# Big Data Mining in the Presence of Concept Drifting


Tabassum Siraj (011 141 027)
Warafta Rasul (011 141 150)
Meher Afroz Chowdhury (011 141 099)
Efrana Jannat (011 141 127)
Department of Computer Science and Engineering
United International University

A thesis submitted for the degree of
*BSc in Computer Science & Engineering*

March 2019

# Abstract

Concept drift in big data mining is an absolute, highly demanding research issue in this digital era. A concept in "concept drift" involved in the field of data mining (DM) and machine learning (ML) studies is referred to the relationship between input features and class variables. In real-life classification problems, a concept can be changed or new concepts may appear over the time. Innumerable data mining classification models, new methods are propounded every day to enhance the solution to this raised issue. In this paper, we have extended our vision to attempt an investigation in resolving the issue of concept drift detection, and carried out analysis for creating a better method to adapt to the newly arriving concepts. We have addressed the issue of concept drifting for mining big data, and presented an evolutionary, concept-adaptive, rule-based approach that classifies novel class instances in "concept drifting". The proposed method clusters the big data and extracts decision rules from each cluster for classifying instances with existing classes. For classifying new unlabelled instances, the proposed method first examines and ensures if the new instances belong to the clusters in existence or not. If the new instance does not appear to belong to the existing clusters, we have considered this instance as a novel class instance. Then, we have extracted new classification rules from the new instances data and added these new rules with existing rules. Performance evaluation tests of the proposed method have been conducted using a number of datasets provided by UCI (University of California, Irvine) machine learning repository. The calculated result proves proposed approach as an effective and efficient means of detecting novel class, which implies concept drift identification at the same time.

We inscribe this paper to our devoted parents and esteemed teachers.

## Acknowledgements

We would earnestly commend our honoured supervisor Dr. Dewan Md. Farid, Associate Professor of United International University, with utmost gratitude for his guidance and direction towards us in working on this paper. His perseverance, patience and constant persuasion has led us to work with success. We would also eagerly express our sincerest respect, recognition and appreciation towards our Professor Dr. Chowdhury Mofizur Rahman, Vice Chancellor, and Dr. Swakkhar Shatabda, Associate Professor and Undergraduate Program Coordinator, United International University for their experiences, their fruitful and helpful guidance towards this area of research. Also, we would not want to miss the chance to thank all our teachers, who played important roles in introducing Data Science to us. Finally, we acknowledge our beloved institution, United International University, for being such an eminent platform and possessing the position of our deepest admiration and devotion.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# Chapter 1

# Introduction

## 1.1 Motivation

Big Data analysis becomes an attractive research area in ML( machine learning) and DM( data mining ) at present time [1–3]. Detection of concept drift in streaming big data captured so much attention of intelligence computational researchers [4, 5]. In big data a large volume of new data generated continuously with time that cannot be restricted to any pre-defined order [6]. Big data are not always accurately stationary. A new data, in a non-stationary environment, could have different knowledge pattern and can probably change the data distribution and makes the current prediction model unsuitable. This circumstance is recognized as concept drift [7]. Concept drift occurs due to difference in the assignment of data in disparate time phases or change the pattern enciphers [8, 9]. Generally, big data should have the following four properties: (1) volume, (2) variety, (3) velocity, and (4) veracity. Volume refers to the amount of generated and stored data. Variety refers to the number of types and nature of data. Velocity is at which pace the data is being processed. Veracity is the quality of data that affects the accurate analysis.

The inauguration of big data applications with huge volumes of high-dimensional instances, often become accessible in stream mode, and impose major obstacles to all learning tasks, among which classification is significant. [10]. On account of numerous reasons, where absent features or acquisition failures are included, the data vectors that are available might often have missing entries, which due to improper treatment, can result in classification. High-dimensional data normally exhibit recognizable patterns

that can be leveraged to effectively embed them into low-dimensional spaces. Classification is the function of learning such embedding maps from a finite number of training data with their correct class labels. Big Data applications are rising high and being researched by the region of computer science, in which applications require distributed and present classification function and pattern recognition of huge data sets accordingly which are collected from many different sensor networks, medical data, online form registrations, video and imaging systems, etc. In our paper, we have proposed a new approach for big data mining in the presence of concept drift. Our proposed approach applies clustering technique with adaptive rule-based classifier to classify the unlabelled instances in big data and also predict the novel class instances in big data. The suggested method clusters the big data initially and then extracts classification rules from each of the cluster. To classify new unlabelled instances, it checks if the instances belong to the existing clusters or not. If the new instance does not belong to any of the existing clusters, it indicates that concept drift happened. Afterwards, it extracted new classification rules from the new instances and added these new rules with existing rules. We have assessed the execution of propound formula on conventional datasets from University of California, Irvine (UCI) machine learning repository [11] and the results that are originated from experiments proved the methodology we suggested as effective and efficient in classifying data with identifying novel class instances in data.

