Reducing Training Spaces in Cluster-Based Data Balancing Ensemble Learning

Judhi Santoso[1], Lenny Putri Yulianti[2], Kridanto Surendro[3] and Agung Trisetyarso[4]

[1.2.3]School of Electrical Engineering & Informatics, Institut Teknologi Bandung, Bandung, Indonesia [4]Department of Mathematics, Bina Nusantara University, Jakarta, Indonesia
judhi@itb.ac.id, 33221007@std.stei.itb.ac.id, endro@itb.ac.id, atrisetyarso@binus.edu

Abstract: Ensemble learning incorporates the predictions of several learners to achieve better performance. The approach most widely used to build it is random subspace generation, where several well-known methods that use random subspace include bagging, boosting, clustering, and, more recently, clustering balancing. Clustering balancing has the potential to create a more accurate and diverse ensemble by generating a large pool of diverse subsets of data. However, involving all the balanced clusters produced for training learners carries the risk of creating similarly trained learners because there is a possibility of similar balanced clusters being produced through oversampling. This not only wastes computational time but also biases the ensemble towards their output. One strategy to handle this issue is by selecting the optimum set of balanced clusters with minimum similarity. Many such similarity indices were introduced to compare clusters from perturbed datasets, but there is a lack of comparison of these measures to select the optimum clusters. This study contributes to handling this issue with the aims to (i) analyze and identify the best similarity index and (ii) design a new approach to reducing the similarity of training spaces in cluster-based data balancing ensemble learning. The proposed approach using nine similarity indices was evaluated on 20 datasets from the UCI repository based on ensemble sizes and accuracy. The results show that the Normalized Mutual Information, Jaccard, and Van Dongen indices could serve as alternatives for measuring similarity in training spaces in ensemble learning. The Normalized Mutual Information employs a probabilistic approach, while the others utilize deterministic approach.

Keywords: Similarity; training space; ensemble learning; clustering balancing

1. Introduction

Ensemble learning combines the predictions of several individual learners, each trained to enhance accuracy and reduce generalization error [1], [2], [3], [4], [5]. Its superiority over individual learning methods lies in the fact that class decisions are not dependent solely on a single learner but on a diverse mixture of accurate learners [6], [7]. The two most popular types of ensemble methods are bagging [8] and boosting [9]. These methods use a traditional approach called “attribute bagging” [6], [10] to generate random subspaces from the input data. The underlying principle of random subspaces is that as learners are trained on distinct sets of input attributes, their ability for generalization becomes independent, enabling them to contribute unique strengths to the ensemble [6], [11].

Another approach involves creating perturbed subspaces using clustering techniques [6], [12]. Clustering is the process of organizing datasets into distinct subsets, It involves the process of organizing datasets into distinct subsets, where the data within the same subset exhibit greater similarity (cohesion) than those in different subsets (separation) [13]. One of the simplest, most renowned, and efficient clustering methods is the K-means algorithm, which is recognized as one of the top ten influential data mining algorithms [6], [14], [15]. The clustering process is not influenced by data classes, and most resulting clusters may have an unequal distribution of samples across data classes, resulting in unbalanced clusters [6], [16]. As a result, these data clusters may not effectively train base learners, resulting in biased outcomes. Hence, to achieve balanced clusters, it is essential to build a set of data clusters where ideally every cluster includes samples from all data classes within the dataset [6], [11],

Received: October 30th, 2023. Accepted: June 19th, 2024
DOI: 10.15676/ijeei.2024.16.2.5

[17], [18]. Balanced clusters play crucial role in attaining the intended performance [6], [17], [19], [20].

Several studies [6], [17], [18], [19] have proposed clustering balancing as a method for perturbing datasets in ensemble learning. There are two common methods for balancing clusters, namely under-sampling and over-sampling [6], [21]. One of the advantages of clustering balancing over other perturbation approaches is the creation of a wide range of diverse data subsets. Nonetheless, there is a potential drawback when employing all of them for learner training. When two clusters contain very similar records, the learners trained from those two clusters will also be very similar [22]. This not only consumes computational time but also results in the votes from these similar learners doubling up when predicting the final label, thereby biasing the ensemble toward their output [16], [22].

There are two common strategies to reduce the similarity between trained learners, which include selecting the optimum set of clusters and/or the optimum set of trained learners with minimum similarity. The first strategy focuses on detecting and pruning non-novel clusters to obtain the optimum set of unique clusters. Several studies [22], [23] have primarily used similarity indices to compare the generated clusters, but considering ensemble accuracy is also one of the cluster selection objectives in other research [16]. The optimum clusters are then used as the subspaces to train base learners. Meanwhile, the second strategy directly focuses on selecting the best set of trained learners with minimum similarity. Several studies have employed diversity measures [24], [25] and ensemble accuracy [16], [17] to select the best set of trained learners. Between these two strategies, this study found that the first strategy can reduce training time because the clusters used as training inputs have been optimized in such a way that they are not redundant. In contrast, the second strategy requires the training process to use all clusters, including redundant ones, before the selection process, which leads to longer training times. This prompted this study to further analyze the first strategy.

As mentioned before, similarity indices are mostly used to compare clusters from perturbed datasets [26]. They are used to evaluate the clustering structure concerning factors such as cluster size, dimensionality, and the quantity of clusters. Several of these indices have emerged throughout the years [26], [27], [28]. Despite the existence of numerous related works on similarity measures for cluster-based ensemble methods using clustering balancing, there remains a need for a comparative evaluation of these measures to select the optimum clusters. This study endeavors to rectify this matter, pursuing two key objectives: (i) analyzing and identifying the best similarity index for the clustering balancing method and (ii) designing a new approach to reduce the similarity of training spaces in cluster-based data balancing ensemble learning.

The remainder of the paper is structured in the following manner: Section 2 explains several similarity indices for cluster-based training spaces. Section 3 outlines the proposed methodology. Section 4 discusses the experimental findings. Section 5 concludes and outlines future research.

2. Related Works

This section outlines the related work about clustering balancing methods for generating ensemble learning and how to measure similarity of training spaces in clustering balancing.

A. Clustering Balancing Method

Clustering balancing is one of the ensemble methods that addressing class imbalance within datasets by ensuring equitable representation of each class in the clustering process. This typically involves either under-sampling, where a substance of instances from the majority class is selected, or over-sampling, where instances from the minority class are replicated to match the size of the majority class [6], [21]. The objective is to create balanced clusters that adequately represent all classes in the dataset. These balanced clusters are then utilized to train classifiers, promoting diversity in ensemble classifiers, and ultimately improving overall accuracy. The illustration of clustering balancing method is presented in Figure 1.

