A REVIEW ON CLASSIFICATION OF DATA IMBALANCE USING BIGDATA

Ramasubramanian and Hariharan Shanmugasundaram

Department of Computer Engineering, Shadan Women's College of Engineering and Technology, Hyderabad, India

ABSTRACT

Classification is one among the data mining function that assigns items in a collection to target categories or collection of data to provide more accurate predictions and analysis. Classification using supervised learning method aims to identify the category of the class to which a new data will fall under. With the advancement of technology and increase in the generation of real-time data from various sources like Internet, IoT and Social media it needs more processing and challenging. One such challenge in processing is data imbalance. In the imbalanced dataset, majority classes dominate over minority classes causing the machine learning classifiers to be more biased towards majority classes and also most classification algorithm predicts all the test data with majority classes. In this paper, the author analysis the data imbalance models using big data and classification algorithm.

KEYWORDS

Data imbalance, Big data, IoT, Data analytics & Classification

1. INTRODUCTION

This section discusses the basic concepts of big data, history of big data, important components of big data, and application areas of big data. The term "data" means raw facts, which can exist in different forms. There are several types of data such as text data, image data, audio and video data. These data may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or a mechanical recording media device which is used for future processing. In the early 60's and 70's decades data were stored in the form of small text files and consumes less storage space; Later, the "flat file" concept was introduced and during 80's Relational Database Management systems (RDBMS) came into existence with the possibility of storing all varieties of data. The RDBMS has its own limitations in storing the amount of data. Only structured data were stored and processed by using SQL (Structured Query Language) tool. Then the concept of "Data Warehouse" has been introduced. Data warehouse is used to collect the data from various sections or departments of any business organization. The goal of Data warehousing is to capture the data from different sources for access and analysis rather than for transaction processing. Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) tools were used to make transaction and querying with the Data Warehouses.

Now, huge data have been generated every micro/nano seconds. The unprecedented data is called as "Big Data" and it leads to many new dimensions in the field of data processing. Big Data is a collection of data that is huge in volume, size and complexity. The main characteristics of Big data are 3 V's: Volume, Velocity and Variety[1] . There are two more V's are also included in recent technological development. These are Veracity and Value. This is shown in Figure 1.

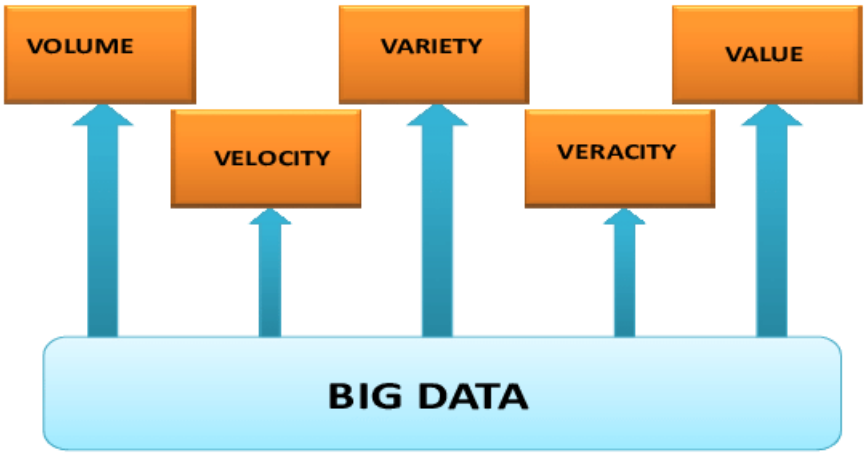

Figure 1. Block diagram of Big data

The veracity refers to the quality or truthfulness of the data under analysis. It also corresponds to the integrity of the data and value refers to the worth of the data being extracted and its impact on the business goals and profit of the organization. Most of the organizations now-a-days consider these two properties as Big Data characteristics [a].

