An interval RSP-based ensemble model for big data analysis

Wenzhu Cai, GengYuan Ao, YiGang Lin, Mark Junjie Li ✉
College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
caiwenzhu2021@email.szu.edu.cn, jj.li@szu.edu.cn

***Abstract*—Ensemble learning for big data has been successful in machine learning and has great advantages over other learning methods. The ensemble model based on Random Sample Partition (RSP) is a prominent method of it. Although the RSP data blocks have the consistent probability distribution function as the whole data, there is some uncertainty in prediction results due to the non-overlapping data between blocks. In this paper, we propose a novel interval ensemble model based on RSP named Inr-RSP, which maps prediction results to interval-valued data by interval modeling and then uses the IAA aggregation method to convert the interval-valued data into fuzzy sets to get a more accurate and stable final result. The experimental classification results from four real datasets also show that the performance of this model is better than that of the traditional RSP ensemble model. And the IAA method usage has a stronger ability to capture uncertainty in prediction than the common majority voting method.**

***Index Terms*—Big data analysis, Ensemble learning, Random Sample Partition, Interval Agreement Approach**

I. Introduction

In recent years, due to the popularity of emerging technologies such as the Internet of Things (IoT), social media, and mobile devices, the scale of data has exploded. Faced with such massive amounts of data, how to efficiently process and analyze them has become an urgent problem to be solved. One of the biggest challenges in big data analysis is how to perform complex computing tasks within a given amount of computing resources. Previously, divide and conquer was the main strategy for big data analysis and calculation, which divide data into small subsets and then processed the subsets independently [1]. The MapReduce [2] and Spark [3], two distributed programming models, are also based on this strategy to process massive data. However, due to the iterative operation, the execution efficiency decreases, and the above models are limited by available memory resources in the calculation and analysis [4]. Therefore, the memory issue becomes a problem for big data analysis.

This problem is mainly alleviated by sampling techniques. Traditional sampling methods such as simple random sampling [5], stratified sampling [6], reservoir sampling [7], and the Record-Level Sampling (RLS) of the Hadoop Distributed File System (HDFS) [8] in distributed architectures, are all based on records. It becomes time-consuming and limited by memory in big data because selecting records with equal probability requires scanning the entire data. Ensemble learning is a common approach when using sampling techniques by dividing the data into many subsets or fitting multiple models using different algorithms, which typically improves the predictive performance of data mining and machine learning algorithms [9]. Nevertheless, the traditional ensemble models including Bagging [10] and Boosting [11] methods can not avoid the bottlenecking of memory resources when using the whole large dataset. Salman et al. propose an appropriate analysis model for large-scale datasets call Random Sampling Partition (RSP), which stores data as ready-to-use blocks of non-overlapping random samples [12]. The generation of RSP blocks is an offline operation, and each block has the consistent probability distribution with the whole data, thus providing the possibility of using a few blocks to approximate the whole big data without the limit of memory.

The existing RSP model generates the RSP blocks using the two-stage data processing (TSDP) [13] algorithm and then obtains the approximate result by processing each block respectively. Although RSP blocks have the consistent probability distribution with the big data, there is some uncertainty in the prediction results due to the non-overlapping of the data between the blocks. The common aggregation strategy is majority voting [14], but it does not consider the effect of the interval values. Besides, the number of learning models is determined by the number of learning models, which is not flexible.

In this paper, we propose an interval ensemble learning model based on RSP named Inr-RSP, which takes into account the uncertainty of the prediction results using interval modeling and uses the Interval Agreement Approach (IAA) to aggregate the final result. Experimental results show that the Inr-RSP model can achieve more accurate and robust classification with minimal information loss. Meanwhile, it presents that a few RSP blocks are enough to achieve the performance of the entire blocks and the number of learning models can be independent of the number of blocks which reduces model costs.

II. Related Works

Big data Sampling is a technology that extracts a sample set from a big dataset to facilitate data processing and analysis. The distribution of the sample data is important to the machine learning models. In the case of random sampling, the distribution of the predicted sampling is similar to that of the overall data. Common sampling methods include Bernoulli Sampling

DOI reference number: 10.18293/DMSVIVA2023-089

[15], Simple Random Sampling [5], Stratified Sampling [6], and so on [16]. Bernoulli Sampling [15] is to randomly select a single sample with moderate probability from the total with variable sample size and prone to sample bias. Simple Random Sampling [5] takes a lot of work when the data size is large or the distribution is more dispersed. Stratified Sampling [6] provides greater statistical precision and reduces sampling error. Similarly, the Bootstrap [17] method requires a large number of replicate samples and traversing the full data each time, which requires large enough memory resources. The RSP model divides large data into ready-to-use disjoint blocks whose distribution is consistent with that of the entire dataset. The use of RSP models can build ensemble models with fewer data, solving the problems of high computation and memory limitation [12].