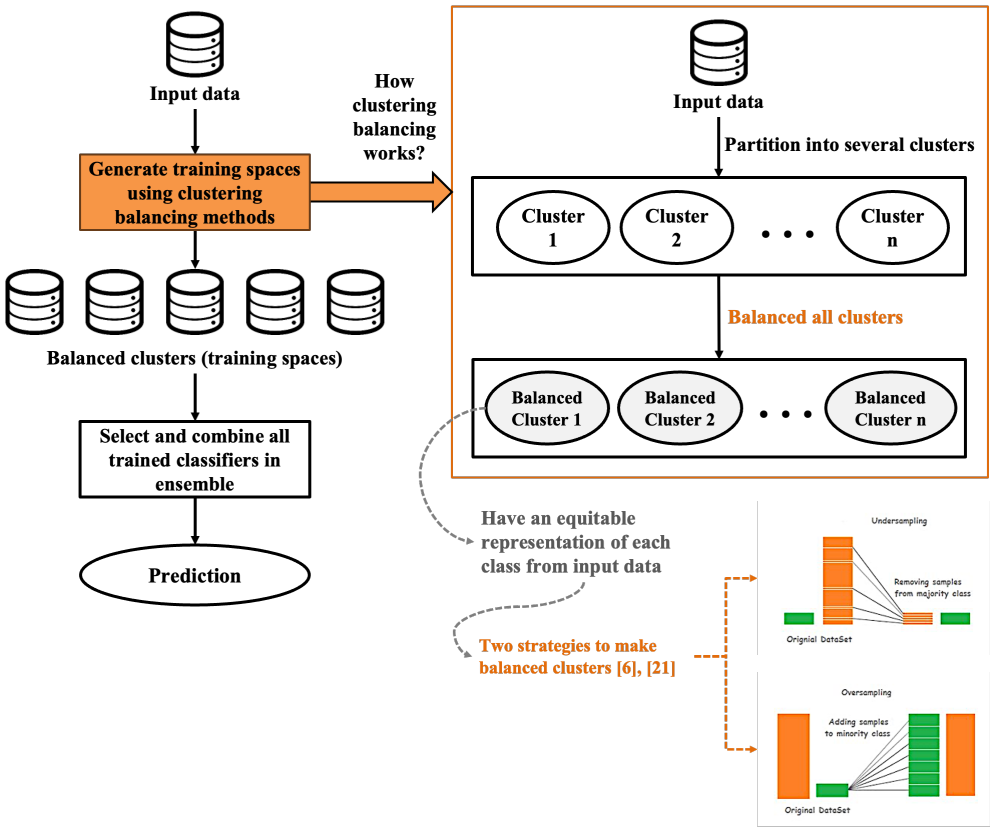

Figure 1. Clustering balancing method

Several studies [6], [11], [17], [18], [19], [29], [30] have introduced methods for balancing clustering to construct ensemble models. Some research [6], [19] employed under-sampling to create balanced clusters within the ensemble's training spaces. Additionally, [18] utilized over-sampling specifically within pure-class clusters (clusters comprising only one class) by minimizing the squared Euclidean distance between clusters. Meanwhile, [11], [17], [29] refined this approach by adopting normalized squared Euclidean distance, citing its scale invariance, normalization, and consistency in comparing distances within feature space, which could enhance over-sampling effectiveness. However, [11] relied solely on one base classifier for training balanced clusters, whereas [17], [29] utilized multiple base classifiers. Moreover, another study [30] generated balanced clusters through over-sampling, optimizing the quality of pure-class clusters through quantum annealing. Nonetheless, a significant limitation remains regarding the scalability of implemented data.

Given the potential issue of generating highly similar balanced clusters [16], [22], several studies [11], [17], [18], [29] proposed pruning methods to refine the ensemble. Most studies [11], [17], [18] focused on pruning trained classifiers using particle swarm optimizations, with the objective of maximizing the accuracy of the pruned ensembles. Meanwhile, one study [29] explored pruning training spaces using simulated annealing, seeking to minimize similarity based on the Jaccard index between balanced clusters. These pruning methods were found to enhance the accuracy and/or diversity of the ensemble.

B. Measuring Similarity of Training Spaces

Similarity measures are used to quantify the similarity between objects [31]. In the context of measurement, a similarity measure typically yields a numerical value falling between 0 and

1, with 0 indicating complete dissimilarity and 1 representing complete similarity [31]. Common similarity metrics can typically be classified into three main categories: i) pair-counting measures, ii) information-theoretic measures, and iii) set-matching measures [13], [27], [28].

Pair-counting measures are calculated by assessing the extent to which the memberships of data points in clusters and classes agree or disagree. Prominent metrics in this group encompass the Rand Index, Adjusted Rand Index, and Jaccard Index. Information-theoretic measures are constructed around various entropic measures derived from information theory. Well-recognized metrics in this category encompass mutual information and Normalized Mutual Information. Set-matching measures quantify the level of similarity between clusters in two partitions using set-theoretic metrics. Noteworthy metrics in this category encompass the van Dongen criterion, the H criterion, and the L criterion.

An appropriate similarity (or distance) measure is determined by the following criteria: it should (i) exhibit symmetry, (ii) meet the requirements of the triangle inequality, and (iii) adhere to the similarity property [31], [32]. The first criterion is met when the distance between objects A and B is identical to that between objects B and A. Furthermore, the second criterion, the triangle inequality, indicates that the sum of the distances from A to C and from C to B should be equal to or greater than the 'direct' distance between A and B. Meanwhile, the similarity property specifies that when the objects are the same, their distance should equal 0. This study also maps some similarity properties from the work of [27] into the three aforementioned properties, as presented in Table 1.

Table 1. Similarity criteria [27], [31]

Properties [27]	Symmetry	Triangle Inequality	Similarity
Symmetry	v		
Distance	v	v	v
Maximal agreement			v
Monotonicity			v
Linear complexity			v
Constant baseline			v

Table 1 shows that all the properties introduced in the work of [27] align with the three general criteria from [31], [32]. This motivated our study to utilize all these properties for the analysis of several similarity indices from three measurement categories, as presented in Table 2. A total of seven pair-counting measures, three information-theoretic measures, and one set-matching measure were selected for further analysis and their expressions can be referred to [13], [27], [33], [34]. The choice of similarity indices is based on their frequent usage or recommendations in previous research. Additionally, this study also assesses how each similarity index satisfies the six similarity properties outlined in Table 1, as presented in Table 3. It reveals that none of the similarity indices can fulfill all of the similarity properties. In fact, some indices only satisfy one or two properties. The diversity within each index category, even when they belong to the same category, presents intriguing opportunities to compare their performance in measuring similarity within cluster-based data balancing ensembles.