## 1.2 Thesis Contribution

In our paper, we have proposed a new approach for big data mining in the presence of concept drift. Our proposed approach applies clustering technique with adaptive rule-based classifier to classify the unlabelled instances in big data and also predict the novel class instances in big data. The suggested method clusters the big data initially and then extracts classification rules from each of the cluster. To classify new unlabelled instances, it checks if the instances belong to the existing clusters or not. If the new instance does not belong to any of the existing clusters, it indicates that concept drift happened. Afterwards, it extracted new classification rules from the new instances and added these new rules with existing rules. We have assessed the execution of propound formula on conventional datasets from University of California, Irvine (UCI) machine

learning repository [11] and the results that are originated from experiments proved the methodology we suggested as effective and efficient in classifying data with identifying novel class instances in data.

## 1.3 Organization of the Thesis

The proposition is explored in the following order:

**Chapter 2** This chapter includes the background of our work with all the related papers we took guidance from.

**Chapter 3** This section discusses about our proposed approach in details and explains the basic theories about concept drift.

**Chapter 4** This chapter explores everything about the experiments, results and analysis of our work.

**Chapter 5** This section concludes the paper with our future plans with this work.

# Chapter 2

# Related Work

In 2017, Liu et al. [12] proposed a concept drift detection method on the basis of subspace learning that mostly depends on angle optimized global embedding (AOGE) and principal component analysis (PCA). AOGE and PCA algorithms analyze the projection angle and the projection variance for concept drift identification. With observation, the change of objective functions for data streams, concept drift is distinguished. If concept drift is present, it is re-organized, based upon subspace learning which utilizes the new data stream patch. After that, the data streams are processed with dimension reduction for classification procedure. Lu et al. [13] proposed a chunk-based incremental learning method called Dynamic Weighted Majority for Imbalanced Learning (DWMIL), which analyzes concept drift in data streams and class imbalance problem. DWMIL avails an ensemble framework that assigns the weight of the base classifiers dynamically with respect to their performance on current data chunk. DWMIL is an incremental learning model and achieves high performance by keeping limited number of classifiers. In 2016, Haque et al. [14] proposed an active (semi-supervised) framework to classify data streams. It can detect concept drift and ensures chunk limit in a robust way by identifying crucial changes in classifier confidence. Additionally, it utilizes confidence scores to acutely choose specified number of data instances that arrives from the newest chunk as flags to update the classifiers later. They hypothesized the use of confidence estimators and the impact of concept drift on classifier confidence.

In 2015, Yang and Fong [15] used three representative singletree induction algorithms (VFDT, ADWIN, iOVFDT) to explore the occurrence of concept drift. iOVFDT balances accuracy, tree size, and learning speed with a compact size of model and less

usage of memory. The results conclude that ioVFDT displayed better performance compared to the other two aforementioned algorithms, due to its adaptive tie threshold, making it suitable for both artificial and real world concept drift data streams. In 2015, Ortiz et al. [16] created an approach named Fast Adapting Ensemble (FAE), which is an ensemble method designed to adapt to concept drifts rapidly and specifically deals with recurring concepts with expertise. FAE harbors a set of inactive base classifiers that change their state to being active within a very short span as the concept they represent re-emerges. FAE uses a drift detector (this helps to process the abrupt concept drifts) to conclude when to construct and include a new base classifier. FAE also uses a weighted majority vote to acquire the global ensemble decision and establishes a formula for adjusting the base classifiers weights, which permits the algorithm to expand or reduce the weights according to their performances. Krawczyk, and Woźniak [17] proposed a novel modification of weighted one-class support vector machine transformed to the relentless streaming data analysis. Their hypothesis prioritized the improved alterations of weights used by WOSVM. They implemented the incremental training of weights, which was followed by forgetting mechanism. They finalized with the theory that the efficient one-class classification model for data streams along with concept drift would be reprized from combination of incremental learning assigned with highest weights to object that arrive from new data chunk with forgetting by sigmoid based function.