The Big Data can exists in either structured, semi-structured or unstructured formats. Machine Learning (ML) models have been used to process for Big Data. There are many approaches and activities for processing Big Data and they may differ from application to application. The most general and widely used categories of activities involved with Big Data processing are[b]:

- Ingesting data into the system
- Persisting the data in storage
- Computing and Analyzing data
- Visualizing the results

Big Data Analytics (BDA) is the process of analysing large datasets, to obtain the hidden patterns contained in the data. These hidden patterns may comprise unknown correlations, user preferences, trends in buying behaviour etc. Analysing and identifying such information aids organizations in making informed decisions. With the improvements in distributed technologies like Hadoop, Spark and improvements in parallelization techniques, identifying hidden patterns in huge datasets (Big Data) becomes a possibility. There lists several applications of Big Data Analytics like weather predictions, medical report analysis, categorizing spam emails, oil-spill detection, network intrusion detection, students academic details prediction, faculty performance predictions, college/university selection and fraud detection in banking products where classification is behind the scenes and producing meaningful predictions [8].

In enterprises application Big Data poses a major challenge and research consideration in context of Business Intelligence and On Line Analytical Processing. There have been innumerable applications in e-commerce and on-line shopping portals. The banking sector performs its own credit score analysis for their existing customers using a wide range of data collected from savings, credits, mortgages and investment of high voluminous data. IoT is one another major thrust research area and variants of IoT Big Data application integrated with medical applications and health care analysis [9].

Sjaak Wolfert et al [10] presented a review using of Big Data analytics for Smart Farming. The potential impact or Big Data applications towards agricultural developments provided thoughts on agricultural based business players and researchers to find significant solution with involvement of robotics. This could also significantly enhances the ability of global food development systems to face the challenge of doubling the food supply by 2050.

The rest of the chapter is organized as follows: Section 2 describes the basic concepts of Big data. Section 3 discusses the data mining techniques used for data analytics. In section 4 brief on the ensemble modeling . In section 5, literature review is elaborated in detail related to the proposed topic of study. Section 6 highlights on the data imbalance classification scheme and solution from big data. Section 7 focus on discussion and finally conclusion is presented in section 8.

2. OVERVIEW OF BIG DATA

2.1. Evolution of Big data

In this modern electronic era, amount of volumes generated has demanded for distributed processing and further development using Hadoop framework by Google. Open source framework like Apache Hadoop [51] aims at operating on massive data for storage and processing. Hadoop mainly aims at enabling distributed processing of large amounts of data on large clusters of commodity servers. The Hadoop main component includes Hadoop Distributed File System (HDFS) and the processing component MapReduce. Several other components were also used in conjunction with Hadoop to enable high level processing (see Figure 2 for detailed components).

<table>
<tr><td colspan="7">Management and Monitoring (Ambari)</td></tr>
<tr><td rowspan="3">Coordination (Zookeeper)</td><td rowspan="3">Workflow and Scheduling (Oozie)</td><td>Scripting (Pig)</td><td>Machine Learning (Mahout)</td><td>Query (Hive)</td><td rowspan="2">No SQL database (HBase)</td><td rowspan="3">Data Integration (Sqoop/REST/ ODBC)</td></tr>
<tr><td colspan="3">Distributed Processing (Map Reduce)</td></tr>
<tr><td colspan="4">Distributed Storage (HDFS)</td></tr>
</table>

Figure 2. Components of Hadoop

HDFS acts as the major data storage which can store huge amount of data of variable format with portability provided across heterogeneous hardware support and support for multiple operating systems. HDFS component enables data replication and usually distributed across different clusters, which can effectively enable reliable and quick data access. MapReduce is a Java-based framework and was created by Google which acts as the processing component which operates on the data from the HDFS store. MapReduce provides smoother Big Data processing with improved performance analysis by breaking up of larger data's into smaller ones. The basic principle of operation of a MapReduce architecture is that the "Map" job or the mapper passes the task for processing. This task is distributed to all the task trackers in the Hadoop cluster. The process is performed in-parallel and the results in the cluster nodes are collected during the "Reduce" phase. This finally results in a single value. Map Task in the Hadoop ecosystem is passed to the node containing the input data. It takes the input data, applies the task to each of the data by splitting them into independent chunks. The resultant chunks are passed as the input for Reduce Task. Similarly, Reduce task reads the input and combines the Mapped data tuples. The process is similar to aggregation and hence results in a smaller set of tuples. Both the input and output of the tasks are stored in the HDFS. MapReduce is responsible for job scheduling, monitoring and re-execution in case of a failed task.

2.1.1. Hadoop 2

The Hadoop 2 replaces the MapReduce processing component with YARN (Yet Another Resource Negotiator). YARN supports parallel processing of huge datasets and MapReduce provides the framework with efficient writing, fault and failure management. The major advantages of Hadoop architecture (Figure 3) are:

- **Scalability:** Allowing for scalability of data storage and support for component architectures.