Based on these literature reviews, this study observed that existing similarity indices can be employed to measure the similarity of training spaces in ensemble learning. Given the multitude of available indices, this study proposes a comparative analysis to identify the most suitable similarity index for ensembles, particularly when employing the clustering balancing method. To streamline the analysis, this study selects similarity indices that satisfy many of the properties. Consequently, two indices, Wallace (W) and Standardized Mutual Information (SMI), are excluded from further analysis as they only fulfill one and two similarity properties, respectively. Details on how to determine the suitability of an index for building an effective ensemble and the methodology for comparison and ensemble construction are provided in Section 3.

3. Methodology

This study employed clustering balancing to generate a pool of balanced clusters, as referenced in [17]. The primary contribution of this study is to improve ensemble accuracy and size by reducing the similarity among balanced clusters, as this can influence bias and redundancy in the aggregation of predictions from trained learners. To achieve these objectives, this study proposes a methodology that utilizes similarity indices and optimization method. The similarity index is designed to quantify the similarity between balanced clusters, serving as a parameter for ensemble reduction. The search process to identify the optimal set of clusters with minimal similarity is conducted using optimization methods, with simulated annealing selected for its capacity to find the global optimum rather than local optimum [35]. The proposed methodology is depicted in Figure 2.

Table 2. Similarity indices [13], [27], [33], [34]

Similarity Indices	Categories
Rand (R)	Pair-counting
Adjusted Rand (AR)	Pair-counting
Jaccard (J)	Pair-counting
Wallace (W)	Pair-counting
Pearson Correlation Coefficient (CC)	Pair-counting
Sokal & Sneath (SS)	Pair-counting
Lopez-Rajski (LR)	Information-theoretic
Normalized Mutual Information (NMI)	Information-theoretic
Adjusted Mutual Information (AMI)	Information-theoretic
Standardized Mutual Information (SMI)	Information-theoretic
Van Dongen (VD)	Set-matching

Table 3. The properties of similarity indices [13], [27], [33], [34]

Properties	Similarity Indices										
	R	AR	J	W	CC	S&S	LR	NMI	AMI	SMI	VD
Symmetry	v	v	v	x	v	v	v	v	v	v	v
Distance	v	x	v	x	x	x	x	x	x	x	x
Maximal agreement	v	v	v	x	v	v	v	v	v	x	v
Monotonicity	v	v	v	x	v	v	v	v	v	x	v
Linear complexity	v	v	v	v	v	v	v	v	x	x	x
Constant baseline	x	v	x	x	v	v	x	x	v	v	v

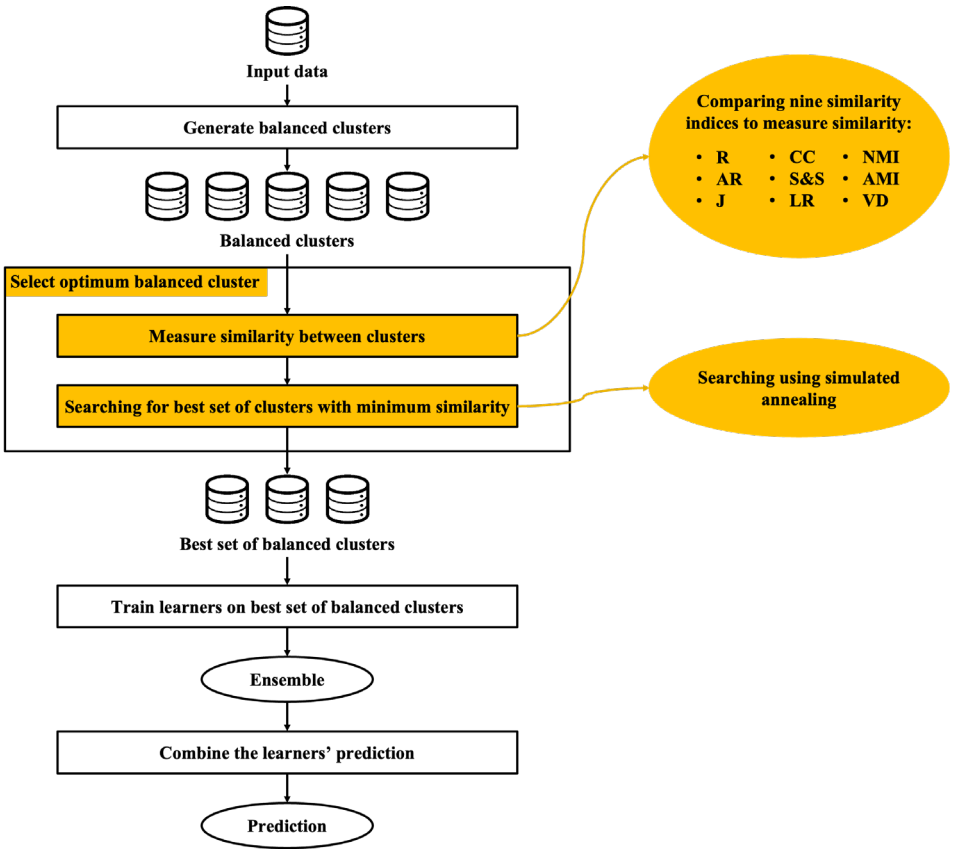

Figure 2. Methodology

From Figure 2, the selection process involves two primary activities: measuring the similarity between clusters serving as training spaces and searching for the optimal set of clusters with minimum similarity. The first activity focuses on quantifying the similarity between two clusters within the pool of balanced clusters. This study utilizes similarity indices because clusters are typically formed based on the similarity measures between patterns [36]. The concept involves creating an NxN matrix if N balanced clusters are generated, which stores all data similarities between clusters. Since nine similarity indices are used, nine matrices are generated in this activity. These matrices function as pools, and the search process, employing simulated annealing, identifies which combinations of clusters can form training spaces with low similarity from each pool. To implement this, the first step is defining the objective function. This study proposes minimizing the similarity between selected clusters, expressed as follows:

$$\min_{\hat{w}} \hat{w}'(Similarity)\hat{w}'^{T} \tag{1}$$

where $\hat{w}'$ represents the selected clusters and Similarity is the general variable representing the similarity matrix derived from the nine indices used in the previous activity. Additionally, this study introduced a constraint to guarantee the selection of at least one cluster, ensuring that the selected training space is not empty. This constraint can be expressed as follows:

$$\alpha\left(1 - \sum_{i=1}^{N} \hat{w}'_i\right) \tag{2}$$

where α represents the penalty value of the constraints, and in this study, α is defined as 1. An example demonstrating the application of the formulas is presented in Figure 3.