The aggregation functions of ensemble learning are methods of combining multiple predictions into a final prediction result. Some of the most classic methods are majority voting [14], weighted voting [18], and stacking [19]. Majority voting [14] is the most common and effective method. Papers [20]–[22] applied the majority voting method to ensemble learning in different applications, and the results of the studies indicated that the majority voting method had a nice performance. In recent years, fuzzy theory has been used to deal with uncertain data, and interval-valued aggregation functions based on fuzzy theory have been proposed for ensemble learning [23]. Paper [24]–[27] proposed interval-valued aggregation functions to capture the uncertainty of data and applied them to ensemble learning. In particular, the Interval Agreement Approach (IAA) [24] converts interval-valued data into fuzzy sets. The IAA method addresses the limitations of the Interval Approach (IA) [28] and the Enhanced Interval Approach (EIA) [29] which only consider fuzzy sets of limited types and cannot handle uncertain intervals. It considers the minimal assumptions of interval data and does not rely on data preprocessing and outlier removal.

III. Preliminaries

In this section, we succinctly review the Random Sample Partition data model and briefly describe the Interval Agreement Approach.

A. Random Sample Partition (RSP)

Random Sample Partition is a distributed data model to facilitate block-level sampling and support big data analysis [12]. In this model, the statistical properties of the data set are preserved in a group of small disjoint data blocks as ready-to-use random samples (RSP blocks) from the entire data. Each RSP block has consistent probability distribution with the whole big data, allowing local results on different data blocks to approximate the global results on the whole big data. Also, it can address the limitation of memory and high computing cost in large-scale data.

With the RSP model, a partitioning of $\mathbb{D}$ into k non-overlapping random sample data blocks $T = \{D_1, D_2, ..., D_k\}$ in advance is represented as RSP blocks if:

- $\bigcup_{i=1}^{k} D_i = \mathbb{D}$
- $D_i \cap D_j = \emptyset,\ \ where\ i, j \in \{1, 2, ..., k\}\ \ and\ i \neq j$
- $E\left[F_i\left(x\right)\right] = \mathbb{F}\left(x\right),\ \ where\ i \in \{1, 2, ..., k\}$

where $F_i(x)$ is the sample probability distribution function of a random variable x in D_i. Accordingly, each block of T is called an RSP block of $\mathbb{D}$. Selecting an RSP block from T equals directly extracting random samples from $\mathbb{D}$. To analyze large-scale data, using such Block-Level Sampling is more efficient than Record-Level Sampling because it not requires scanning the entire data.

The RSP-based ensemble model for big data analysis uses a few selected RSP blocks to obtain approximate results. First, a block-level sample is selected from the RSP. Second, a sequential algorithm is applied parallel to each selected RSP block. Third, the outputs of these blocks are combined to produce an approximate result for the entire data (i.e., the majority voting in a classification task or the average response in a regression task). The ensemble process for the classification task is shown in Figure 1.

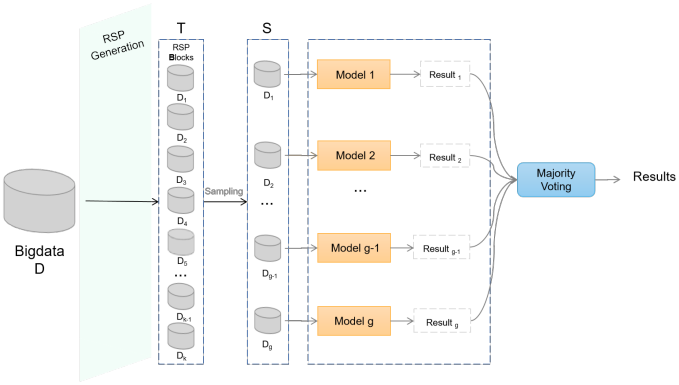

Fig. 1. The RSP-based ensemble model for big data analysis

B. Interval Agreement Approach (IAA)

The Interval Agreement Approach is a novel approach to generating fuzzy sets from interval-valued data and is accurately modeled by aggregating collective information captured by intervals [24]. An interval is denoted as $\bar{A} = [l_{\bar{A}}, r_{\bar{A}}]$, where $l_{\bar{A}}$ shows the left endpoint and $r_{\bar{A}}$ represents the right endpoint. Let $\mathcal{A} = \left\{\bar{A}_1, ..., \bar{A}_n\right\}$ be a set of intervals and a Type-1 Fuzzy Set (T1 FS) named A in IAA. The membership function μ_A of A is defined as:

$$\begin{aligned}\mu_A = {} & y_1 / \bigcup_{i_1=1}^{n} \bar{A}_{i_1} \\ & + y_2 / \left(\bigcup_{i_1=1}^{n-1} \bigcup_{i_2=i_1+1}^{n-1} \left(\bar{A}_{i_1} \cap \bar{A}_{i_2}\right) \right) \\ & + \dots \\ & + y_n / \left(\bigcup_{i_1=1}^{1} \dots \bigcup_{i_n=n}^{n} \left(\bar{A}_{i_1} \cap \dots \cap \bar{A}_{i_n}\right) \right)\end{aligned} \tag{1}$$

where $y_i = i/n$. Equation(1) represents the common notation of membership for fuzzy sets and / refers to the degree of membership rather than division. That means a value of μ_A shows the number of that value within all the intervals in $\mathcal{A}$. When the y_i is equal to 1, it indicates that all intervals are intersected.

