Active Learning with an Adaptive Classifier for Inaccessible Big Data Analysis

Sadia Jahan*, Md Rafiqul Islam†, Khan Md. Hasib‡, Usman Naseem§, Md. Saiful Islam¶
*Department of CSE, City University, Bangladesh
†Advanced Analytics Institute (AAi), University of Technology Sydney, Australia
‡Department of CSE, Ahsanullah University of Science and Technology, Bangladesh
§School of Computer Science, University of Sydney, Australia
¶School of Information and Communication Technology, Griffith University, Australia
{sadiajhn, rafiqulislam.cse24, khanmdhasib.aust, engr.usmannaseem87}@gmail.com, saiful.islam@griffith.edu.au

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9534046

***Abstract*—Supervised machine learning (ML) approaches effectively derive valuable insights from big data. These approaches, on the other hand, require an extensive amount of high quality annotated data for training, created manually by domain experts through a costly and time-consuming process. To overcome this challenge, active learning (AL) is a promising approach, which can support a fast, cost-efficient and common strategy to deal with big data with limited labeling effort. Instead of annotating a large pool of unlabeled data, as in standard supervised learning, AL reduces the volume of data that requires manual annotation by effectively selecting subsets of highly informative samples for manual annotation within an iterative process. In this paper, we aim to present a robust approach utilizing AL to mitigate the aforementioned challenges and help the decision-makers. To be precise, we propose a framework involving a support vector machine (SVM) technique in AL for mining big data to manage inaccessible data situations. The proposed approach is tested on five different semi-supervised data sets. The performance of the proposed framework is evaluated using traditional ML classifiers such as Naïve Bayes (NB), Decision Tree (DT), Sequential Minimal Optimization (SMO), Random Forest (RF), Bagging and Adaboost. Among the reported classifiers, bagging achieves the best outcome, delivering 99.19% accuracy. According to the results of the experiment conducted we find that the proposed method increases the efficiency of the classifiers in AL with fewer training instances.**

***Index Terms*—Big data, Active learning and Machine learning.**

I. Introduction

There has been an enormous increase in the volume of multimedia based data accessible in recent years. This is attributed mainly to the technical advancement and evolution of the internet which has contributed to the emergence of a significant new media range. Managing and organizing such a vast volume of data is an expensive job that requires to be streamlined as much as possible. Machine learning (ML) algorithms have been used for automated recognition of digital objects to facilitate this role. However, achieving reasonable classification performance by ML algorithms involves a vast volume of labeled data for algorithm testing. Categorizing information needs to be performed periodically, but it's time to do so — a consuming and costly job of its own. Consequently, methods that aim to exploit unlabeled data to improve the classification performance have been of primary interest in recent years.

Active learning (AL), which uses a limited collection of labeled information to save human effort for successful modeling in the face of large data, is one of the promising approaches [1]. AL is a human-in-the-loop method with the capacity to substantially reduce human inclusion compared with the conventional supervised ML methods that require a massive amount of labeled data at the start [2], [3]. Queries that adopt a data selection strategy appear frequently in AL, requiring a response from an oracle or a human annotator. The standard question in each iteration of AL is to label a few unlabeled data that may be difficult to determine their label details through premature models [3]. Unlabeled data collection techniques involve using parameters such as complexity, diversity, or representativeness, to name a few, to determine which data might be crucial in cases where we need to know their label details for successful modeling.

AL in supervised ML is a method of achieving high classification outcomes using fewer training instances to learn a definition that can often be much smaller than that needed in traditional supervised learning. Interactively, it asks a professional to unmarked label instances. Huge amounts of data are produced in the era of big data from a variety of sources, including the cloud, business management, and various machines and devices. However, vital data is often inaccessible to users due to its incomplete and unstructured type [4], [5]. Furthermore, managing and organizing a vast volume of data is a time-consuming process that should be automated as much as possible [6]. Thus, there is a space to seek practical solutions that will enable them to manage and arrange out a large amount of data, improve business performance, and create new and useful data-driven business models.

Machine learning (ML), such as supervised algorithms, benefits from the fact that it simplifies the role of automatic data classification, which requires a large amount of labeled data for training [7], [8]. A recent study has demonstrated that performance will bias efficiency through the disproportionate distribution of class examples in the learning method. This means that the class provides limited specificity's for the

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

minority class, while the class offers great precision [9], [10]. However, data labeling is time-consuming and expensive task. Since the last few decades, semi-supervised learning approaches have been used to increase the accuracy of unlabeled data classification, and it is also challenging to obtain label data from unlabeled data sources [11]. It is noted that without labeling data, we can predict or generate the same accuracy. However, by labeling a small percentage from the massive amount of data, it will be beneficial and productive [12]. Therefore, previous works have some limitations as for finding approach of informative unlabeled instances and label from big data. Again, there will need improvement in the previous work classification accuracy, and a small percentage of label data is a timely concern.