In 2015, Wang and Abraham [18] concentrated on concept drift detection that affects binary classification models and demonstrates a framework (Linear Four Rates, LFR), which can identify the concept drift and distinguishes the updated data points of the new concept. The wide-range vision of LFR makes it possible to function for both batch and stream datasets, which are imbalanced datasets. It can also use user-specified parameters that are automatically easy to understand, different from other methods of detecting concept drift. LFR stands out with better performance than the existing popular approaches in respect of early concept drift detection, low false alarm rate and high detection rate across the various sorts of concept drifts. Mirza et al. [19] presented an Ensemble of Subset Online Sequential Learning Machine (ESOS-ELM) to learn from class imbalanced data that occurs from drifting data streams. ESOS-ELM is an ensemble approach for classification in imbalanced environment. It's also an ELM-Store module, which harbours information of old concepts and a change detector that

can instantly detect concept drifts. In ESOS-ELM, the ensemble is processed with the equitable subsets of streaming data. In 2015, Kanoun and Schaar [20] addressed the limitations of streaming big data by suggesting an online energy-efficient scheduler that can promptly learn the environmental changes by reinforcement learning techniques to increase the QoS (i.e. throughput and output quality) even with restrictions for resource use. They identified the scheduling problem as a Stochastic Shortest Path problem (SSP) and introduced a learning algorithm that reinforces itself to adapt to environmental changes and resolve this barrier even with recurrence of concept drift. L. Rutkowski et al.(2015) [21] proffered a new criteria that apply the theory of splitting nodes of decision trees for classifying data streams. The nodes are split according to the count of impurity, otherwise addressed popularly as misclassification error. The proposed measure or criteria observes if the selected ideal one among all the resident properties of current series of elements in node is equally the best in the entire stream or not. Also, this criterion in question was blended along with the criterion developed from Gini index. The synthesized algorithm reproduced quite efficient results while processing data streams in different environments.

In 2013, Farid et al. [22] introduced an adaptive ensemble learning classifier to handle deficiencies that particularize concept drift, infinitesimal length, minimal labeled data, and concept evolution for detecting novel class in data streams with concept drifting. The suggested model frequently upgrades itself with freshly arrived data points to keep up with most up-to-date concepts in data stream. The majority of weighted predictions among the classifiers in model classifies unlabeled observations. The model, M, labels the instances present in each sub-stream and a new classifier are adapted with the most contemporary dataset.

# Chapter 3

# Proposed Method

## 3.1 Concept Drift

Big Data is a vast source of instances that appear in an organized order. So, it enforces certain restraints on the learning system, which are unable in accomplishing the destined result with the help of the orthodox algorithms from this realm. Suppose, a stream of data made up of some states which are grouped, $G = \{G_1, G_2, \cdots, G_n\}$, where $G_i$ is originated by a diffusion $D_i$. If we consider streaming data which is stationary where the order of instances are symbolized as $D_j = D_{j+1}$, the transition can be marked as $G_j \rightarrow G_{j+1}$. Meanwhile, the behaviour of data may change in time due to many reasons in our modern and realistic problems. This event is addressed as concept drift. The fundamental problem is the alteration in the input and output(I/O) with time. A "drift" or change from a standard concept occurs because of the "subtle changes" in attribute values of big data. Concept drift possibly appears where data is collected over time, with or without predictions on supervised learning problems. Typically, these are known as online learning problems, where change is expected in the data with time. Concept drift is basically bifurcated into (1) real concept drift, and (2) virtual concept drift. The type of concept drift that can alter the decision boundaries also known as posterior probabilities and can also have an effect on unconditional probability density function is called real concept drift which imposes a threat on learning system. Another type of concept drift is virtual that does not affect the decision boundaries, but transform conditional probability density functions, thus it cannot overpower the recently used learning methods. But, it is mandatory for us to identify concept drift.

The four prior applications that are popularly used to deal with concept drift are: concept drift detectors, online learners, sliding windows, and ensemble learners. We might need different concept drift detection and handling procedures depending on different situations. Several ways to handle concept drift are: periodically update, detect and choose model, learn the change, weight data, and data preparation.