- **Cost Efficiency:** Aim at resulting at huge reduction in cost per terabyte of storage with parallel computation for affordable users.

- **Flexibility:** Capacity to operate on any type of data with any number of formats from any number of data sources.

- **Fault tolerance:** Providing built-in fault tolerant feature that enables data recovery due to computation failures

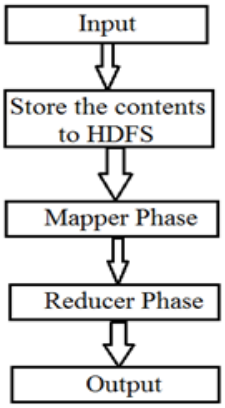

Figure 3. Block diagram of Map Reduce

The major components used in Hadoop eco systems are: Sqoop, Flume, Pig, HIVE, HBASE, SPARK which are detailed here in this section [52].

- **SQOOP :** An open source data tool for efficiently data processing between Hadoop file system and relational databases such as Teradata, Ntezza, Oracle, MySQL, Postfres and HSQLDB. In the file system, each of the rows will be treated as a record in HDFS. All records are stored as text data in text files or as binary data in Avro and Sequence files.

- **FLUME(Apache)** collects aggregate data and transporting large amounts of streaming data, such as log files, events etc., from several sources with centralized data repository. Flume is co be considered to be a more reliable, distributed, and easily configurable tool which provides services for handling the data from different data streaming sources like Twitter, Facebook, Cloud and Web Servers.

- **PIG** is a open source tool and was developed by Yahoo which is high-level declarative language similar to SQL and it accepts structured, semi structured and unstructured data. Pig is mainly used to reduce the development time of programs and also simplifies the

coding task. It also accepts the code written in other languages like Python, JavaScript, Ruby, PHP and Perl. Pig has its own scripting language called "***Pig Latin"***, which can store all commands and can be executed sequentially.

- **HIVE** developed by Facebook analyzes Terabytes of data which consists of huge data warehouse package and built on top of Hadoop. It allows conventional business intelligence (BI) applications to run queries on a Hadoop cluster.

- **HBASE** is an open source, distributed database which is written in Java and developed by Apache Software foundation which stores massive amounts of data from Tera to Peta bytes and it is used for online analytics applications.

Apache SPARK Spark is open source parallel processing framework developed as an ecosystem project for operating with the Hadoop architecture. Spark was designed for operating on large-scale applications that involve data analytics. The operations are usually performed across clustered computers. The major advantage of using a Spark architecture is that it can handle batch processing operations and also has the ability to handle real-time analytics on real-time streaming data. Spark Core API, is the major component of the Spark architecture.

3. DATA MINING TECHNIQUES FOR DATA ANALYTICS

Data mining models are classified as Supervised and Unsupervised based models. Supervised learning models are categorized into Classification models and Regression models. A supervised training model operates on data with labels, while unsupervised models operate on data without labels. In this paper, the authors focused on creating a supervised learning model for prediction/classification using Big Data. Classification is a categorization of data mining domain, which deals with supervised identification of class labels, given a large training dataset. Classifiers learn patterns contained in the given training data to predict unseen data.

Classification has several applications in the real-time systems beginning from image, sound and video classification, anomaly detection, Intrusion detection, bank fraud detection, detection of cancer, etc., The ratio between the number of instances in majority class and number of instances in minority classes is called the imbalance ratio. A class is considered to be balanced if its imbalance ratio is 1 and when increasing the imbalance ratio leads to increase in the ratio. There are several issues due to data imbalances [53]. The major issue is that data imbalances tend to bias the prediction process of a classifier, hence making the classifier more reliant towards predicting the majority classes. There are two ways to create models for imbalanced learning. The first is to modify the objective function of the model to concentrate more on the minority classes, the second is to resample the data to reduce imbalance. Resampling is performed by over-sampling the minority classes, under-sampling the majority classes or by hybrid sampling methods.

