the other hand, removing inconsistent datasets for this IR-sensitive classifier is even more advantageous to PA improvement because noisy data may negatively influence the performance of bagging and boosting approaches, or combined approaches.

These findings are relevant because existing literature has not addressed in depth the links among granulation techniques, IR, and classification performance. Although PA depends on various factors such as application domains, data characteristics, preprocessing techniques and strategies, learning mechanisms, and testing criteria, the experimental results demonstrate the importance of IR resulting from granulation techniques and the effect on learning performance. Moreover, despite the difficulty in arguing which combinations with settings perform best on any dataset regarding granulation technique (unsupervised or supervised), detection and removal of inconsistent data with which level or specific rate, simple or

Table 11
IR and improvement of PA with MDL by five classifiers.

DSN	ID3		SVM		RF		AdaBoost		UMC-SGD	
	IR	IPA	IR	IPA	IR	IPA	IR	IPA	IR	IPA
D04	5.53 %	3.48 %	6.55 %	3.59 %	6.55 %	4.65 %	4.53 %	4.12 %	6.55 %	4.11 %
D06	9.69 %	5.24 %	9.69 %	4.61 %	9.69 %	4.61 %	13.12 %	4.61 %	13.12 %	4.61 %
D08	0.73 %	2.87 %	AGBR	AGBR	BG	BG	BG	BG	0.73 %	−1.01 %
D09	1.38 %	0.79 %	BG	BG	BG	BG	BG	BG	1.67 %	1.77 %
D10	59.71 %	N/A	N/A	N/A	NA	N/A	N/A	N/A	N/A	N/A
D11	62.73 %	29.28 %	62.73 %	42.73 %	62.73 %	43.33 %	62.73 %	40.61 %	62.73 %	35.99 %
D12	2.60 %	1.78 %	3.68 %	10.96 %	1.99 %	8.98 %	1.99 %	10.49 %	2.60 %	5.36 %
D13	0.24 %	0.29 %	0.24 %	0.28 %	BG	BG	BG	BG	0.24 %	−0.25 %
D14	43.23 %	28.35 %	43.23 %	26.52 %	43.23 %	26.96 %	43.23 %	26.96 %	43.23 %	43.48 %
D15	23.10 %	12.13 %	16.98 %	20.98 %	16.98 %	14.16 %	16.98 %	14.07 %	23.10 %	10.09 %
D16	1.74 %	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
D17	35.05 %	16.41 %	35.05 %	24.83 %	35.05 %	26.49 %	35.05 %	20.24 %	18.22 %	33.29 %
D18	31.10 %	13.63 %	88.29 %	16.67 %	16.39 %	16.67 %	16.39 %	14.44 %	16.39 %	16.67 %
D19	2.81 %	1.18 %	1.60 %	1.96 %	40.64 %	1.69 %	1.31 %	1.82 %	1.60 %	1.62 %
D20	60.10 %	28.25 %	57.69 %	27.01 %	57.69 %	29.89 %	60.10 %	31.03 %	57.69 %	27.01 %
D21	7.33 %	4.66 %	3.33 %	4.44 %	3.33 %	4.44 %	3.33 %	2.22 %	3.33 %	4.44 %
D22	AGBR	AGBR	AGBR	AGBR	BG	BG	BG	BG	0.03 %	1.09 %
D23	3.06 %	1.04 %	3.06 %	3.33 %	3.06 %	7.30 %	3.06 %	7.30 %	3.06 %	3.90 %
D24	2.98 %	1.18 %	1.99 %	0.64 %	1.99 %	1.20 %	1.99 %	0.58 %	2.98 %	1.15 %
D27	1.50 %	2.00 %	1.50 %	3.47 %	1.15 %	2.25 %	1.15 %	4.56 %	1.50 %	3.71 %
D28	0.43 %	0.47 %	0.74 %	0.27 %	0.74 %	1.57 %	BG	BG	0.74 %	1.26 %
D29	0.01 %	0.08 %	0.02 %	0.04 %	0.02 %	0.12 %	BG	BG	0.02 %	0.02 %
D30	10.87 %	9.13 %	10.87 %	8.73 %	11.94 %	9.74 %	10.87	7.74 %	11.94 %	3.98 %
D31	8.38 %	4.67 %	8.38 %	10.43 %	BG	BG	BG	BG	5.25 %	4.60 %
D33	2.87 %	2.82 %	2.87 %	2.69 %	2.87 %	2.75 %	2.87 %	2.75 %	2.87 %	2.69 %
D35	42.65 %	28.46 %	42.65 %	36.68 %	42.65 %	36.94 %	42.65 %	32.98 %	0.4265	0.3119
D36	32.28 %	11.53 %	32.28 %	27.18 %	32.28 %	22.92 %	32.28	20.66 %	0.4314	0.1047
Mean		8.74 %		12.64 %		13.33 %		13.73 %		10.03 %
Std		10.14 %		13.02 %		13.23 %		12.30 %		13.40 %
CoR	0.9683		0.7870		0.8636		0.9607		0.8409	
n	24		22		20		18		25	

DSC: Dataset code, IR: Inconsistency rate, IPA: Improvement of prediction accuracy, ID3: iterative dichotomiser 3, SVM: Support vector machine, RF: Random forest, AdaBoost: Adaptive boosting, UMC-SGD; Updatable multiclass classifier updatable with stochastic gradient descent, BG: Before granulation, AGBR: After granulation but before removal of inconsistency, Std: Standard deviation, CoR: Coefficient of correlation, n: the number of datasets used.

ensemble classifiers with various approaches and settings, and testing scenario, the experimental findings show that managing IR is one of the alternatives worth trying, especially for cases when the PA of BG datasets can still be improved.