## 3.2 Proposed Method

If concept drift is existent in big data, classifying it will be quite an arduous task. Because, machine learning models are trained on fix number of class labels and learning models misclassify the novel class instance as existing classes in classification process. It's very common in real-life classification problem that observations with latest class may manifest itself over the time. A rule-based learning model has been propounded to classify large data with clustering technique that can classify unknown instances and detect novel class instances. The proposed method divides the big data into small sub-datasets, so that each of the sub-data can fit into the computer memory and easy to manage. Then we cluster the each sub-data into several clusters. Data clustering is an approach of segmenting data observations that are similar to each other [23]. Then classification rules are extracted from clusters using decision tree (C4.5) algorithm [24], as extracting classification rules from trees are very well known approach [25]. C4.5 decision tree algorithm is a top-down divide and conquer recursive process, which classify the unseen instances based on the several seen instances [26, 27]. C4.5 decision tree algorithm uses Gain Ratio to select the best splitting attribute. Generation of a classification rule from individual leaf of the decision tree. From the root to leaf defines a path that corresponds to a rule-set. The perk of extracting rules from tree is that it's easy to marge and can be executed in any order. Latest rules can be merged to previously extracted rules. Where as decision tree merging is quite complex process. For classifying new unlabelled instances, the proposed method first checks the new instances belong to the existing clusters or not. If the contemporary observations does not fit in to the existing clusters, then we have considered this observations as a novel class instance. Then, we have extracted new classification rules from the new instances and added these new rules with existing rules. Table 3.1 presents the frequently used notations and terminologies in this paper. Fig. 3.1 describes the proposed method

for mining big data in concept drifting employing clustering with rule-based classifier, which is illustrated in algorithm 1.

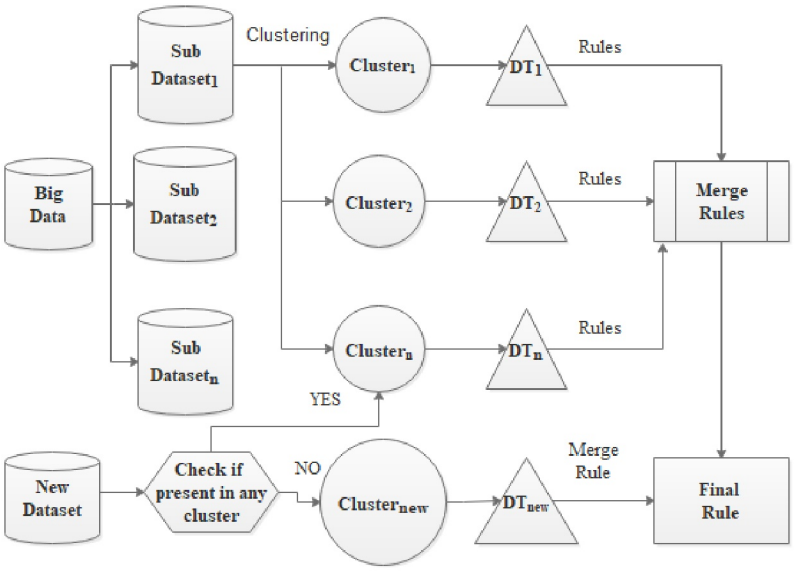


**Figure 3.1:** Propounded method for mining big data in presence of concept drift.

**Table 3.1:** Frequently used notations and terminologies.

| Notation | Terminology |
|---|---|
| $D_{Big}$ | Big Data |
| $D_i$ | Small data that into the memory |
| $C_j$ | Cluster |
| $DT_j$ | Decision tree |
| $rule$ | Classification rule |
| $x_{New}$ | New data point/ instance |
| $x_{Novel}$ | New instance with novel class |

**Algorithm 1** Big Data Mining with Concept Drift

**Input:** *Big Data, $D_{Big}$; C4.5 Classifier;*

**Output:** *Rule-set*; // Classification rules.

**Method:**

1: *Rule-set* $= \emptyset$;
2: divide $D_{Big}$ into small sub-data sets $D_1, D_2, \cdots, D_N$:
3: **for** i = 1 to N **do**
4: cluster $D_i$ into several clusters $C_1, C_2, \cdots, C_M$;
5: **for** j = 1 to M **do**
6: build decision tree, $DT_j \leftarrow C_j$ using C4.5 classifier;
7: **if** $error(DT_j) \leq 0.5$ **then**
8: *Rules* $\leftarrow DT_j$; // extracting the rules from $DT_j$.
9: **end if**
10: **end for**
11: *Rule-set* $=$ *Rule-set* $\cup$ *Rules*;
12: **end for**