There are two ways to simplify equation(1). One is that the membership of any value x can be calculated as the count of intervals which x contained like

$$\mu_A(x) = \frac{1}{n}\sum_{i=1}^{n}\mu_{\bar{A}_i}(x)$$
$$\textbf{where } \mu_{\bar{A}_i}(x) = \begin{cases} 1 & l_{\bar{A}_i} \le x \le r_{\bar{A}_i} \\ 0 & \text{else} \end{cases} \tag{2}$$

The other way to show the membership function is to subtract the number of left endpoints less than x in $\mathcal{A}$ from the number of right endpoints in $\mathcal{A}$ less than x as

$$\mu_A(x) = \frac{1}{n}\left(\sum_{i=1}^{n}\left(l_{\bar{A}_i} \le x\right) - \sum_{i=1}^{N}\left(r_{\bar{A}_i} \le x\right)\right) \tag{3}$$

Thus, A Type-1 Fuzzy Set can be generalized over $\mu_A(x)$.

IV. Proposed Model

In this section, we introduce the new interval RSP-based ensemble model named Inr-RSP which uses interval modeling and the IAA aggregation method to capture the uncertainty of prediction results and decrease the information loss. The main process of Inr-RSP is shown in Figure 2.

A. Generate RSP

Let $\mathbb{D}$ be a multivariate data set of N records and M features where N is large. A partitioning of $\mathbb{D}$ into k small disjoint data blocks $\{D_1, D_2, ..., D_k\}$ is regarded as a generation of RSP. The two-stage data processing (TSDP) algorithm for generation is as [13]:

- Sequentially cut $\mathbb{D}$ into p non-overlapping subsets called a partition of $\mathbb{D}$. Each subset has the same size with n records. Randomize each subset into i.i.d and cut it into an RSP of k parts independently to generate P data blocks.
- From each RSP block, select its corresponding RSP block, for 1 to k, to generate a new data block. Repeat this merging operation k times to generate a new partition $\{D_1, D_2, ..., D_k\}$, which is an RSP of $\mathbb{D}$.

The RSP model generates ready-to-use non-overlapping data blocks with consistent probability distribution of the entire data. It only needs to be executed once, which achieves Write-Once-Use-Many-Times(WOUM) strategy.

B. RSP Blocks Sampling

In this part, select g blocks from RSP data blocks $T = \{D_1, D_2, ..., D_k\}$ without replacement to form a sample set S as

$$S = \{D_1, D_2, ..., D_g\}$$

where $g \le k$. Thus, memory and communication costs depend on g, not k. The sampled RSP blocks are the same as the samples of the whole big data used for the following big data analysis.

C. Build Different Models

According to the analysis task, the base model can choose different learning models for the same task or one learning model with different parameters. For example, different base models, e.g. decision tree, support vector machine, and logistic regression, can be used if the task is classification.

D. Generate Uncertain Intervals and Aggregate

The key idea of the proposed model is capturing uncertainty by uncertain intervals from different data samples and models. Firstly, the selected RSP blocks are processed in different built models. In this part, intermediate results $\{Result_{i\_j}\}_{j=1}^{g}$ can be generated for each model i. Then, to avoid the influence of outliers on the results, we used Tukey's Test to process the intermediate result. Consider DL_i as $Q1_i - k * 1.5$ and UL_i as $Q3_i + k * 1.5$, the uncertain interval for each model i is shown as:

$$I_i = [DL_i, UL_i] \tag{4}$$

where the $Q1_i$ is the first quartile of $\{Result_{i\_j}\}_{j=1}^{g}$ and $Q3_i$ is the third quartile of $\{Result_{i\_j}\}_{j=1}^{g}$ for model i. Also, k is the difference between $Q3_i$ and $Q1_i$.

As mentioned before, IAA is used to generate a Type-1 Fuzzy Set (T1 FS), which is able to capture variation in the opinion of a particular decision model and divergence between the individual views of a group of decision models. Using uncertain intervals I_i and equation(2), a T1 FS is defined as:

$$A = \{((l_i, r_i), u_i)\}_{i=1}^{z} \tag{5}$$

where l_i is the left point, r_i is the right point and u is the membership function value of regions.