Similar conditions may produce different results in real-life cases probably because decision attributes are not easily amenable to modeling. However, the same decision attributes should lead to the same predictive conclusion when training a reliable classifier [31]. Even if the evidence the research presented above shows that a higher IR leads to a lower PA, the continued gap between theory and practice still requires solutions to improve the merits of machine learning. To do so, programs of machine learning should manage theoretical improvement in general and practical feasibility in particular.

The research limitations are as follows. First, instead of presenting a novel granulation solution conducive to classification performance for specific datasets because a program machine learning usually involves various situations, the research from the experimental results revealed the effect of IR on PA using the same selected classifiers with the same combinations and settings on selected datasets. The findings were restricted to their interpretations within the defined context. Second, the datasets were collected from a public repository that is not frequently updated, so more updated datasets may be used in future studies to widen the testing scale. In addition, only two granulation techniques (EWI and MDL) were tested. Alternative dynamic and static, global and local, or even domain expertise involvement granulation techniques may produce different outcomes.

Third, the five classifiers including one tree-based and four ensemble-based were selected as the classification models for examining the effect of IR on PA. Although the research selected classifiers with different approaches such as bagging, boosting, and SGD, alternatives and other combinations and settings should be considered in future studies to extend the testing scale. Finally, despite the improvement of PA and correlation coefficient with IR being revealed most clearly at L1 than at other levels, the effect of IR on various aspects of machine learning, such as overfitting issues, selection and reduction of attribute dimensions, and prediction robustness, and reliability of classifiers as well, should be more thoroughly investigated in future studies.

6. Conclusion

The research has drawn the importance and existence of IR in granulation processes. An experimental analysis was conducted to reveal the effect of IR on classification performance. Both unsupervised (EWI) and supervised (MDL) granulation

techniques using five classifiers were used in 36 datasets. A dataset separation mechanism was created to detect IR and divide original granulated dataset into consistent subset and inconsistent one. Five classifiers were employed to derive PA for BG datasets, AGBR datasets, and AR datasets. The experimental results confirmed the research argument that IR influences the classification performance and PA can be generally improved by using the five consistency acceptance levels. It is suggested that granulated datasets be examined with consistency confirmation to properly manage inconsistent data when PA is concerned. In summary, the study contributes to a better understanding of the relationship between granulation inconsistency and classification performance in the domain of machine learning.

CRediT authorship contribution statement

ChienHsing Wu: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing.

Data availability

The authors do not have permission to share data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author would like to express appreciations to National Science and Technology Council, Taiwan for the partially financial support under the Grand 110-2410-H-390-001-MY2.

Indexing

Knowledge discovery, granulation, data inconsistency, prediction accuracy.