Consider a scenario with three clusters (A, B, C) and the following similarity matrix:

	A	B	C
A	1	0.8	0.2
B	0.8	1	0.6
C	0.2	0.6	1

if $\widehat{w}'$=[1,0,1] (cluster A and C are selected), the objective function:

$$\widehat{w}'\,(Similarity)\widehat{w}'^{T} = [1 \quad 0 \quad 1]\begin{bmatrix} 1 & 0.8 & 0.2 \\ 0.8 & 1 & 0.6 \\ 0.2 & 0.6 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \frac{12}{5}$$

Meanwhile, the constraint is -1.

Similarly, perform calculations of the objective function value and constraint for various possible cluster selection scenarios $\widehat{w}'$, resulting in a total of 8 potential solutions.

Subsequently, choose the solution with the minimum value.

Figure 3. Example illustrating formula application in cluster selection

The implementation of simulated annealing employed the Simulated Annealing sampler from D-Wave. To embed the objective function and constraint into the solver, they needed to be formulated in Quadratic Unconstrained Binary Optimization (QUBO). This study utilized the PyQUBO library to generate the QUBO from equations (1) and (2). The QUBO formulation was then embedded and annealed using the D-Wave solver. The best set of clusters was determined based on the minimum solution energy produced during the annealing process. It's important to note that the resulting cluster combinations between one index and another index can differ due to variations in parameters and methods for measuring similarity. This study analyzes how these differences impact ensemble size and accuracy.

4. Results and Discussion

This section outlines the experimental settings and presents the evaluation results of the proposed methodology. It involves a comparison of nine similarity indices within the proposed methodology and assesses them alongside benchmark methods that employ particle swarm optimization (PSO) for pruning trained learners, as described in [17]. The analysis primarily focuses on ensemble size and accuracy.

A. Datasets and Experimental Settings

This study utilized 20 benchmark datasets from UCI repository, chosen for their diversity in terms of attributes, instances, and classes. Table 4 provides an overview of the dataset characteristics. Each dataset served as input data for the proposed methodology. The proposed methodology utilized several base learners to train the best set of balanced clusters generated from perturbed datasets. The base learners included artificial neural networks, support vector machines, k-nearest neighbors, naïve Bayes, linear discriminant analysis, and decision trees. The selection of these base learners was driven by the understanding that their distinct learning strategies and structures could collectively enhance the model's performance through increased diversity. The parameters were adopted from [17] with some modifications made to the neural network and linear discriminant. Specifically, the neural network employed three solvers: lbfgs, sgd, and adam from the Scikit-Learn library, with a randomized number of epochs ranging between 500 and 1000. Conversely, the linear discriminant used default parameters.

B. Ensemble Sizes

The ensemble sizes represent the number of trained learners used to construct an ensemble model. These trained learners were obtained by training m base learners within the n selected training spaces. In other words, the ensemble size is the result of multiplying m (the number of base learners) by n (the number of selected training spaces). The ensemble sizes generated for each method and dataset are presented in Figure 4.

Figure 4 demonstrates that the proposed pruning method effectively reduces the ensemble size by eliminating similar training spaces. Several similarity indices, including R, S&S, and VD, also exhibit a significant reduction in ensemble size compared to the benchmark method. However, others only show significant reductions in six datasets: cryotherapy, divorce, e-coli, Parkinson's, Somerville, and sonar. This variation is influenced by the cluster sizes generated using the clustering balancing method. Specifically, the more clusters that are produced, the more effective the proposed pruning method is in reducing the number of clusters and, consequently, the ensemble size. This study revealed that the proposed pruning method can significantly reduce the ensemble size when the cluster sizes generated exceed 11 clusters.

Another interesting observation is made in the case of the 'segment' dataset, where a larger number of clusters does not lead to better reduction in ensemble size with the proposed method compared to the benchmark method. This can be attributed to the dataset's high number of classes, which results in the generation of a larger number of pure-class clusters. Additionally, the over-sampling process tends not to produce redundant clusters due to the substantial amount of data used as input for over-sampling. Consequently, the proposed method is less effective in reducing ensemble size when cluster similarity tends to be low.

Table 4. Datasets

Dataset	Number of Instances	Number of Attributes	Number of Class
Banknote	1372	4	2
Bone marrow	187	38	2
Ca-cervix	72	19	2
Coimbra	116	9	2
Cryotherapy	90	6	2
Diabetes	768	8	2
Divorce	170	54	2
E-coli	336	7	8
Haberman	306	3	2
Hayes-roth	132	5	3
Hepatitis	80	18	2
Ionosphere	351	34	2
Iris	150	4	3
Parkinson’s	195	22	2
Segment	2310	18	7
Somerville	143	6	2
Sonar	208	60	2
Spambase	4601	57	2
Vehicle	846	18	4
Wine	178	13	3