To classify new instance, $x_{New}$:

1: **if** $x_{New}$ belongs to any $C_j$ **then**
2: classify $x_{New}$ with existing rules;
3: **else**
4: extract *Rule* $\leftarrow x_{New}$;
5: **end if**
6: add new rule with existing rules;

# Chapter 4

# Experimental Analysis

To provide valid assessment of our hypothesis, we conducted experiments with big data and demonstrated the performance of our improved method. We implemented our proposed approach in Java. We introduced the adaptive rule method for detection of novel class, instigating concept drift in Big Data.

## 4.1 Datasets

Datasets are assembled from UCI machine learning repository [11]. Table 4.1 provides a brief interpretation of the datasets. NSL-KDD dataset is used quite frequently as a standard for assessment of concept drift detection approaches. This dataset is the updated form that has solved the shortcomings present in the KDD99 dataset. In Table 4.3, we presented the variety instructed samples of different genre used in the NSL-KDD dataset. There are 683 data instances and 35 categorical attributes with 19 class values in the large soybean dataset. Another dataset employed in our work is image segmentation data of 7 outdoor images, where instances were drawn randomly. Categorization with respect to each pixel was generated from handpicked images. The 3X3 region was assigned for every instances in the employed dataset. The basic target for using this specific dataset is accomplishing a factual reasoning for image segmentation research. This database has 19 features and 7 categories and 2310 data instances.

**Table 4.1:** Dataset details.

| Dataset Name | Features | Category | Observation | Target |
|---|---|---|---|---|
| NSL-KDD | 41 | Real & Categorical | 25192 | 23 |
| Soybean-Dataset | 35 | Categorical | 683 | 19 |
| Segment-Test | 19 | Numeric | 2310 | 7 |

**Table 4.2:** The NSL-KDD dataset.

| Category | Target | Observation |
|---|---|---|
| Normal | normal | 13449 |
| Denial_of_Service (DoS) | neptune, land, back teardrop, smurf, pod | 9234 |
| Remote_to_User (R_2_U) | warezmaster, guess_passwd, impa, spy, ftp_write, warezclient, multihop, phf | 209 |
| User_to_Root (U_2_R) | rootkit, buffe_overflow, loadmodule, perl | 11 |
| Probing | protsweep, nmap ipsweep, satan | 2289 |
| | Total = 23 | Total = 25192 |

### 4.1.1 NSL-KDD Dataset

NSL-KDD dataset is used quite frequently as a standard for assessment of concept drift detection approaches. The NSL-KDD dataset is the new version of the KDD99 dataset, which has been updated as an improved version that has solved the shortcomings present in the previous one (Travallee, Bagheri, Lu, & Ghorbani, 2009). Despite the issues raised by McHugh (McHugh, 2000), the NSL-KDD dataset has important merit, that is, the training and test datasets are free of redundancy. Each record in this data consists of 41 attributes and one class label. The number of training instances of normal and intrusion classes in NSL-KDD dataset is presented in Table 4.3.

### 4.1.2 Large soybean Dataset

One of the other datasets enlisted for our analysis is a database of large soybeans. There are 19 classes in this dataset. Although there are 35 categorical attributes present in this data, some appear as nominal and some as ordered. We have nominalized every attribute using string values instead of numerical values. There are 307 data instances

**Table 4.3:** NSL-KDD Dataset.

| Types | Classes | Instances |
|---|---|---|
| Normal | normal | 13449 |
| Denial of Service (DoS) | back, land, neptune, pod, smurf, tenddrop | 9234 |
| Remote to User (R2U) | ftp_write, guess_passwd, impa, multihop. phf, spy, warezclient, warezmaster | 209 |
| User to Root (U2R) | buffe_overflow, perl, loadmodule. rootkit | 11 |
| Probing | ipsweep, nmap, satan, protsweep | 2289 |
| | Total = 23 | Total = 25192 |

here in total.

### 4.1.3 Image Segmentation Dataset

Another dataset employed in our work is a database of 7 outdoor images where instances were drawn randomly. The images were hand-segmented to create a classification for every pixel and each instance is a 3X3 region. The basic target for using this dataset is to accomplish an empirical basis result to research on image segmentation, which we used here for the evaluation of our model detecting concept drift. There are 19 attributes and 7 class values and 2310 data instances in this database.