Existing studies used methods like random sampling, representative sampling, uncertainty sampling, local uncertainty sampling, global uncertainty sampling, and others to try to minimize unlabeled data. [13]. Many properly labeled training instances are needed in supervised learning or classification to achieve accurate predictive models [14]. Incorrectly labeled or inconsistent examples reduce the output of research models [15]. A detailed and comprehensive review is essential for the labelling of unlabeled training results [16]. In this regards, AL is the process of labelling unlabeled data, and to classifying semi-supervised data in ML [17], [18]. In general, AL in data mining strengthening is a successful solution to this problem [19], [20]. In the active training phase, an expert or oracle is used to tag unlabeled data in order to improve classification accuracy by asking the fewest possible questions of the user or expert. The number of training courses needed to learn an AL definition is also substantially smaller than in the conventional supervised learning phase [21], [22]. However, unlisted Big Data is manual tagging is very complicated and costly. In particular, big data safety is confronted with the security related challenges like Big data privacy and safety mechanism. Data quality, inconsistency and incompleteness with scalability, timeliness and data security are also the challenges of big data. For that reason many methods also applied for security purposes [23]. By using AL, it is also possible to acquire the most valuable unlabeled data from imbalance drifting data stream [24], [25]

In this study, the main contribution is to improve the classifier's performance with the assistance of professional expertise [26]. AL interactively queries to oracle or human experts for labeling the unlabeled instances. To find the best method for selecting the unlabeled samples, which are calling informative samples is the research challenges of AL [6]. In our proposed method, we apply Support Vector Machine (SVM) for collecting those informative instances in the AL process as we know that SVM separates the data into classes by creating a line or hyper-plane. SVM classifiers are exceptionally well adapted for AL related to their versatile mathematical properties. They conduct linear classification, usually in a kernel induced feature space, which simplifies the distance of the data point from the decision boundary [27]. By applying SVM, it is easy to collect those instances.

Therefore, the key **contributions** of this research are as follows:

- We apply SVM to collect informative instances from big unstructured data.
- We use different supervised and semi-supervised benchmark data sets from UCI ML repository and KEEL (Knowledge Extraction based on Evolutionary Learning) dataset repository with multiple classes.
- The performance is evaluated with different ML techniques. Our experimental outcomes served as an alternative source of knowledge to fill the traditional significant data reports and surveys' gaps.

The remaining part of the paper has been arranged accordingly. The successful learning and related work is discussed in Section II. The approach suggested in Section III is added. Section IV offers experimental findings with datasets with measurements. Section V eventually ends the conclusions and recommended recommendations for future work.

II. Related Work

Big data analysis using active learning is not a new idea. However, there has been no systematic investigation into how they should be analyzed for inaccessible big data. Again, in machine learning, active learning is a technique for achieving high classification results by using fewer training instances to learn a concept that is often much smaller than that used in conventional supervised learning. In this section, first we describe:

- Active learning techniques, and
- Big data analysis using active learning.

Thus, the research of each approach will be introduced below respectively.

A. Active Learning Techniques

The current approaches for big data analysis mainly include two directions: 1) manually feature extraction and building traditional ML techniques for classification, and 2) deep learning techniques to automatically extract features and construct. In this study, we describe several existing approaches that have been proposed to collect new issues, such as random sampling, representative sampling, sampling of uncertainties, local selection of uncertainties, global sampling of uncertainties, feature extraction, committee-based active learning, etc. for inaccessible big data analysis [28], [29]. Among the earliest and most common, active learning based is the sampling best solution because it offers a variety of discovery strategies and optimizations [26]. For example, from the existing study in 2018, Jahan [30] suggested to identify and mark the insightful imbalance instances by experts/ users to enhance the successful learning process considering the closest neighbour data centre instances to select informative cases, including the cluster centre. AL is also used successfully in interactive ways to solve multitask classification problem by taking decision from density initialization,evidential data and from weighted instances [31]–[35]. However, it is challenging to pick a