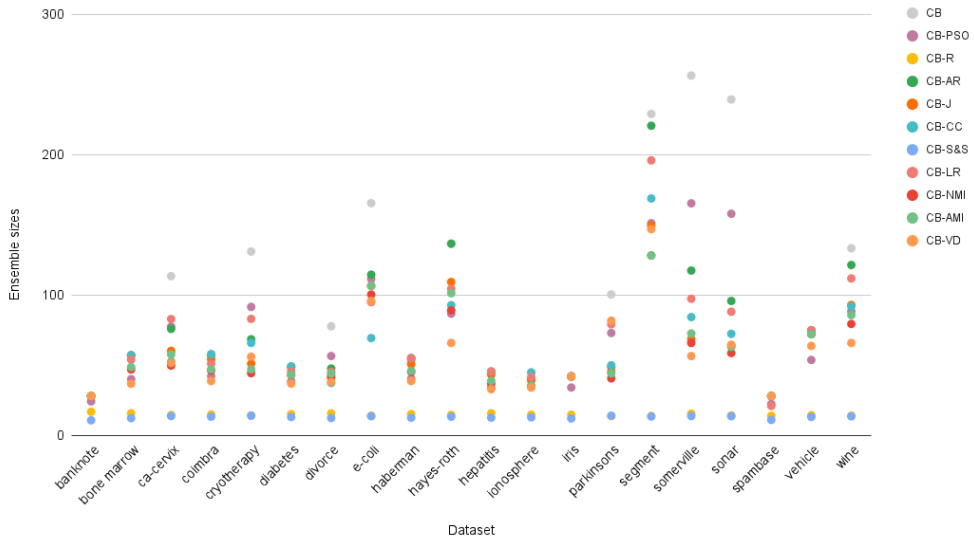

Figure 4. Ensemble sizes

C. Accuracy

Reducing ensemble size should not compromise ensemble accuracy; in fact, it should enhance it. Table 5 presents the average accuracies achieved using nine similarity indices in the proposed methodology across 20 datasets. The highest accuracies obtained for each dataset are highlighted in bold and orange. This study found that NMI and VD consistently delivered the best accuracy in 7 out of 20 datasets. They were followed by R, which outperformed others in 4 out of 20 datasets. On the other hand, J, CC, LR, S&S, and AMI exhibited superior accuracy in only one dataset each. Interestingly, AR did not outperform the other indices in any of the datasets. Additionally, it was observed from Table 5, it is also observed that information-theoretic index methods are generally suitable for large datasets. Although LR and AMI only excelled in one dataset, both datasets were large. This indicates that a probabilistic approach to assessing the extent of information sharing between two clusters is effective. Even LR, which is based on conditional entropy, has a positive effect on datasets with the largest sizes and attributes. Meanwhile, pair-counting index methods are generally suitable for small to medium-sized datasets due to the simplicity of their measurement processes.

Furthermore, it's noteworthy that although both NMI and VD outperformed others in more datasets, NMI boasted the highest average accuracy at 66.60%, followed by AMI, J, VD, AR, LR, CC, S&S, and R with 66.02%, 65.62%, 65.37%, 65.23%, 64.73%, 64.71%, 63.10%, and 62.40% respectively. It's evident that NMI and AMI, both belonging to the IT category of indices, consistently delivered better accuracy than the others. Interestingly, although J and AMI only outperformed others in one dataset each, their overall average accuracy ranks second and third highest, surpassing VD. This suggests that J and AMI can remain competitive even though they do not excel in many specific datasets.

This study also compared the highest results from Table 5 with clustering balancing and the benchmark pruning method using clustering balancing [17], as shown in Table 6. The results indicated that the two benchmark methods could not outperform the proposed method. This highlights the significance of the selection of the best set of training spaces in enhancing accuracy in ensemble learning. Furthermore, this study grouped the accuracy scores obtained from the nine similarity indices into pair-counting, information-theoretic, and set-matching indices (see Figure 5). These indices were averaged within their respective groups and compared with the benchmark methods.

Table 5. The average accuracy using nine similarity indices

Dataset	R	AR	J	CC	S&S	LR	NMI	AMI	VD
Banknote	0.955	0.989	0.989	0.990	0.928	0.989	0.990	**0.991**	0.989
Bone marrow	0.550	0.659	0.658	0.627	0.590	0.645	0.655	0.680	**0.718**
Ca-cervix	0.673	0.729	0.753	0.725	0.634	0.657	**0.771**	0.757	0.630
Coimbra	0.544	0.584	0.586	0.580	0.568	0.593	0.597	0.584	**0.611**
Cryotherapy	**0.494**	0.461	0.458	0.467	0.467	0.467	0.467	0.456	0.461
Diabetes	0.412	0.460	0.453	0.452	0.470	0.479	0.478	0.480	**0.528**
Divorce	**0.978**	0.976	0.976	**0.978**	0.972	0.974	0.976	0.976	0.974
E-coli	0.802	0.839	0.840	0.812	0.788	0.801	**0.842**	0.837	0.802
Haberman	0.504	0.441	0.445	0.440	0.530	0.438	0.500	0.483	**0.546**
Hayes-roth	0.421	0.407	0.412	0.404	0.408	0.395	**0.464**	0.422	0.373
Hepatitis	0.472	0.441	0.453	0.441	**0.478**	0.450	0.459	0.453	**0.478**
Ionosphere	0.507	0.589	0.574	0.551	0.549	0.574	**0.601**	0.599	0.591
Iris	0.908	0.970	0.968	0.970	0.897	0.970	**0.972**	0.968	**0.972**
Parkinsons	**0.760**	0.755	0.755	0.755	0.757	0.755	0.755	0.755	0.755
Segment	0.853	0.925	0.917	0.920	0.845	0.921	0.921	0.918	**0.932**
Somerville	0.440	0.430	0.449	0.442	0.439	0.418	**0.454**	0.422	0.421
Sonar	**0.501**	0.469	0.489	0.467	0.482	0.468	0.498	0.482	0.469
Spambase	0.504	0.590	0.591	0.594	0.632	**0.694**	0.593	0.594	0.595
Vehicle	0.531	0.576	0.583	0.576	0.523	0.574	**0.587**	0.585	0.569
Wine	0.669	0.753	**0.773**	0.751	0.662	0.682	0.740	0.762	0.656

Table 6. The average accuracy of clustering balancing (CB), PSO, and proposed method

Dataset	CB	PSO	Proposed Similarity Indices							
			R	J	CC	S&S	LR	NMI	AMI	VD
Banknote	0.988	0.987	0.955	0.989	0.990	0.928	0.989	0.990	**0.991**	0.989
Bone marrow	0.642	0.708	0.550	0.658	0.627	0.590	0.645	0.655	0.680	**0.718**
Ca-cervix	0.658	0.689	0.673	0.753	0.725	0.634	0.657	**0.771**	0.757	0.630
Coimbra	0.584	0.595	0.544	0.586	0.580	0.568	0.593	0.597	0.584	**0.611**
Cryotherapy	0.467	0.461	**0.494**	0.458	0.467	0.467	0.467	0.467	0.456	0.461
Diabetes	0.453	0.464	0.412	0.453	0.452	0.470	0.479	0.478	0.480	**0.528**
Divorce	0.976	0.976	**0.978**	0.976	**0.978**	0.972	0.974	0.976	0.976	0.974
E-coli	0.817	0.815	0.802	0.840	0.812	0.788	0.801	**0.842**	0.837	0.802
Haberman	0.440	0.503	0.504	0.445	0.440	0.530	0.438	0.500	0.483	**0.546**
Hayes-roth	0.412	0.422	0.421	0.412	0.404	0.408	0.395	**0.464**	0.422	0.373
Hepatitis	0.441	0.472	0.472	0.453	0.441	**0.478**	0.450	0.459	0.453	**0.478**
Ionosphere	0.557	0.587	0.507	0.574	0.551	0.549	0.574	**0.601**	0.599	0.591
Iris	0.967	0.968	0.908	0.968	0.970	0.897	0.970	**0.972**	0.968	**0.972**
Parkinsons	0.755	0.755	**0.760**	0.755	0.755	0.757	0.755	0.755	0.755	0.755
Segment	0.923	0.929	0.853	0.917	0.920	0.845	0.921	0.921	0.918	**0.932**
Somerville	0.432	0.440	0.440	0.449	0.442	0.439	0.418	**0.454**	0.422	0.421
Sonar	0.467	0.467	**0.501**	0.489	0.467	0.482	0.468	0.498	0.482	0.469
Spambase	0.593	0.678	0.504	0.591	0.594	0.632	**0.694**	0.593	0.594	0.595
Vehicle	0.570	0.577	0.531	0.583	0.576	0.523	0.574	**0.587**	0.585	0.569
Wine	0.716	0.769	0.669	**0.773**	0.751	0.662	0.682	0.740	0.762	0.656