E. Defuzzification of Fuzzy Sets

There are many defuzzification methods to calculate the centroid of the Type-1 Fuzzy Sets. In this part, the computation approach of [30] is used to acquire the centroid as follows:

$$c = \frac{u_1 * (l_1 + r_1) + u_2 * (l_2 + r_2) + \cdots + u_z * (l_z + r_z)}{2(u_1 + u_2 + \cdots + u_z)} \tag{6}$$

where c is the centroid which will be applied as the final result. For binary classification, if the centroid is equal to or upper than 0.5, the final class will be class zero, and if it's not, it is class one. Similarly, the multiclass classification result is the primary class, and the other case is the secondary class.

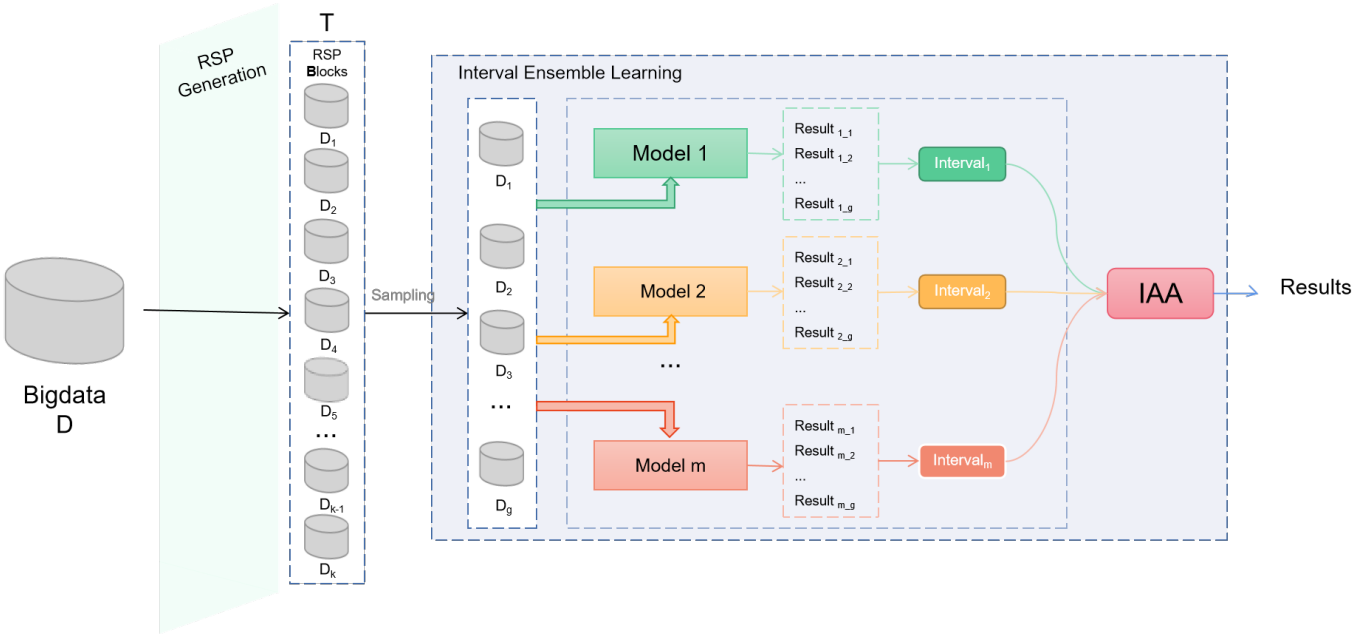

Fig. 2. The Inr-RSP ensemble model for big data analysis

V. Experiments and Results

To demonstrate the classification performance of the proposed model for big data analysis, we conducted several experiments on four datasets. First, we show the characteristics of datasets, experiment settings, and evaluation methods used in our experiments. Then, we evaluate the performance of the proposed model in classification, compared with the traditional RSP analysis model and interval RSP model which are applicable to majority voting. Also, we run our model on different sampling sizes, the various number of selected RSP blocks and learning models to obtain the sensitivity of the proposed model.

A. Datasets

We evaluate the proposed model on four datasets from the University of CaliforniaIrvine (UCI) [1] machine learning repository. As the optimization problem is an ensemble learning under big data, the number of records in the selected datasets is relatively large. In general, each of the four datasets differs in size, features and classes. The properties are described in Table I.

TABLE I
Properties of the datasets used in experiments

Dataset	Records(N)	Features(M)	Classes
Covertype	581,012	54	7
Watch_acc	3,777,046	5	18
SUSY	5,000,000	18	2
HIGGS	11,000,000	28	2

B. Experiment settings

The experiments focus on the classification task, so the decision tree is used as the base classifier. To generate different classifiers, with M as the number of features in each data, each decision tree is generated by abandoning a random feature that has not been ignored. Therefore, the maximum number of classifiers m cannot exceed the number of data features.

Notably, the classifiers' outputs in Inr-RSP are main class probabilities, not labels. For binary classification, the specified primary class is class Zero, and the secondary is class One. Also, the proposed model is suitable for multiclass classification. Consider the class of maximum probabilities as the primary class and the second maximum as the secondary. The multiclass classification ensemble problem is converted to determine the main classification class.