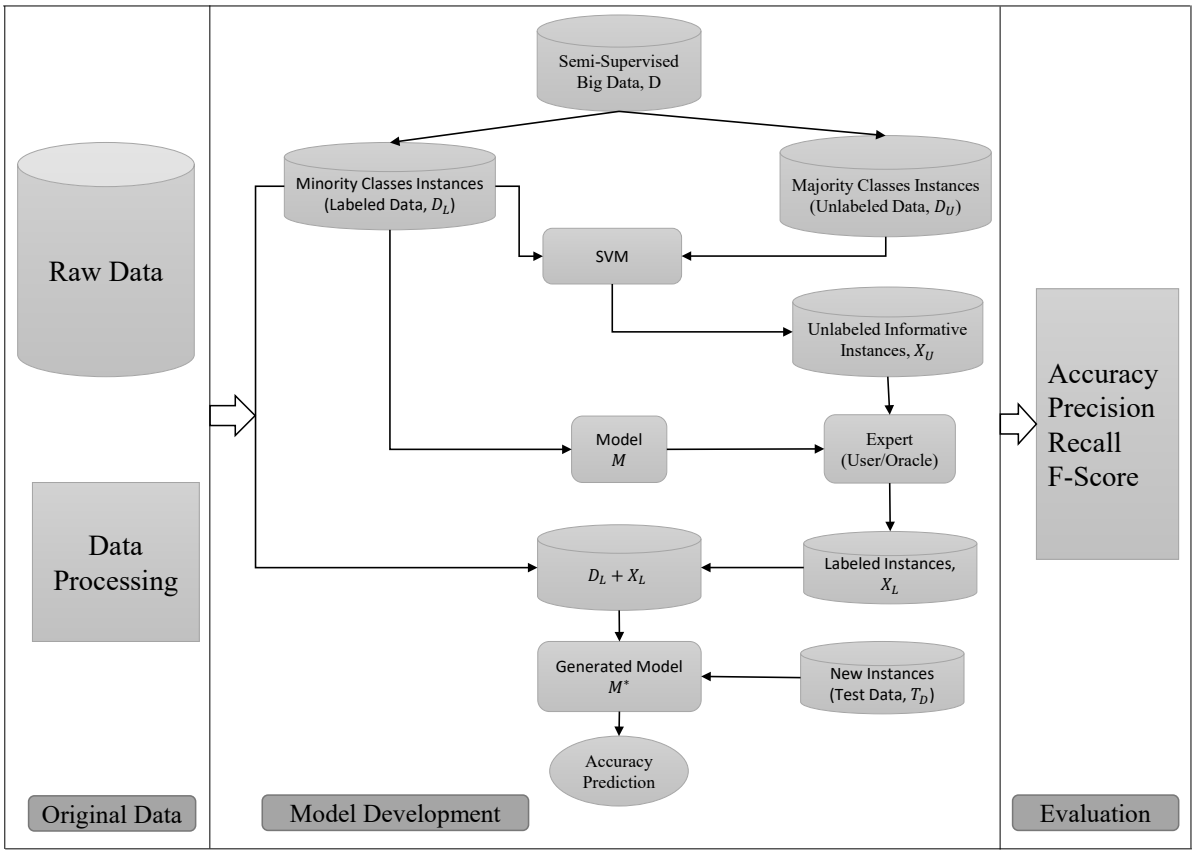

Fig. 1. The proposed framework for applying SVM in active learning for inaccessible big data analysis

minimal number of insightful unmarked instances from big unmarked data. Additionally, requiring experts to classify unmarked data in each iteration makes the learning process sluggish and expensive. Therefore, an adaptive classifier with active learning is referred to as pick instances from unmarked results [36].

B. Big Data Analysis using Active Learning

Active learning is mainly for considering to label unmarked instances. Among them, the modeling of the unlabeled information is the main research direction. To solve this problem, Hong [37] suggested an algorithm that was used euclidean weighted distance, SVM and radial integral kernel function wide packet classification dimensionality reduction algorithm. Hu et al. developed unsupervised active learning in 2009 [38] which was ambivalent for hybrid active learning because pre-labeled examples focused on random selection from the initial set are required. Zliobaite et al. [39] proposed a new strategy in 2014 that worked well for limited labeling budgets. Edwin Lughofer [40] suggested a new data-driven classification method in 2012 that did not include any initial labeling or learning. Yang et al [41] raised a multi-class active learning algorithm in semi-supervised batch mode to test data vulnerability on visual idea recognition from the active pool in 2015 [41]. In 2015, Hajmohammadi et al. [42] introduced the paradigm, incorporating a semi-conservation with an adaptive learning approach focused on ambiguity.

However, the existing study on inaccessible big data management using active learning is very limited. Although a large studies have analysed for various issues using various techniques, there is no study on active learning with an adaptive classifier. Thus, to our best knowledge, this is the first work using active learning with an adaptive classifier to address the various data risk management issues.