4. ENSEMBLE BASED MODELING

Ensemble learning is an essential part in Machine Learning (ML) algorithms used to enhance the predictive accuracy of classifier models. Machine learning methodologies are deployed in many real-time applications like email, online shopping, banking, gaming, internet telephony, streaming etc. They are complex scenarios, hence using single model-based mechanisms are not sufficient to capture the subtlety of the intrinsic patterns associated with these domains. Complex predictive structures such as ensemble models are required to provide acceptable performances. Ensemble techniques are popularly known as Classifier Combiner. In Predictive data modelling,

it is common to have training data and test data. Initially the training data will be trained with a base classifier to construct the prediction model. The learned model may be called as a ***hypothesis*** or a ***learner.*** Then the test data will be tested by using the learned or trained model to measure the prediction accuracy. The higher the prediction accuracy, the prediction model is more stable and accurate.

There are several classifiers are available like BF Tree, ID3, C4.5(J48 in WEKA Tool), Decision Stump, CART, Naïve Bayes, SVM, KNN, Neural Network etc.,. Decision tree based classifiers are called decision tree inducers. BF Tree, ID3, J48, Decision Stump and CART are examples of decision tree inducers. In the past few decades, ensemble methods are widely applied in many domains of sciences and provide promising research directions. Ensemble technique can be applied for both Classification and Clustering tasks. The basic idea of an ensemble method is to divide the whole dataset into so many subsets and then applying the same or different base classifiers on the subsets of a dataset and finally combining the accuracies produced by different subsets using majority voting method. Classification and Ensembling techniques fall under supervised learning methods. Classifier ensemble techniques like bagging, boosting, Random Forest, Random Subspace and Rotation Forest are used for classification of high dimensional dataset.

5. LITERATURE REVIEW

The purpose of the literature review is to identify the research gap between the chosen field. In recent years, huge volumes of data are used in many fields and these data are usually laden with imbalances. Using machine learning algorithm, classification of imbalanced data are processed by data cleaning techniques and are usually balanced by oversampling or under-sampling techniques. These models usually operate on clean data.

Li et al[11] proposed an adaptive swarm based classification model that effectively operates on imbalanced data. This model performs optimization using stochastic swarm fusion heuristics to perform optimization in the prediction process. Lee et al [12] describes an overlap sensitive classifier using support vector machines and k-nearest neighbor algorithms. A dissimilarity based classifier to handle imbalance data was proposed which features for elimination based model that eliminates unnecessary features to enhance the prediction process [13]. The feature selection and construction of dissimilarity space to provide effective predictions with imbalanced data classifier for binary classification was also proposed which is based on Neural Networks [14]. This has claimed significant improvements and performance with neural networks. This model initially creates classifier rules using neural networks and refines them using geometric mean.

Ensemble modelling is the process of creating multiple models and combining their individual predictions to formulate the final predictions. This procedure is done in several types such as, bagging, boosting, stacking and even with combination of several other techniques. A boosting based ensemble model concentrating on handling data imbalance levels was found to be more interesting [15]. This model utilizes the ROSE sampling technique as a base to handle imbalance, while high performances are provided by the boosting methodology. Ryan Hoens.T et al. [16] have suggested many sampling techniques, application of skew-intensive classifiers, Nura Muhammad Baba et al. [17] have done an extensive review study on current issues in ensemble methods and its applications covering various domains. The author emphasizes the incorporation of optimization algorithms like ACO, GA and PSO along with the ensemble methods would optimize the classification models.

Zhongbin Sun et al. [18] have proposed a novel study on ensemble method for classifying highly imbalanced data sets. The authors elaborate that the proposed method does not alter the original

class distribution ratio and does not suffer from information loss or unexpected mistakes that may be caused by other conventional methods via increasing the minority class instances or decreasing the majority ones. Uma et al. [19] presented a study on classifier ensemble design for imbalance data classification through a hybrid approach. The authors have mixed up both data level approach and also by incorporating classifier ensemble techniques to achieve better prediction performance. Area Under ROC Curve (AUC) has been suggested for measuring the performance accuracy. A Comparative analysis of predictive performance of various classifiers for multiclass problem was proposed in [20]. The author suggested that ROC curves are best tool for visualizing various classifiers behaviour. A comparison of multiple ensemble models for imbalance data prediction was proposed by Galar et al [21].