We use the abbreviations below for simplicity. RSP is to represent the traditional RSP analysis model which processes the RSP blocks independently and then aggregates them by majority voting method. Then, using interval RSP to present interval modeling of RSP blocks and aggregation by majority voting. Finally, Inr-RSP is proposed by this paper to represent the interval modeling of RSP blocks but aggregation using the IAA.

In Inr-RSP, the maximum number of classifiers is equal to the number of data features. In the preliminary experiment, to facilitate fair comparisons with other models, the size

[1] https://archive.ics.uci.edu/ml/index.php

of each RSP block n, the number of RSP blocks(g=5) and classifiers(m=5) are fixed for each dataset. In the parameter influence experiment, each dataset is divided into two RSP block sizes. The number of RSP blocks varies from 2 to 20 with intervals of 2, and the number of classifiers differs according to the data features. Each experiment only changes one parameter to reflect the influence of the parameter. To eliminate chance, the experiments are repeated 10 times and the average results are reported.

C. Evaluation Methods

To get convincing results, we use the same testing data to test models for proposed and compared models. Also, using the following two matrices, Accuracy and Kappa, to measure the performance of classification tasks.

$$\text{Accuracy } = \frac{TP + TN}{TP + FN + TN + FP} \tag{7}$$

where TP is the number of true positive predictions, TN is the number of true negative predictions, FP is the number of false positive predictions, and FN is the number of false negative predictions. It shows the proportion of accurate results among the total number of testing.

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{8}$$

where P_o is the overall accuracy and P_e is the chance consistency error. It represents the percentage of errors reduced if the classification were completely random.

D. Preliminary Results

Table II shows the four classification matrices results of the proposed Inr-RSP model compared with the interval RSP model and the RSP model both aggregated by majority voting method in four datasets. For fair comparisons, we used the same sampling size, the number of selected RSP blocks, and the number of learning models for each model. That is, the difference in the results is the performance improvement.

As seen in Table II, both interval ensemble models are superior to the RSP model in most datasets, which means that interval modeling can combine multiple predictions into more accurate interval predictions to preserve the uncertainty. Also, the Inr-RSP ensemble model using the IAA aggregation method successfully outperforms the contrast model in all datasets because the IAA algorithm can well consider the impact of interval values and process uncertain data, which can also transform interval values into more stable and reliable results, so as to make more accurate and robust classification and minimize information loss.

E. Influence of Parameters

In this section, we experiment with the sensitivity of Inr-RSP to changes in parameters, including the RSP sampling size, the number of selected RSP blocks, and the number of classifiers. Note that in evaluating the selected parameters, all other parameters remain fixed during the experimental run.

1) RSP sampling size n:

Figures 3 and 4 present the classification accuracy of the Inr-RSP model on four datasets for two different RSP sampling sizes n (shown as solid lines). As n affects the amount of data, too small n will not allow the classifier to capture enough specific patterns, and so large n may increase the risk of overfitting. It is observed that the RSP sampling size affects the accuracy of the Inr-RSP model, a larger value of n generally leads to better classification.

2) Number of RSP blocks g:

Figure 3 also shows the classification accuracy of the Inr-RSP model on four datasets for different RSP block numbers. Since the traditional RSP model has the same number of classifiers as the blocks, it is not compared without fixed m. As shown in Figure 3, the classification accuracy of the proposed Inr-RSP model increases with the number of RSP blocks at a fixed m = 5 and maintains convergence at a certain number of blocks, indicating that a stable model can be built with a few blocks. In contrast, most of the Inr-RSP models aggregated by the majority voting method are unstable and have poor accuracy as it does not consider the uncertainty of interval data.

3) Number of classifiers m:

Figure 4 represents the classification accuracy of the Inr-RSP model on four datasets for distinct classifier numbers. Because the number of classifiers in the traditional RSP model is the same as the number of blocks, only display the results of g=5 and m=5. The results show that the accuracy of the Inr-RSP model does not have a significant effect so fewer classifiers can build a stable model but no limit to the number of blocks. Meanwhile, it outperforms the comparison model when the g increases.

VI. Conclusion

This paper presents a novel ensemble model for big data named Inr-RSP, which better captures the uncertainty of the traditional RSP-based ensemble model through interval modeling and interval aggregation methods. The Inr-RSP model uses the IAA aggregation method to transform the interval-valued data generated by RSP data blocks interval modeling into fuzzy sets and then obtains the final result through centroid calculation. This model can reduce information loss and obtain more accurate and robust ensemble results. The new model outperforms the traditional RSP model on four real datasets and is also superior to the majority voting method using the IAA.

References

[1] Bo-Wei Chen, Wen Ji, and Seungmin Rho. Divide-and-conquer signal processing, feature extraction, and machine learning for big data. *Neurocomputing*, 174:383, 2016.