III. METHODOLOGY

To analyse inaccessible big data, in this section, we propose a novel framework, as shown in Figure 1. This framework mainly consist of three parts such as 1) Data collection and processing, 2) Model development and 3) performance evaluation. Additionally, we describe the details of active learning for mining big data. Therefore, we first introduce the relevant symbol and terms used throughout the paper, as shown in the Table I. Afterwards, we describe active learning to build a classification model.

In machine learning, active learning is the method of achieving high classification accuracy with a smaller number of training instances than is expected in traditional supervised learning [26]. It interactively queries an expert to label the unlabeled instances. The aim is to improve the performance of a classifier by training it with the aid of expert knowledge [6].

TABLE I
COMMONLY USED SYMBOLS AND TERMS

Symbol	Term
D	The whole dataset, which includes all instances
D_L	Instances in which the label is recognized
D_U	Instances where the label is unfamiliar
X_U	A set of unlabeled informative instances
X_L	A set of label instances
M	A learning model/classifier
M^*	A classifier/ learning re-generated model
T_D	A set of testing instances

Input: Semi-supervised Big Data, D;
Output: An active learning model, M^* ;
Method:

1: divide D into D_L, and D_U;
2: create a model, M using D_L;
3: create unlabeled informative instances, X_U by applying SVM on both D_L and D_U, $X_U \leftarrow$ SVM on D_L and D_U;
4: Labeled instances, $X_L \leftarrow X_U$ by expert/ oracle;
5: $D_L \leftarrow D_L + X_L$;
6: generated model, M^* using D_L;
7: return M^*;

Algorithm 1: SVM in Active Learning

Assume we have a data set, D, that contains respectively labeled, D_L and unlabeled , D_U instances. Initially, D is bifurcated into D_L and D_U. An ensemble learning model, M^* is trained using labeled data, D_L. In contrast, a subset of unlabeled instances, $X_U \in D_U$ is chosen from the unlabeled data D_U and requests the expert/ user to label $X_U \rightarrow X_L$. Finally, X_L is added to D_L and the ensemble model, M^* is re-trained. This procedure is repeated until the consumer is fully satisfied. The most difficult part of active learning is selecting a subset of unlabeled instances from the original unlabeled data [4]. The active learning mechanism is shown in Figure 1.

It's difficult to select a small number of insightful unlabeled instances from large amounts of unlabeled data [43]. Pool-based active learning is most commonly used to pick instances from unlabeled data. However, querying the experts for labeling unlabeled data in each iteration makes the learning process slow and costly. We also need to improve classification accuracy by using the smallest number of possible training instances. Several methods for selecting unlabeled instances have been proposed in recent decades, including random sampling, representative sampling, uncertainty sampling, local uncertainty sampling, global uncertainty sampling, active mining, and committee-based active learning, among others [40].

IV. EMPIRICAL EVALUATION

This section presents the information about dataset collection and process, experimental setup, and analysis of the result.

A. Dataset Descriptions

In this experiment, we have used 5 different semi-supervised data set from the UCI ML and KEEL dataset repository such as 1) Abalone Dataset (Abalone), 2) Nursery Data Set (Nursery), 3) MAGIC Gamma Telescope Dataset (Magic), 4) Thyroid Disease Dataset (Thyroid), and 5) Two norms Data Set (Two norms). Each dataset consists of training instances where both labeled and unlabeled instances exist and testing instances separately. labeled instances from the training dataset are used to train a model, and by using that model, we labeled those unlabeled informative instances. Testing instances are used to predict the accuracy of the model. Table II presents the details of the data sets.

B. Experimental Setup

We have used jupyter notebook web application and python code for our experiment. We have used different types of scikit-learn library for data loading, processing, for the classifier, and finding accuracy. The attribute types of most of our dataset were just categorical or categorical and real. We have encoded those categorical data into an integer by using a prepossessing library for label encoding. To evaluated the proposed method, we have used accuracy, precision, recall, and F-score. We have used six different ML algorithms like Naïve Bayes, Decision Tree, SMO (Sequential minimal optimization) Random Forest, Bagging, and Adaboost for classification. The accuracy, precision, recall, and F-score are shown in Eq. 1 to Eq. 4 where TP, TN, FP, and FN are true positives, true negatives, false positive, and false negatives, respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - score = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

The above metrics are described briefly below.