The findings indicate that the information-theoretic index consistently yielded the highest accuracy across most datasets, surpassing the other two indices. This study delves into how different measurement approaches influence these variations in results. The information-

theoretic index employs a probabilistic approach to measure similarity, offering flexibility and sensitivity to the complex structure of training spaces in ensemble learning. On the other hand, the set-matching index, while capable of competing with the information-theoretic index, employs additional mathematical operations beyond pair counting, making it a potential alternative for deterministic similarity measurement.

Furthermore, this study also analyzes the correlation between accuracy and ensemble size, as illustrated in Figure 6. The X-axis represents the methods, while the Y-axis indicates the values of accuracy and ensemble size, normalized to a range of 0 to 1. In Figure 6, two lines are depicted: the blue line represents the accuracy values of each method, and the orange line represents the ensemble size values of each method. The results indicate that J and CC exhibit the most significant decrease in ensemble size. Interestingly, when considering average accuracy, J ranks third highest, while CC ranks third lowest among other indices. Meanwhile, NMI, with the highest average accuracy, has the third highest average ensemble size compared to other indices.

Comparison with the baseline method (CB) reveals a significant decrease in ensemble size across all indices, with almost all methods showing an increase in accuracy. This suggests that the proposed pruning of training spaces using various similarity indices has a positive impact on both accuracy and ensemble size. However, when comparing nine similarity indices, some indices (J and CC) contribute significantly to the reduction in ensemble size, while others (NMI and AMI) prioritize increasing accuracy, albeit with less significant changes in ensemble size. This indicates that a significant reduction in ensemble size can adversely affect ensemble accuracy, making it less optimal. Meanwhile, achieving optimal ensemble accuracy tends not to lead to a significant decrease in ensemble size.

5. Conclusion

Ensemble learning incorporates the predictions of numerous learners to achieve better performance. The most widely used approach to build it is random subspace generation, which includes several well-known methods such as bagging, boosting, clustering, and more recently, clustering balancing. Clustering balancing has the potential to create a more accurate and diverse ensemble by generating a wide range of diverse data subsets. However, utilizing all the balanced clusters produced for training learners carries the risk of creating similar trained learners, as similar balanced clusters may result from over-sampling. This not only consumes computational resources but also biases the ensemble toward their output.

One strategy to address this issue is to select the optimal set of balanced clusters with minimal similarity. While numerous similarity indices have been introduced to compare clusters from perturbed datasets, there is an absence of comparative analysis among these measures to select the best clusters. This study aims to address this issue by (i) analyzing and identifying the best similarity index and (ii) designing a new approach to reduce the similarity of training spaces in cluster-based data balancing ensemble learning.

The proposed approach, which employs nine similarity indices, was evaluated across 20 datasets from the UCI repository based on ensemble sizes and accuracy. Engaging diverse datasets aimed to examine the method's consistency across various scenarios and gain deeper insights into how similarity indices behave and perform under different dataset conditions. The results revealed that the R, S&S, and VD indices significantly reduced ensemble size compared to other indices and benchmark methods in various datasets. However, upon examining the overall averages, it was found that J and CC provided the best average reduction in ensemble size. Interestingly, these five indices all utilize deterministic measurements (pair-counting and set-matching indices). Moreover, in terms of accuracy, NMI and VD outperformed the others in most datasets, with NMI exhibiting the highest average accuracy. When considering the overall average accuracy, AMI and J were found to be viable alternatives, ranking second and third highest, respectively.