[2] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

TABLE II
CLASSIFICATION ACCURACY AND KAPPA RESULTS ON FOUR DATASETS

Dataset	n	Accuracy			Kappa		
		RSP	Interval RSP	Inr-RSP	RSP	Interval RSP	Inr-RSP
Covertype	44150	0.8377	0.8375	**0.8619**	0.7372	0.7369	**0.7761**
Watch_acc	35800	0.6605	0.6970	**0.7323**	0.6411	0.6801	**0.7166**
SUSY	50000	0.7444	0.7736	**0.7930**	0.4835	0.5397	**0.5787**
HIGGS	10890	0.6757	0.6836	**0.7109**	0.3499	0.3657	**0.4193**

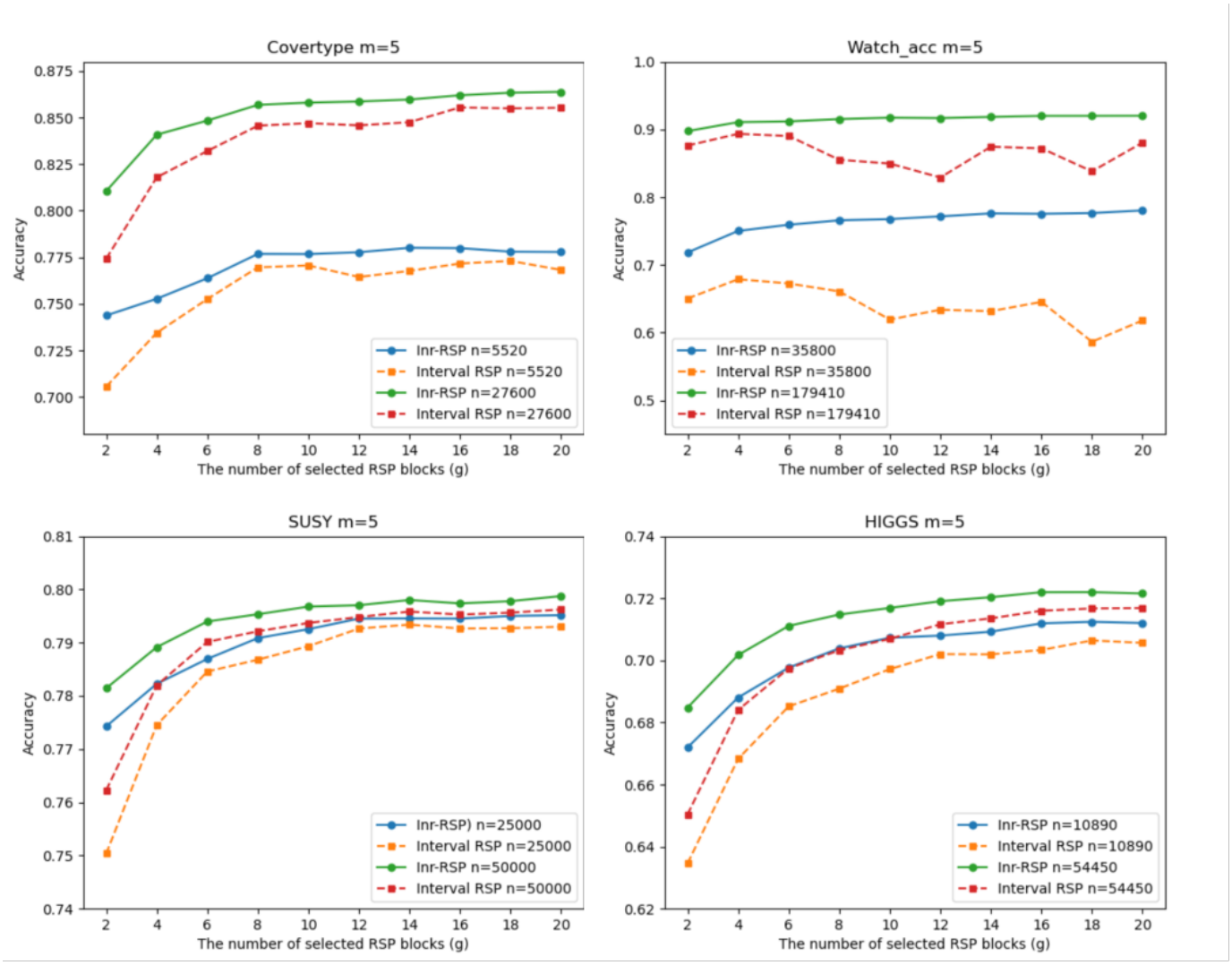

Fig. 3. Compare classification accuracy for the increasing number of RSP blocks with two RSP sampling sizes