- *True Positive (TP)* is the number of instances classified correctly as a positive value/yes (1) by the generated model M* after updating the labeled instances.
- *True Negative (TN)* is the number of instances classified correctly as a negative value/no (0) by the generated model.
- *False Positive (FP)* is the number of instances classified incorrectly, that is, machine predicts the value as positive/yes (1) but its actual value is negative/no (0) by the generated model M* after updating the labeled instances.
- *False Negative (FN)* is the number of instances classified incorrectly, that is, machine predicts the value as negative/no (0), but its actual value is positive/yes (1) by the generated model.

TABLE II
DATASETS DESCRIPTION

No.	Name of Datasets	No. of Features	Attribute Type	Total Instances	Training Instances	Testing Instances	Class Attributes
1	Abalone	9	Categorical, Real	7508	3756	3752	28
2	Nursery	9	Categorical	12960	11664	1296	5
3	Magic	9	Real	53721	36603	17118	2
4	Thyroid	22	Categorical, Real	7200	6480	720	3
5	Two Norm	9	Categorical, Real	7400	6660	740	2

TABLE III
EXPERIMENTAL RESULTS

Datasets	Training Instances	Informative Instances	Reduced rate (%)	Algorithm	Result with all instances				Result with informative instances			
					Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
Abalone	1506 (Label) 2250 (Unlabel)	816	63.73	Naïve Bayes	24.79	21.01	24.69	21.65	24.37	20.94	24.33	21.62
				Decision Tree	54.02	55.13	53.79	53.61	53.05	53.51	53.11	52.89
				SMO	24.82	20.42	25.22	17.87	24.62	17.93	24.96	17.17
				Random Forest	52.42	54.43	53.01	52.80	53.08	55.45	53.42	53.20
				Bagging	53.19	54.22	52.77	52.56	**53.21**	**54.86**	**53.28**	**53.03**
				Adaboost	21.18	07.53	20.93	8.87	21.32	07.53	21.20	09.21
Nursery	2333 (Label) 9331 (Unlabel)	4933	47.13	Naïve Bayes	67.02	82.80	69.86	73.33	67.47	83.24	69.32	73.04
				Decision Tree	93.54	91.31	91.53	91.12	93.71	91.82	91.91	91.69
				SMO	74.18	72.89	74.21	73.34	74.29	73.07	74.40	73.52
				Random Forest	91.80	91.00	91.33	90.72	91.23	91.19	91.37	90.64
				Bagging	94.06	91.69	91.90	91.38	**94.33**	**92.47**	**92.48**	**92.19**
				Adaboost	73.06	77.30	75.69	73.74	76.64	78.31	76.23	74.40
Magic	17705 (Label) 18898 (Unlabel)	392	**97.92**	Naïve Bayes	72.61	72.28	72.60	69.90	72.54	72.16	72.54	69.86
				Decision Tree	98.25	98.27	98.24	98.23	**98.97**	**99.01**	**99.01**	**99.01**
				SMO	73.45	72.67	66.25	63.08	73.67	72.34	66.48	64.40
				Random Forest	98.53	98.51	98.50	98.49	98.12	98.17	98.15	98.14
				Bagging	98.52	98.51	98.50	98.49	98.37	98.06	98.04	98.03
				Adaboost	84.79	84.42	84.56	84.27	84.76	84.57	84.70	84.41
Thyroid	1296 (Label) 5184 (Unlabel)	3571	31.14	Naïve Bayes	12.47	92.81	10.33	12.46	09.28	91.18	10.69	13.66
				Decision Tree	98.72	98.54	98.53	98.51	99.08	98.94	98.92	98.89
				SMO	93.17	88.16	93.14	90.38	93.14	88.06	93.06	90.23
				Random Forest	98.86	98.82	98.75	98.74	98.63	98.96	98.92	98.90
				Bagging	98.83	98.83	98.81	98.80	**99.19**	**99.03**	**99.03**	**99.01**
				Adaboost	96.94	97.75	97.56	97.60	97.28	97.76	97.61	97.63
Two Norm	1332 (Label) 5328 (Unlabel)	2023	62.03	Naïve Bayes	97.08	97.13	97.11	97.11	**97.22**	**97.24**	**97.22**	**97.22**
				Decision Tree	84.11	84.31	84.11	84.08	83.92	83.88	83.81	83.80
				SMO	96.59	96.93	96.86	96.86	97.05	97.04	97.00	97.00
				Random Forest	92.78	93.30	93.00	92.99	93.08	92.32	92.03	92.14
				Bagging	91.78	91.92	91.51	91.49	91.92	92.36	92.05	92.04
				Adaboost	94.89	94.95	94.84	94.83	95.08	94.94	94.86	94.86

- **Accuracy**: Accuracy is the percentage of the test set, which is classified correctly.
- **Precision**: Precision is the measurement of the exactness of the actual class.
- **Recall**: Recall is the measurement of completeness of the actual class.
- **F-Score**: F1-score or F-measurement is the combination of both precision and recall that it is taken both false positives and false negatives into accounts.