### 4.1.4 Experimental Setup

We have enacted the proposed hypothesis in Java using NetBeans IDE 8.2. NetBeans is an open-source integrated development environment that gives assistance to the development for all types of Java application. The code for decision tree (C4.5) is adopted from Weka 3.8, which is a data mining software in Java [28].

## 4.2 Experimental Results

Tables 4.4 to 4.6 describe how the presence of novel classes have an impact on the performance of the datasets that we have used. The outcomes represent that our proposed method, which is taught with a stable amount of sets, displays satisfactory

fitness in learning and classifying new concepts, and overshadows the conventional C4.5 decision tree induction algorithm. Table 4.7 shows the summary of some useful metrics of success of prediction when the classes are very imbalanced. It presented the values of the metrics called Precision, Recall and F-score respectively on each of the datasets that are shown in Eq. 4.1 to Eq. 4.3. Fig. 4.1 to Fig. 4.3 show the classification accuracy between proposed method and C4.5 classifier in the appearance in novel classes on the NSL-KDD, Soybean, Segment-test datasets respectively.

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.2}$$

$$F - score = \frac{2 \times precision \times recall}{precision + recall} \tag{4.3}$$

**Table 4.4:** Performance of propounded method in the novel class manifestation using the NSL-KDD dataset.

| Training Observations | Testing Observations | Enduring Classes | Novel Classes | Accuracy |
|---|---|---|---|---|
| 13449 | 9234 | 1 | 6 | 85.68 % |
| 22683 | 220 | 7 | 10 | 75.64 % |
| 22903 | 2289 | 19 | 4 | 66.85 % |

**Table 4.5:** Performance of propounded method in the novel class manifestation using the Soybean dataset.

| Training Observations | Testing Observations | Enduring Classes | Novel Classes | Accuracy |
|---|---|---|---|---|
| 192 | 90 | 5 | 2 | 86.97 % |
| 364 | 172 | 10 | 3 | 85.16 % |
| 630 | 53 | 15 | 4 | 83.03 % |

**Table 4.6:** Performance of propounded method in the novel class manifestation using the Segment-test dataset.

| Training Observations | Testing Observations | Enduring Classes | Novel Classes | Accuracy |
|---|---|---|---|---|
| 180 | 50 | 6 | 1 | 90.85 % |
| 987 | 315 | 5 | 2 | 78.12 % |
| 1343 | 417 | 4 | 3 | 65.57 % |

**Table 4.7:** The standard optimum of the precision, recall and f-measure of proposed method on datasets derived from calculations.

| Dataset | Precision | Recall | F-measure |
|---|---|---|---|
| The NSL-KDD | 0.88 | 0.72 | 0.79 |
| The Soybean | 0.78 | 0.69 | 0.73 |
| The Segment-test | 0.82 | 0.95 | 0.88 |

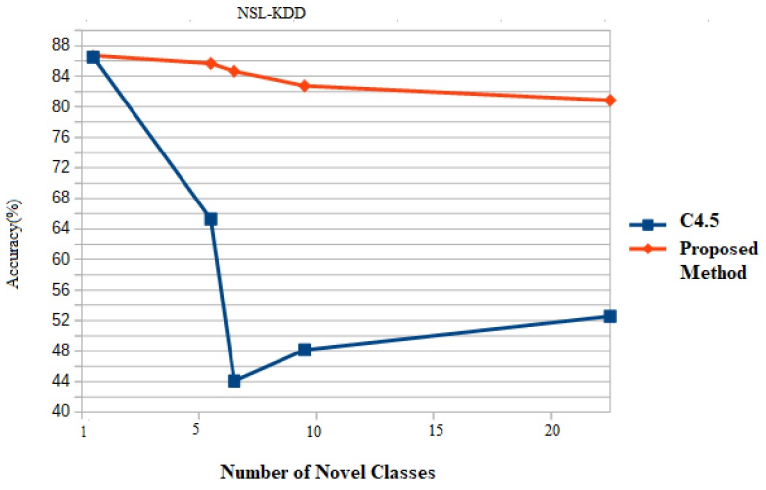


**Figure 4.1:** Performance of propounded method and C4.5 classifier in the novel class manifestation using the NSL-KDD dataset.