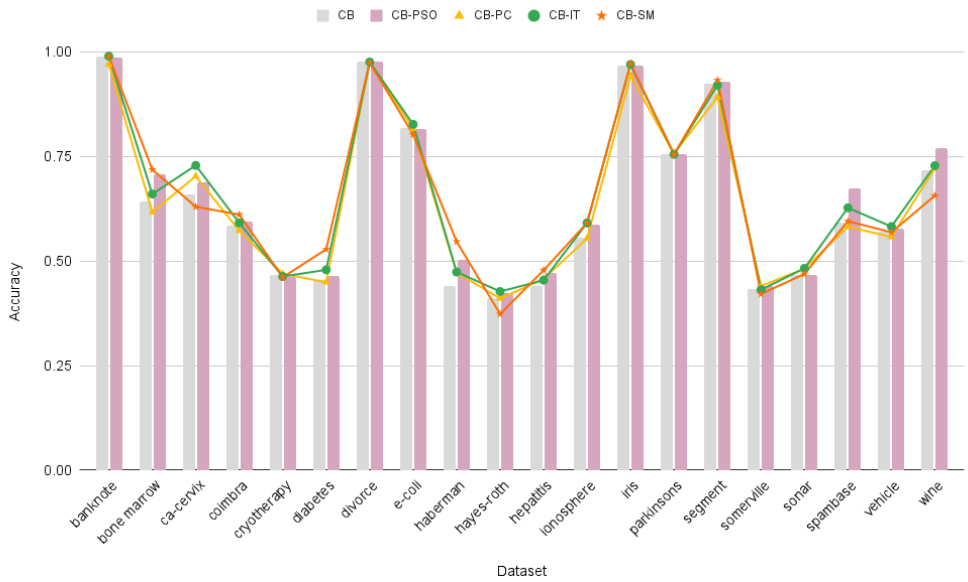

Figure 5. The average accuracy of three categories of similarity indices

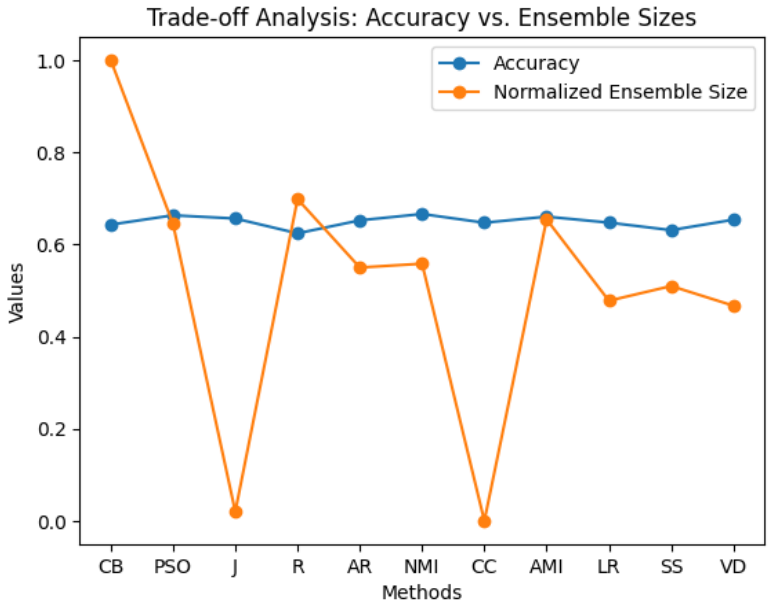

Figure 6. The correlation between ensemble accuracy and sizes

In conclusion, this study found that when focusing on accuracy, NMI emerges as the best alternative compared to other methods. However, J and VD can also serve as alternative methods for measuring similarity in training spaces in ensemble learning, where J excels in terms of overall averages and VD for specific datasets. For specific problem domains with moderate-sized data, the combination of NMI and J proves to be a suitable solution, whereas for larger datasets, the combination of NMI and VD could be an alternative. NMI represents a probabilistic approach, while J and VD represent a deterministic one, allowing for a combination of methods to measure similarity between training spaces both probabilistically and deterministically.

Moreover, the exploration of the proposed pruning method using various similarity indices in specific real-world problems presents an intriguing avenue for further research. This approach offers unique opportunities to assess the method's efficacy and adaptability in addressing the complexities and nuances of real-world datasets. Additionally, the implementations can gain deeper insights into the method's performance under diverse conditions, paving the way for targeted enhancements and optimizations.

6. Acknowledgment

This study is supported by Institut Teknologi Bandung.

7. References

[1] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. New York, USA: Springer, 2012.

[2] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. London, New York: CRC Press, 2012.

[3] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*, Cagliari, Italy: Springer, 2000, pp. 1–15.

[4] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pp. 1-18, 2018.

[5] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Computing Surveys*, vol. 45, no. 1, pp. 1-40, 2012.

[6] S. A. Mousavian Anaraki, A. Haeri, and F. Moslehi, "Generating balanced and strong clusters based on balance-constrained clustering approach (strong balance-constrained clustering) for improving ensemble classifier performance," *Neural Comput & Applic*, vol. 34, no. 23, pp. 21139–21155, Dec. 2022, doi: 10.1007/s00521-022-07595-6.

[7] L. P. Yulianti, J. Santoso, A. Trisetyarso, and K. Surendro, "Hybrid Classical-Quantum Optimization for Ensemble Learning," in *2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, Tokoname, Japan: IEEE, Sep. 2022, pp. 1–6. doi: 10.1109/ICAICTA56449.2022.9932950.

[8] L. Breiman, "Bagging predictors," *Mach Learn*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.

[9] Y. Freund, "Boosting a Weak Learning Algorithm by Majority," *Information and Computation*, vol. 121, no. 2, pp. 256–285, Sep. 1995, doi: 10.1006/inco.1995.1136.

[10] P. Bühlmann, "Bagging, Boosting and Ensemble Methods," in *Handbook of Computational Statistics: Concepts and Methods*, London, New York: Springer, 2012, pp. 985–1022.

[11] Z. Jan and B. Verma, "Multiple strong and balanced cluster-based ensemble of deep learners," *Pattern Recognition*, vol. 107, p. 107420, Nov. 2020, doi: 10.1016/j.patcog.2020.107420.

[12] L. P. Yulianti, A. Trisetyarso, J. Santoso, and K. Surendro, "Comparison of Distance Metrics for Generating Cluster-based Ensemble Learning," in *Proceedings of the 2023 12th International Conference on Software and Computer Applications*, Kuantan Malaysia: ACM, Feb. 2023, pp. 26–33. doi: 10.1145/3587828.3587833.

[13] D. Pfitzner, R. Leibbrandt, and D. Powers, "Characterization and evaluation of similarity measures for pairs of clusterings," *Knowl Inf Syst*, vol. 19, no. 3, pp. 361–394, Jun. 2009, doi: 10.1007/s10115-008-0150-6.

[14] M. K. Gupta and P. Chandra, "An Empirical Evaluation of K-Means Clustering Algorithm Using Different Distance/Similarity Metrics," in *Proceedings of ICETIT 2019*, vol. 605, P. K. Singh, B. K. Panigrahi, N. K. Suryadevara, S. K. Sharma, and A. P. Singh, Eds., in Lecture Notes in Electrical Engineering, vol. 605. , Cham: Springer International Publishing, 2020, pp. 884–892. doi: 10.1007/978-3-030-30577-2_79.

[15] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[16] Z. Md. Jan and B. Verma, "Evolutionary Classifier and Cluster Selection Approach for Ensemble Classification," *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 1, pp. 1–18, Feb. 2020, doi: 10.1145/3366633.

[17] Z. Jan, J. Munoz, and A. Ali, "A Novel Method for Creating an Optimized Ensemble Classifier by Introducing Cluster Size Reduction and Diversity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 7, pp. 3072-3081, 2022.

[18] Z. Jan and B. Verma, "Ensemble Classifier Generation Using Class-Pure Cluster Balancing," in *Neural Information Processing*, vol. 1143, T. Gedeon, K. W. Wong, and M. Lee, Eds., in Communications in Computer and Information Science, vol. 1143. , Cham: Springer International Publishing, 2019, pp. 761–769. doi: 10.1007/978-3-030-36802-9_80.