C. Results and Discussion

In Table III, we present the overall classification accuracy result (in percentage) with precision, recall, and f-score. This table also compares the result with all instances and the result with informative instances of our proposed framework. As our proposed framework's key goal is to reduce the unlabeled instances as much as possible. Instead of the product of all unlabeled instances, we can achieve nearly the same or better accuracy by labeling only the insightful unlabeled instances. Our experiment used Naïve Bayes, Decision tree, SMO, Random Forest, Bagging, and Boosting ML algorithms. We compare the result's accuracy with all instances where all labeled and unlabeled instances are used. The result with only informative cases in which labeled and a small number of unlabeled instances are used. For the magic dataset in training instances, 17705 instances were labeled, and 18898 instances were unlabeled. According to our proposed method, we get the result using all instances where all unlabeled instances are being labeled. By applying SVM, we collect our informative

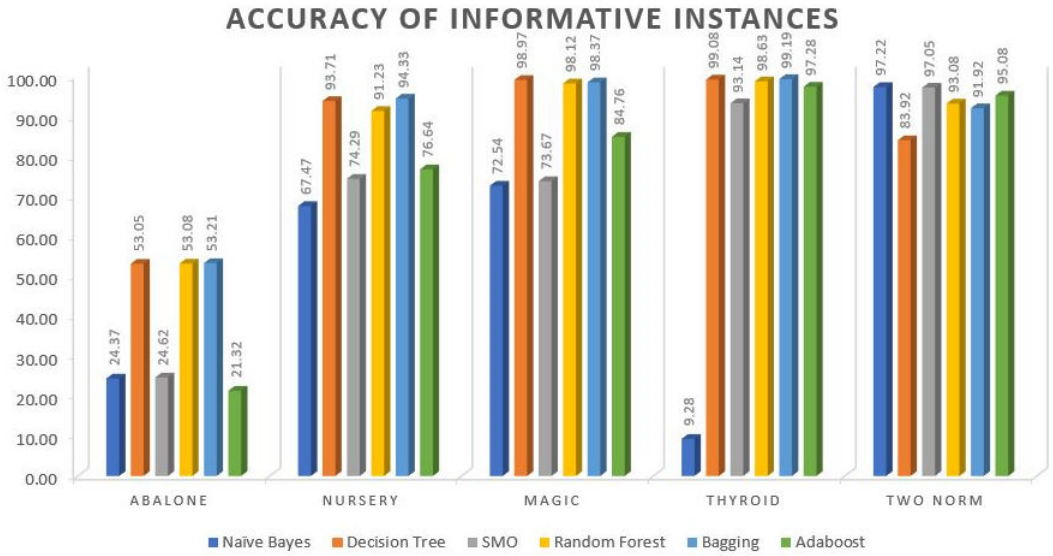

Fig. 2. The experimental results on the benchmark datasets using various classifiers. After evaluating the results, it is determined which classifiers are feasible to achieve the best outcome

instances, and for the magic dataset, we can reduce almost 97.92% unlabeled instances. That means only 2.08% unlabeled instances (almost 3931 from 18898) need to be labeled for acquiring the nearly same or better accuracy. Therefore, we can easily exclude 14967 unlabeled instances that have no effect in our model or for measuring accuracy for the magic datasets.

The reduced rate of non-informative instances is not the same for all datasets. Like, Abalone dataset, we can reduce 63.73% unlabeled instances. For the Nursery dataset, the reduced rate is 47.13%, for thyroid, it is 31.14%, and for the two norms dataset, we can reduce 62.03% unlabeled instances. The most reduced instances comes form Magic dataset which was 97.92%. From Table III we can see that the number of unlabeled training instances was large in magic dataset and its reduce rate also high than other dataset.

Moreover, Figure 2 represents the comparison of six existing machine learning methods which is applied in our proposed method of five different datasets. We can see from this experiment that the accuracy result of the other ML algorithms is good for different datasets. Like for Abalone, we achieve good results from bagging algorithm, for Nursery data good result comes from bagging, for Magic the result of the decision tree is good, bagging providing the good result for Thyroid and for Two Norms dataset the result of Naïve Bayes algorithm is good than another ML algorithm. As bagging algorithms provides good result for three datasets and DT and Naïve Bayes provides good result for one dataset so we can say that in this experiment bagging ML algorithms works better than other five ML algorithms.