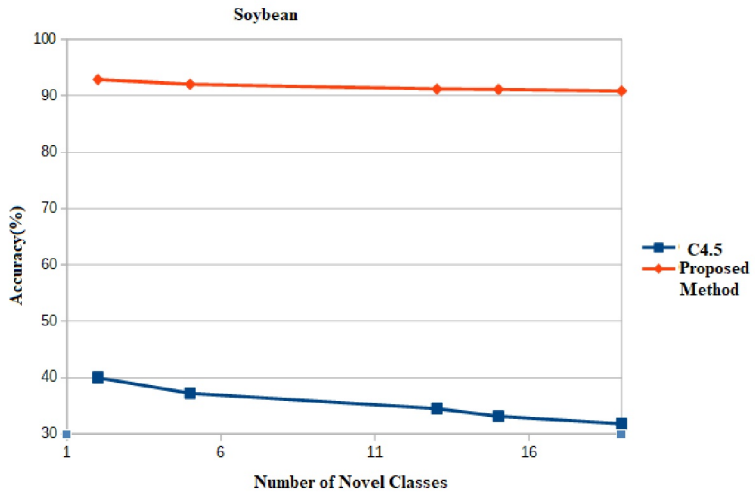


**Figure 4.2:** Performance of propounded method and C4.5 classifier in the novel class manifestation using the Soybean dataset.

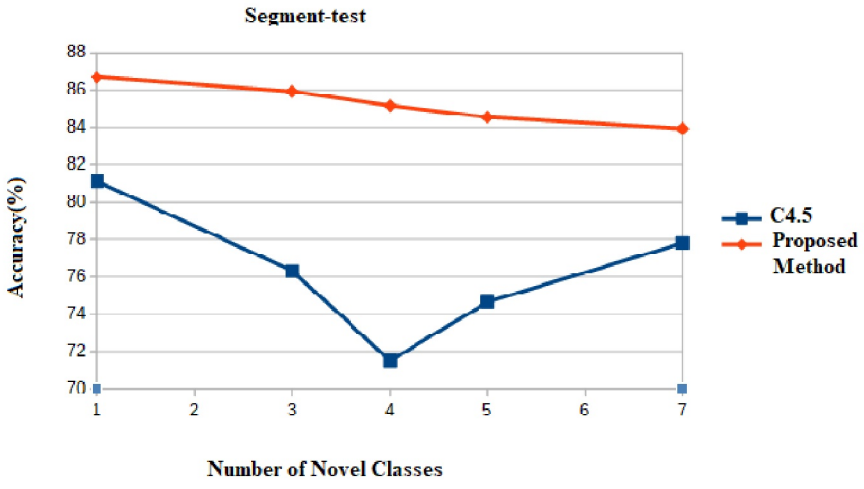


**Figure 4.3:** Performance of propounded method and C4.5 classifier in the novel class manifestation using the Segment test dataset.

# Chapter 5

# Conclusions and Future Work

## 5.1 Conclusions

In this paper, we successfully attempted in introducing a standard approach for mining big data in the presence of concept drift employing clustering with rule-based classifier. A major problem of dealing with big data is managing and storing it into machine's computer memory. The most popular technique is the partitioning of big data into small sub-datasets, so that each sub-data can easily be alloted to the memory of a computer. The proposed system clusters each of the sub-datasets to group the instances that are similar to each other. We then extracted classification rules from each cluster using decision tree induction technique. Merging decision trees or any other learning classifiers is a quite difficult task, whereas merging rules are quite easy owing to the fact that rules do not require any definite order of execution. We can add new rules to previously extant rules in our system without disrupting them. To find the concept drifting in data we have considered clustering technique. If any new data point does not belong to any clusters existing in our learning method, we have considered this instance as a novel class instance. Novel class instances come with new class labels in the data, generally in streaming data environment. Our conducted experiments demonstrated the performance of the proposed approach as efficacious in manner of generating better and accurate results with minimized error.

## 5.2 Future Work

We plan on extending our work in future under dynamic feature environment.

# Bibliography


[1] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314–347, August 2014. 1

[2] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, pp. 1–21, February 2015.