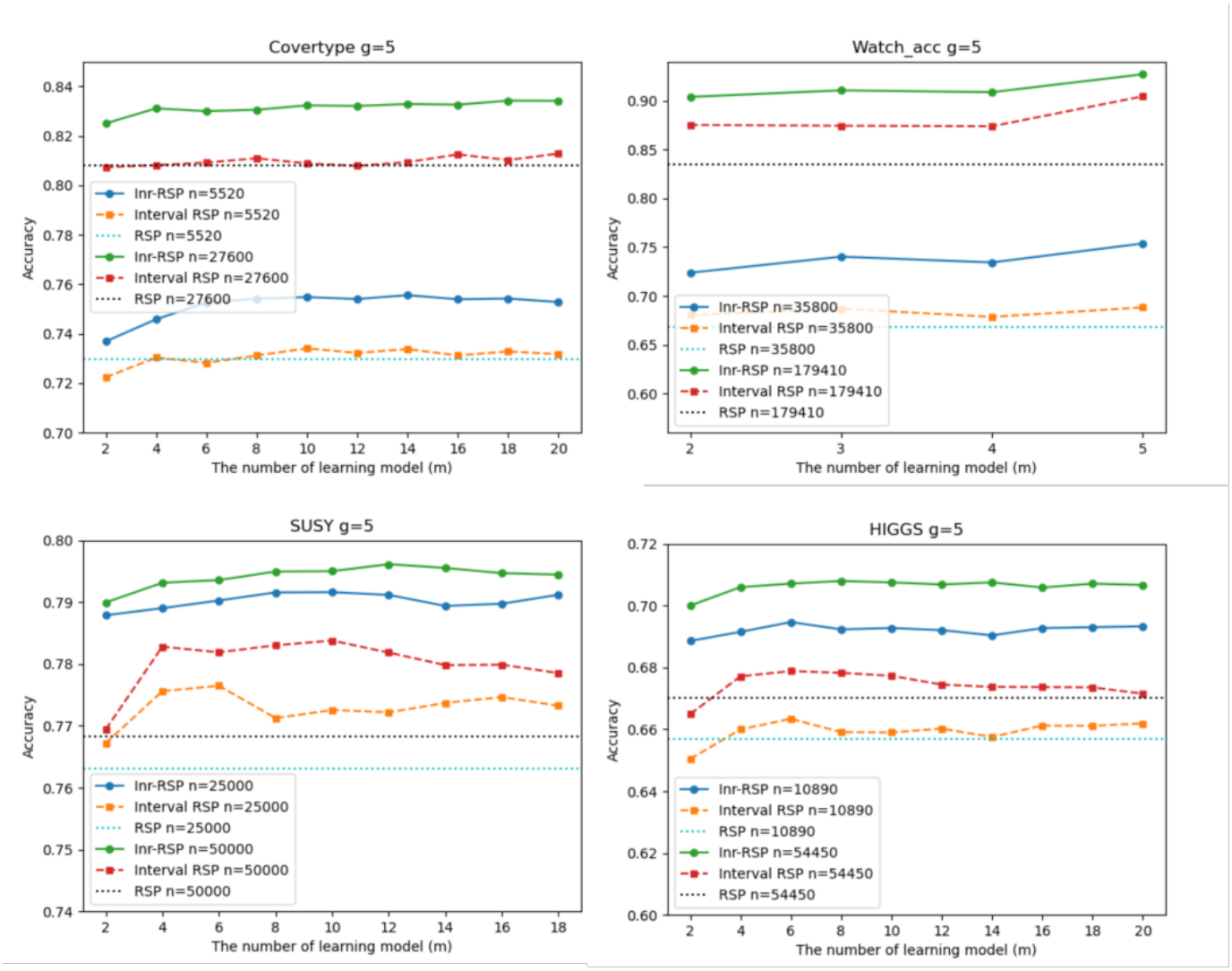

Fig. 4. Compare classification accuracy for the increasing number of classifiers with two RSP sampling sizes

[3] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. In Erich M. Nahum and Dongyan Xu, editors, *2nd USENIX Workshop on Hot Topics in Cloud Computing, HotCloud'10, Boston, MA, USA, June 22, 2010*. USENIX Association, 2010.

[4] Lei Gu and Huan Li. Memory or time: Performance evaluation for iterative operation on hadoop and spark. In *10th IEEE International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing, HPCC/EUC 2013, Zhangjiajie, China, November 13-15, 2013*, pages 721–727. IEEE, 2013.

[5] Cem Kadilar and Hulya Cingi. Ratio estimators in simple random sampling. *Appl. Math. Comput.*, 151(3):893–902, 2004.

[6] Peter J Bickel and David A Freedman. Asymptotic normality and the bootstrap in stratified sampling. *The annals of statistics*, pages 470–482, 1984.

[7] Jeffrey Scott Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985.

[8] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*, pages 1–10. Ieee, 2010.

[9] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers Comput. Sci.*, 14(2):241–258, 2020.

[10] Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.