In summary, studying this result, we can conclude that from any semi-supervised big dataset we do not need to label all unlabeled instances for acquiring the better result. And we also do not need to think which instances are suitable for keeping and which one is bad for rejection. As we are applying SVM so automatically those unlabeled instances are collected as informative instances which are close to hyper-plane and this method is more efficient, less time consuming and cost-effective than other methods for collecting informative instances.

V. Conclusion and Future Work

This paper provides a new methodology involving Support Vector Machine (SVM) in Active Learning (AL) to manage the unavailable data challenge in mining big data. Five semi-supervised data set are decomposed for exploring relevant information regarding a timely solution. Afterwards, six-ML classifiers are applied for evaluation. Experimental findings demonstrate that the bagging classification offered a superior rating performance of 99.19%. The results also suggests that bagging is more capable of achieving higher classification efficiency than other classifiers. The proposed approach provides new stable and reliable approaches to semi-supervised big data mining to create a classification scheme with fewer chosen SVM-technical training instances due to the incompleteness, unstructured data and without sacrificing the classification outcome. We assume that the AL architecture that has been established will have an impact on AL research for big data.

In the future, an instance-weighting methodology and a gray box model of AL will be developed to classify descriptive data instances accurately and enhance the classification outcomes of semi-supervised learning. A deep learning algorithm will also be developed to extract features for big data efficiently.

References

[1] L. Korycki, A. Cano, and B. Krawczyk, "Active learning with abstaining classifiers for imbalanced drifting data streams," in *IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 2334–2343.

[2] F. K. Nakano, R. Cerri, and C. Vens, "Active learning for hierarchical multi-label classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1496–1530, 2020.

[3] U. Naseem, M. Khushi, S. K. Khan, K. Shaukat, and M. A. Moni, "A comparative analysis of active learning for biomedical text mining," *Applied System Innovation*, vol. 4, no. 1, p. 23, 2021.

[4] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 79.

[5] J. Kremer, K. Steenstrup Pedersen, and C. Igel, "Active learning with support vector machines," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 4, pp. 313–326, 2014.

[6] D. M. Farid, A. Nowé, and B. Manderick, "Combining boosting and active learning for mining multi-class genomic data," in *25th Belgian-Dutch Conference on Machine Learning (Benelearn), Kortrijk, Belgium*, 2016, pp. 1–2.

[7] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health information science and systems*, vol. 6, no. 1, pp. 1–12, 2018.

[8] M. R. Islam, A. R. M. Kamal, N. Sultana, R. Islam, M. A. Moni *et al.*, "Detecting depression using k-nearest neighbors (knn) classification technique," in *International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*. IEEE, 2018, pp. 1–4.

[9] K. M. Hasib, M. Iqbal, F. M. Shah, J. A. Mahmud, M. H. Popel, M. Showrov, I. Hossain, S. Ahmed, O. Rahman *et al.*, "A survey of methods for managing the classification and solution of data imbalance problem," *arXiv preprint arXiv:2012.11870*, 2020.

[10] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models," *arXiv preprint arXiv:2010.15036*, 2020.

[11] M. R. Islam, S. Liu, R. Biddle, I. Razzak, X. Wang, P. Tilocca, and G. Xu, "Discovering dynamic adverse behavior of policyholders in the life insurance industry," *Technological Forecasting and Social Change*, vol. 163, p. 120486, 2021.

[12] C. Campbell, N. Cristianini, A. Smola *et al.*, "Query learning with large margin classifiers," in *ICML*, vol. 20, no. 0, 2000, p. 0.

[13] M. H. Popel, K. M. Hasib, S. A. Habib, and F. M. Shah, "A hybrid under-sampling method (husboost) to classify imbalanced data," in *21st International Conference of Computer and Information Technology (ICCIT)*. IEEE, 2018, pp. 1–7.

[14] H. W. Ian and F. Eibe, "Data mining: Practical machine learning tools and techniques," 2005.

[15] D. M. Farid and C. M. Rahman, "Assigning weights to training instances increases classification accuracy," *International Journal of Data Mining & Knowledge Management Process*, vol. 3, no. 1, p. 13, 2013.

[16] M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: a survey and new perspectives," *Social Network Analysis and Mining*, vol. 10, no. 1, pp. 1–20, 2020.