References

[1] M. Bello, G. Napoles, K. Vanhoof, R. Bello, Data quality measures based on granular computing for multi-label classification, Inf. Sci. 560 (2021) 51–67.
[2] L. Bottou, Large-scale learning with stochastic gradient descent, in: K.-.-R. Muller, G. Montavon, G.B. Orr (Eds.), Neural Networks: Tricks of the Trade, Springer, Reloaded, 2013.
[3] L. Breiman, Random forest, Mach. Learn. 45 (1) (2001) 5–32.
[4] H. Chen, T. Li, X. Fan, C. Luo, Feature selection for imbalanced data based on neighborhood rough sets, Inf. Sci. 483 (2019) 1–20.
[5] J. Dougherty, R. Kohavi, M. Sahami. Supervised and unsupervised discretization of continuous features, in: A. Prieditis, S. Russell (Eds.), in Proceedings of 1995 International Conference on Machine Learning, Morgan Kaufmann, Los Altos, CA, (1995) 194–202.
[6] Z. Duan, H. Zou, X. Min, S. Zhao, J. Chen, Y. Zhang, An adaptive granulation algorithm for community detection based on improved label propagation, Int. J. Approx. Reason. 114 (2019) 115–126.
[7] U.M. Fayyad, K.B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning, Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93), Chambery, 28 August-3 September 1993, 1022-1027.
[8] A. Fernandez, V. Lopez, M. Galar, M.J. del Jesus, F. Herrera, Analyzing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches, Knowl.-Based Syst. 42 (2013) 97–110.
[9] B. Franay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE Trans. Neural Networks Learn. Syst. 25 (2014) 845–869.
[10] Y. Freund, R. Schapire. Experiments with a new boosting algorithm, Machine Learning: Proceedings of the Thirteenth International Conference, July, 1996, 48–156.
[11] Grunwald, P. The Minimum Description Length Principle. 2007, MIT Press, Cambridge, MA.
[12] S. Kandanaarachchi, M.A. Munoz, R.J. Hyndman, K. Smith-Miles, On normalization and algorithm selection for unsupervised outlier detection, Data Min. Knowl. Disc. 34 (2020) 309–354.
[13] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy, Improvements to Platt's SMO algorithm for SVM classifier design, Neural Comput. 13 (3) (2001) 637–649.
[14] F.L. Liu, B.W. Zhang, D. Ciucci, W.Z. Wu, F. Min, A comparison study of similarity measures for covering-based neighborhood classifiers, Inf. Sci. 448–449 (2018) 1–17.
[15] M. Liu, F. Stella, A. Hommersom, P.J.F. Lucas, L. Boer, E. Bischoff, A comparison between discrete and continuous time Bayesian networks in learning from clinical time series data with irregularity, Artif. Intell. Med. 95 (2019) 104–117.
[16] N. Manwani, P.S. Sastry, Noise tolerance under risk minimization, IEEE Trans. Cybern. 43 (3) (2013) 1146–1151, https://doi.org/10.1109/TSMCB.2012.2223460.
[17] N. Natarajan, I.S. Dhillon, P.K. Ravikumar, A. Tewari, Learning with noisy labels, Adv. Neural Inf. Process. Syst. (2013) 1196–1204.
[18] D.F. Nettleton, A. Orriols-Puig, A. Fornells, A study of the effect of different types of noise on the precision of supervised learning techniques, Artif. Intell. Rev. 33 (4) (2010) 275–306.
[19] M. Nikolic, M. Bierlaire, Data-driven spatio-temporal discretization for pedestrian flow characterization, Transp. Res. Procedia 23 (2017) 188–207.
[20] W. Pedrycz, A dynamic data granulation through adjustable fuzzy clustering, Pattern Recogn. Lett. 29 (2008) 2059–2066.
[21] N. Peker, C. Kubat, Application of Chi-square discretization algorithms to ensemble classification methods, Expert Syst. Appl. 185 (2021), https://doi.org/10.1016/j.eswa.2021.115540.
[22] J.C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines, in Advances in Kernel Method: Support Vector Learning, Scholkopf, Burges, and Smola, Eds. Cambridge, MA: MIT Press, 1998, 185–208.
[23] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[24] S. Ramirez-Gallego, S. Garcia, F. Herrera, Online entropy-based discretization for data streaming classification, Fut. Gener. Comput. Syst. 86 (2018) 59–70.
[25] T.R. Rao, P. Mitra, R. Bhatt, A. Goswami, The big data system, components, tools, and technologies: a survey, Knowl. Inf. Syst. 60 (2019) 1165–1245, https://doi.org/10.1007/s10115-018-1248-0.
[26] M.A.N.D. Sewwandi, Y. Li, J. Zhang, Automated granule discovery in continuous data for feature selection, Inf. Sci. 578 (2021) 323–343.
[27] C.F. Tsai, Y.C. Chen, The optimal combination of feature selection and data discretization: an empirical study, Inf. Sci. 505 (2019) 282–293.
[28] UCI Machine Learning Repository, Accessed date: August, 2021, https://archive.ics.uci.edu/ml/index.php.
[29] J.L. Velazquez-Rodriguez, Y. Villuendas-Rey, C. Yanez-Marquez, I. Lopez-Yanez, O. Camacho-Nieto, Granulation in rough set theory: a novel perspective, Int. J. Approx. Reason. 124 (2020) 27–39.
[30] A.W. Wijayanto, J.J. Choong, K. Madhawa, T. Murata, Towards robust compressed convolutional neural networks, in: In 2019 IEEE International Conference on Big Data and Smart ComputIng (BigComp), IEEE, 2019, pp. 1–8.
[31] C.H. Wu, S.C. Kao, K. Okuhara, Examination and comparison of conflicting data in granulated datasets: Equal width interval vs. equal frequency interval, Inf. Sci. 239 (2013) 154–164.
[32] L. Yue, W. Chen, X. Li, W. Zuo, M. Yin, A survey of sentiment analysis in social media, Knowl. Inf. Syst. 60 (2019) 617–663.
[33] F. Zhang, S. Qi, Q. Liu, M. Mao, A. Zeng, Alleviating the data sparsity problem of recommender systems by clustering nodes in bipartite networks, Expert Syst. Appl. 149 (2020), https://doi.org/10.1016/j.eswa.2020.113346.
[34] Z. Zhao, L. Chu, D. Tao, J. Pei, Classification with label noise: a Markov chain sampling framework, Data Min. Knowl. Disc. 33 (2019) 1468–1504.