[3] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," *Big Data Research*, vol. 2, no. 2, pp. 59–64, June 2015. 1

[4] D. Brzezinski and J. Stefanowski, "Prequential AUC: properties of the area under the roc curve for data streams with concept drift," *Knowledge and Information Systems*, vol. 52, no. 2, pp. 531–562, 2017. 1

[5] J. Gao, W. Fan, J. Han, and P. S. Yu, "A general framework for mining concept-drifting data streams with skewed distributions," in *Proceedings of the 2007 SIAM International Conference on Data Mining*. SIAM, 2007, pp. 3–14. 1

[6] G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen, and F. Petitjean, "Characterizing concept drift," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 964–994, 2016. 1

[7] I. liobait, M. Pechenizkiy, and J. Gama, "An overview of concept drift applications," in *Big data analysis: new algorithms for a new society*. Springer, 2016, pp. 91–114. 1

[8] B. Ramakrishna and S. K. M. Rao, "Concept drift detection in data stream mining: The review of contemporary literature," *Global Journal of Computer Science and Technology*, 2017. 1

[9] G. Song, Y. Ye, H. Zhang, X. Xu, R. Y. Lau, and F. Liub, "Dynamic clustering forest: an ensemble framework to efficiently classify textual data stream with concept drift," *Information Sciences*, vol. 357, pp. 125–143, 2016. 1

[10] M. Hammoodi, F. Stahl, and M. Tennant, "Towards online concept drift detection with feature selection for data stream classification," 2016. 1

[11] M. Lichman, "UCI machine learning repository," University of California, Irvine, School of Information and Computer Sciences, 2013. [Online]. Available: http://archive.ics.uci.edu/ml 2, 3, 11

[12] S. Liu, L. Feng, J. Wu, G. Hou, and G. Han, "Concept drift detection for data stream learning based on angle optimised global embedding and principal component analysis in sensor networks," *Computers & Electrical Engineering*, vol. 58, pp. 327–336, 2017. 4

[13] Y. Lu, Y. ming Cheung, and Y. Y. Tang, "Dynamic weighted majority for incremental learning of imbalanced data streams with concept drift," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 2393–2399. 4

[14] A. Haque, L. Khan, M. Baron, B. Thuraisingham, and C. Aggarwal, "Efficient handling of concept drift and concept evolution over stream data," in *IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE, 2016, pp. 481–492. 4

[15] H. Yang and S. Fong, "Countering the concept-drift problems in big data by an incrementally optimised stream mining model," *Journal of Systems and Software*, vol. 102, pp. 158–166, 2015. 4

[16] A. O. Diaz, J. del Campo-Avila, G. Ramos-Jimenez, I. F. Blanco, Y. caballero Mota, A. M. Hechavarria, and R. Morales-Bueno, "Fast adapting ensemble: A new algorithm for mining data streams with concept drift," *The Scientific World Journal*, vol. 2015, 2015. 5

[17] B. Krawzczyk and M. Wozniak, "One-class classifiers with incremental learning and forgetting for data streams with concept drift," *Soft Computing*, vol. 19, no. 12, pp. 3387–3400, 2015. 5

[18] H. Wang and Z. Abraham, "Concept drift detection for streaming data," in *International Joint Conference on Neural Networks*. IEEE, 2015, pp. 1–9. 5

[19] B. Mirza, Z. Lin, and N. Liu, "Ensemble of subset online sequential extreme learning machine for class imbalance and concept drift," *Neurocomputing*, vol. 149, pp. 316–329, 2015. 5

[20] K. Kanoun and M. van der Schaar, "Big-data streaming applications scheduling with online learning and concept drift detection," in *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*. EDA Consortium, 2015, pp. 1547–1550. 6

[21] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "A new method for data stream mining based on the misclassification error," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1048–1059, 2015. 6

[22] D. M. Farid, L. Zhang, A. Hossain, C. M. Rahman, R. Strachan, G. Sexton, and K. Dahal, "An adaptive ensemble classifier for mining concept drifting data streams," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5895–5906, 2013. 6

[23] D. M. Farid, A. Nowé, and B. Manderick, "A feature grouping method for ensemble clustering of high-dimensional genomic big data," in *Future Technologies Conference*, San Francisco, United States, December 2016, pp. 260–268. 8

[24] J. R. Quinlan, "Induction of decision tree," *Machine Learning*, vol. 1, no. 1, pp. 81–106, August 1986. 8

[25] C. Apté and S. Weiss, "Data mining with decision trees and decision rules," *Future Generation Computer Systems*, vol. 13, no. 2-3, pp. 197–210, November 1997. 8

[26] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Chapman and Hall/CRC, 1984. 8

[27] J. R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993. 8

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, June 2009. 13