[17] X.-Y. Zhang, S. Wang, and X. Yun, "Bidirectional active learning: A two-way exploration into unlabeled and labeled data set," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 12, pp. 3034–3044, 2015.

[18] A. Ghasemi, H. R. Rabiee, M. Fadaee, M. T. Manzuri, and M. H. Rohban, "Active learning from positive and unlabeled data," in *IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 244–250.

[19] A. Krishnakumar, "Active learning literature survey," *Tech. rep., Technical reports, University of California, Santa Cruz.*, vol. 42, 2007.

[20] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis," *IEEE Transactions on Computational Social Systems*, 2021.

[21] T. Reitmaier, A. Calma, and B. Sick, "Transductive active learning–a new semi-supervised learning approach based on iteratively refined generative models to capture structure in data," *Information Sciences*, vol. 293, pp. 275–298, 2015.

[22] U. Naseem, I. Razzak, and P. W. Eklund, "A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter," *Multimedia Tools and Applications*, pp. 1–28, 2020.

[23] I. Jahan, N. N. Sharmy, S. Jahan, F. A. Ebha, and N. J. Lisa, "Design of a secure sum protocol using trusted third party system for secure multi-party computations," in *International Conference on Information and Communication Systems (ICICS)*. IEEE, 2015, pp. 136–141.

[24] U. Aggarwal, A. Popescu, and C. Hudelot, "Active learning for imbalanced datasets," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1428–1437.

[25] C. H. Park and Y. Kang, "An active learning method for data streams with concept drift," in *IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 746–752.

[26] M. Zuluaga, A. Krause, and M. Püschel, "ε-pal: an active learning approach to the multi-objective optimization problem," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3619–3650, 2016.

[27] H. Xu, L. Li, and P. Guo, "Semi-supervised active learning algorithm for svms based on qbc and tri-training," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–14, 2020.

[28] B. Demir and L. Bruzzone, "A multiple criteria active learning method for support vector regression," *Pattern recognition*, vol. 47, no. 7, pp. 2558–2567, 2014.

[29] H. Guo and W. Wang, "An active learning-based svm multi-class classification model," *Pattern recognition*, vol. 48, no. 5, pp. 1577–1597, 2015.

[30] S. Jahan, S. Shatabda, and D. M. Farid, "Active learning for mining big data," in *21st International Conference of Computer and Information Technology (ICCIT)*. IEEE, 2018, pp. 1–6.

[31] Y. Xiao, Z. Chang, and B. Liu, "An efficient active learning method for multi-task learning," *Knowledge-Based Systems*, vol. 190, p. 105137, 2020.

[32] L. Ma, S. Destercke, and Y. Wang, "Online active learning of decision trees with evidential data," *Pattern Recognition*, vol. 52, pp. 33–45, 2016.

[33] C. Dou, D. Sun, G. Li, and R. K. Wong, "Active learning with density-initialized decision tree for record matching," in *International Conference on Scientific and Statistical Database Management*, 2017, pp. 1–12.

[34] M.-R. Bouguelia, Y. Belaïd, and A. Belaïd, "An adaptive streaming active learning strategy based on instance weighting," *Pattern Recognition Letters*, vol. 70, pp. 38–44, 2016.

[35] K. Dimitriadou, O. Papaemmanouil, and Y. Diao, "Aide: an active learning-based approach for interactive data exploration," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 2842–2856, 2016.

[36] S. B. Hamida, H. Hmida, A. Borgi, and M. Rukoz, "Adaptive sampling for active learning with genetic programming," *Cognitive Systems Research*, vol. 65, pp. 23–39, 2021.

[37] D. Dheeru and E. K. Taniskidou, "Uci machine learning repository," 2017.

[38] W. Hu, W. Hu, N. Xie, and S. Maybank, "Unsupervised active learning based on hierarchical graph-theoretic clustering," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 5, pp. 1147–1161, 2009.

[39] I. Žliobaitė, A. Bifet, B. Pfahringer, and G. Holmes, "Active learning with drifting streaming data," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 1, pp. 27–39, 2013.

[40] E. Lughofer, "Hybrid active learning for reducing the annotation effort of operators in classification systems," *Pattern Recognition*, vol. 45, no. 2, pp. 884–896, 2012.

[41] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *International Journal of Computer Vision*, vol. 113, no. 2, pp. 113–127, 2015.

[42] M. S. Hajmohammadi, R. Ibrahim, A. Selamat, and H. Fujita, "Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples," *Information sciences*, vol. 317, pp. 67–77, 2015.

[43] P. Jain and A. Kapoor, "Active learning for large multi-class problems," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 762–769.