[11] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In Lorenza Saitta, editor, *Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, July 3-6, 1996*, pages 148–156. Morgan Kaufmann, 1996.

[12] Salman Salloum, Joshua Zhexue Huang, and Yulin He. Random sample partition: A distributed data model for big data analysis. *IEEE Trans. Ind. Informatics*, 15(11):5846–5854, 2019.

[13] Chenghao Wei, Salman Salloum, Tamer Z. Emara, Xiaoliang Zhang, Joshua Zhexue Huang, and Yu-Lin He. A two-stage data processing algorithm to generate random sample partitions for big data analysis. In Min Luo and Liang-Jie Zhang, editors, *Cloud Computing - CLOUD 2018 - 11th International Conference, Held as Part of the Services Conference Federation, SCF 2018, Seattle, WA, USA, June 25-30, 2018, Proceedings*, volume 10967 of *Lecture Notes in Computer Science*, pages 347–364. Springer, 2018.

[14] Aytug Onan, Serdar Korukoglu, and Hasan Bulut. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Syst. Appl.*, 62:1–16, 2016.

[15] CT Fan, Mervin E Muller, and Ivan Rezucha. Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*,

57(298):387–402, 1962.

[16] Zhicheng Liu and Aoqian Zhang. A survey on sampling and profiling over big data (technical report). *CoRR*, abs/2005.05079, 2020.

[17] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael Jordan. The big data bootstrap. *arXiv preprint arXiv:1206.6415*, 2012.

[18] Ludmila I Kuncheva and Juan J Rodríguez. A weighted voting framework for classifiers ensembles. *Knowledge and information systems*, 38:259–275, 2014.

[19] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[20] Adib Ashfaq A Zamil, Sajib Hasan, Showmik MD Jannatul Baki, Jawad MD Adam, and Isra Zaman. Emotion detection from speech signals using voting mechanism on classified frames. In *2019 international conference on robotics, electrical and signal processing techniques (ICREST)*, pages 281–285. IEEE, 2019.

[21] Rahma Atallah and Amjed Al-Mousa. Heart disease detection using machine learning majority voting ensemble method. In *2019 2nd international conference on new trends in computing sciences (ictcs)*, pages 1–6. IEEE, 2019.

[22] Zhiyong Lv, Tongfei Liu, Cheng Shi, Jon Atli Benediktsson, and Hejuan Du. Novel land cover change detection method based on k-means clustering and adaptive majority voting using bitemporal remote sensing images. *Ieee Access*, 7:34425–34437, 2019.

[23] Mikel Uriz, Daniel Paternain, Iris Dominguez-Catena, Humberto Bustince, and Mikel Galar. Unsupervised fuzzy measure learning for classifier ensembles from coalitions performance. *IEEE Access*, 8:52288–52305, 2020.

[24] Christian Wagner, Simon Miller, Jonathan M. Garibaldi, Derek T. Anderson, and Timothy C. Havens. From interval-valued data to general type-2 fuzzy sets. *IEEE Trans. Fuzzy Syst.*, 23(2):248–269, 2015.

[25] Urszula Bentkowska and Barbara Pekala. Diverse classes of interval-valued aggregation functions in medical diagnosis support. In Jesús Medina, Manuel Ojeda-Aciego, José Luis Verdegay Galdeano, Irina Perfilieva, Bernadette Bouchon-Meunier, and Ronald R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications - 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018, Proceedings, Part III*, volume 855 of *Communications in Computer and Information Science*, pages 391–403. Springer, 2018.

[26] Krzysztof Dyczkowski. *Intelligent Medical Decision Support System Based on Imperfect Information - The Case of Ovarian Tumor Diagnosis*, volume 735 of *Studies in Computational Intelligence*. Springer, 2018.

[27] Urszula Bentkowska, Jan G. Bazan, Wojciech Rzasa, and Lech Zareba. Application of interval-valued aggregation to optimization problem of k-nn classifiers for missing values case. *Inf. Sci.*, 486:434–449, 2019.

[28] Feilong Liu and Jerry M. Mendel. Encoding words into interval type-2 fuzzy sets using an interval approach. *IEEE Trans. Fuzzy Syst.*, 16(6):1503–1521, 2008.

[29] Simon Coupland, Jerry M. Mendel, and Dongrui Wu. Enhanced interval approach for encoding words into interval type-2 fuzzy sets and convergence of the word fous. In *FUZZ-IEEE 2010, IEEE International Conference on Fuzzy Systems, Barcelona, Spain, 18-23 July, 2010, Proceedings*, pages 1–8. IEEE, 2010.

[30] Christian Wagner and Hani Hagras. zslices - towards bridging the gap between interval and general type-2 fuzzy logic. In *FUZZ-IEEE 2008, IEEE International Conference on Fuzzy Systems, Hong Kong, China, 1-6 June, 2008, Proceedings*, pages 489–497. IEEE, 2008.