[19] M. Amini, J. Rezaeenour, and E. Hadavandi, "A Cluster-Based Data Balancing Ensemble Classifier for Response Modeling in Bank Direct Marketing," *Int. J. Comp. Intel. Appl.*, vol. 14, no. 04, p. 1550022, Dec. 2015, doi: 10.1142/S1469026815500224.

[20] M. Asafuddoula, B. Verma, and M. Zhang, "An incremental ensemble classifier learning by means of a rule-based accuracy and diversity comparison," in *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, USA: IEEE, May 2017, pp. 1924–1931. doi: 10.1109/IJCNN.2017.7966086.

[21] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan: IEEE, Apr. 2020, pp. 243–248. doi: 10.1109/ICICS49469.2020.239556.

[22] S. Fletcher and B. Verma, "Pruning High-Similarity Clusters to Optimize Data Diversity when Building Ensemble Classifiers," *Int. J. Comp. Intel. Appl.*, vol. 18, no. 04, p. 1950027, Dec. 2019, doi: 10.1142/S1469026819500275.

[23] H. Khalili, M. Rabbani, and E. Akbari, "Clustering ensemble selection based on the extended Jaccard measure," *Turk J Elec Eng & Comp Sci*, vol. 29, no. 4, pp. 2215–2231, Jul. 2021, doi: 10.3906/elk-2010-91.

[24] Q. Dai, R. Ye, and Z. Liu, "Considering diversity and accuracy simultaneously for ensemble pruning," *Applied Soft Computing*, vol. 58, pp. 75-91, 2017.

[25] A. M. Mohammed, E. Onieva, M. Woźniak, and G. Martínez-Muñoz, "An analysis of heuristic metrics for classifier ensemble pruning based on ordered aggregation," *Pattern Recognition*, vol. 124, p. 108493, Apr. 2022, doi: 10.1016/j.patcog.2021.108493.

[26] A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko, "On Similarity Indices and Correction for Chance Agreement," *Journal of Classification*, vol. 23, no. 2, pp. 301–313, Sep. 2006, doi: 10.1007/s00357-006-0017-z.

[27] M. Gösgens, A. Tikhonov, and L. Prokhorenkova, "Systematic Analysis of Cluster Similarity Indices: How to Validate Validation Measures," in *Proceedings of the 38 th International Conference on Machine Learning*, Virtual Event: PMLR, 2021, pp. 3799–3808.

[28] S. Zhang, Z. Yang, X. Xing, Y. Gao, D. Xie, and H.-S. Wong, "Generalized Pair-Counting Similarity Measures for Clustering and Cluster Ensembles," *IEEE Access*, vol. 5, pp. 16904–16918, 2017, doi: 10.1109/ACCESS.2017.2741221.

[29] L. P. Yulianti, A. Trisetyarso, J. Santoso, and K. Surendro, "Annealing-Based Optimization for Selecting Training Space in Ensemble Learning," in *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, Lombok, Indonesia: IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICAICTA59291.2023.10390142.

[30] L. P. Yulianti, A. Trisetyarso, J. Santoso, and K. Surendro, "A hybrid quantum annealing method for generating ensemble classifiers," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 10, p. 101831, Dec. 2023, doi: 10.1016/j.jksuci.2023.101831.

[31] V. U. Thompson, C. Panchev, and M. Michael Oakes, “Performance Evaluation of Similarity Measures on Similar and Dissimilar Text Retrieval:,” in *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Lisbon, Portugal: SCITEPRESS - Science and and Technology Publications, 2015, pp. 577–584. doi: 10.5220/0005619105770584.
[32] M. P. Oakes, *Literary detective work on the computer*. in Natural language processing (NLP), no. volume 12. Amsterdam ; Philadelphia: John Benjamins Publishing Company, 2014.
[33] Y. Lei, J. C. Bezdek, S. Romano, N. X. Vinh, J. Chan, and J. Bailey, “Ground truth bias in external cluster validity indices,” *Pattern Recognition*, vol. 65, pp. 58–70, May 2017, doi: 10.1016/j.patcog.2016.12.003.
[34] M. Rezaei and P. Franti, “Set Matching Measures for External Cluster Validity,” *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2173–2186, Aug. 2016, doi: 10.1109/TKDE.2016.2551240.
[35] L. P. Yulianti and K. Surendro, “Implementation of Quantum Annealing: A Systematic Review,” *IEEE Access*, vol. 10, pp. 73156–73177, 2022, doi: 10.1109/ACCESS.2022.3188117.
[36] K. Golalipour, E. Akbari, S. S. Hamidi, M. Lee, and R. Enayatifar, “From clustering to clustering ensemble selection: A review,” *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104388, Sep. 2021, doi: 10.1016/j.engappai.2021.104388.

Judhi Santoso received the B.Sc. degree in mathematics from Institut Teknologi Bandung (ITB), Indonesia, in 1987, the master’s degree in computer science from Universitas Indonesia (UI), in 1991, and the Ph.D. degree in informatics from the School of Electrical Engineering and Informatics, ITB, in 2006. He is currently an Associate Professor of Computer Science with Institut Teknologi Bandung (ITB), Indonesia. His primary research interests include modeling and simulation, soft computing, and numerical analysis.

Lenny Putri Yulianti (Member, IEEE) received B.Sc. in Information System and Technology from Institut Teknologi Bandung (ITB) in 2017 and M.Sc. in Informatics from ITB in 2018. Since 2019, She is a Lecturer at School of Electrical Engineering and Informatics ITB, Indonesia. She is currently a Ph.D. student in Informatics from ITB. Her current research interests are information system, machine learning, and quantum annealing.

Agung Trisetyarso is a Faculty Member at Department of Computer Science, Doctoral Programme, Bina Nusantara University (2015-present) and was a Faculty Member at Department of Informatics, Telkom University (2011-2015). He was awarded Bachelor of Science (Thesis: "Application of Darboux Transformation to solve Multisoliton Solution on Non-linear Schroedinger Equation"; Supervisor: Alexander Iskandar, Ph.D) and Master of Science from Department of Physics, Institut Teknologi Bandung. He got Ph.D from Keio University, in the field of Quantum Information and Computation under supervision of Prof. Kohei M. Itoh and Prof. Rodney V. Meter.

Kridanto Surendro received the PhD degree in Computer Science from Keio University, Japan, in 1999. He is currently working at Institut Teknologi Bandung, Indonesia. He is working in the area of information systems, information & knowledge analytic, and information governance. His research interests include the application of artificial intelligence to enterprise engineering and quantum computing.