A study on bagging ensemble method combined with the widely popular decision tree inducers like BF Tree, J48, Decision stump and CART was carried out in [22]. This study was implemented in WEKA data mining tool and concluded that the use of base decision tree inducers along with bagging yields higher classification accuracy. Yongjun Piao et.al [23] proposed a framework to implement the ensemble method for classification of high dimensional data. Akhlaqur Rahman etal [24] presented a brief study on the various types of ensemble techniques and their applications . An ensemble classifier was designed in [25] using many data mining techniques to discretize the continuous values. However, the classifier consumed large amount of time for classification. Hina Anwar et al [26] proposed a global optimization ensemble model for classification to increase the accuracy. However, the classification accuracy was not improved to the expected level. Asma Gul et al [27] subsequently suggested a Random Forests (RF) algorithm termed as xRF, chooses features in learning RFs for high-dimensional data.

Alireza Osareh and Bita Shadgar [28] addressed gene classification problems with RotBoost ensemble methodology. Rotation Forest and AdaBoost techniques preserved features of ensemble architecture. For choosing the exact subset of informative genes, feature selection algorithms were taken. But, prediction accuracy remained unaddressed. Yongjun Piao et al [29] introduced a new ensemble method with portioning feature space of high-dimensional data. Thomas Dietterich.G [30] presented a study on ensemble methods in Machine learning. The study presents many methods for constructing ensembles using Bayesian averaging classifier. Several experimental analysis have been carried out to compare ensemble methods. It is observed that ADA BOOST performs well in many cases. The behavior of classification algorithms for imbalanced data sets were discussed in [31].

Sampling is the process of balancing the data prior to the model training phase. This aids us in using existing classifier models for processing, rather than designing a new model or architecture. Sampling can be performed as oversampling, under-sampling or hybrid sampling. Oversampling is the process of increasing the number of minority records to balance the number of majority records contained in the dataset. The process of additional data generation is performed by considering two instances and generating the new instance that corresponds to an intermediate point between the two instances. The major drawback of this model is that it tends to introduce multiple representations of similar data, sometimes leading to over fitting. Under sampling is the process of eliminating some majority class instances to balance the class levels of a dataset. This elimination can be random or selective. Random sampling has the risk of eliminating significant entries in the data, however, it is fast and computationally less complex. Selective sampling eliminates data based on certain rules. These rules are domain dependent and are defined by the domain expert. This type of sampling tends to increase computational complexity levels, due to the logical reasoning involved in the process. Hybrid sampling uses a combination of oversampling and under-sampling to balance the data. Currently, hybrid mechanisms are on the raise, as it provides the advantage of both the sampling models.

Several methods for handling imbalanced data with the help of sampling techniques [32]. A boosting based model was proposed which is a sampling based method that uses multiple sampling models to achieve the desired balance [15]. This model also proposes a boosting technique for effective prediction of classes. Yu et al [33] describes a credit classification method to handle imbalanced data, which is a rebalancing mechanism that utilizes bagging and re-sampling models to perform predictions. Borsos et al [34] describes the study on imbalanced data and metrics to measure their performances.

Cost sensitive learning is the prediction process which is extended as a solution during the classification of imbalanced nature of real-time data. Cao et al [35] proposed a cost sensitive SVM, an imbalance handling model by Wang et al. in [36] and a rule based learning model [37]. The impact of varied imbalance levels on datasets and processing them with different categories of classifiers in the preview of Big Data have been presented by the authors in [38]. Alberto Fernandez et al. [39] studied the diverse challenges and future direction behind the imbalanced Big Data classification. A bagged ensemble specifically designed for credit card fraud detection was proposed and this model proposes a bagging methodology for effective detection of fraudulent cases in credit card transactions [40]. A cost sensitive model to handle imbalance is a probability estimation based classifier model, aimed to effectively handle data imbalance [41].

The literature reviews presented in this section are categorized into four groups namely individual model based techniques, ensemble based techniques, sampling based techniques and cost sensitive learning models. All these techniques were applied particularly on imbalanced datasets with varied imbalance ratio. The prime motivation of all these literature leads to increase the prediction accuracy in their chosen task. Each technique and model has its own pros and cons.

6. DATA IMBALANCE CLASSIFICATION AND SOLUTION FROM BIG DATA

The data is the key asset for discussion and it need to be balanced one. Imbalanced data existence is one of the characteristics that exhibit a huge dominance over the other classes. In multi-class dataset, majority class represents highest number of instances and minority class refers to the class with the lowest instance levels. Imbalance is usually referred in terms of ratio between the total instances in the majority class to the number of instances in the minority class. Classifiers are based on supervised learning, i.e., it requires appropriate training prior to the prediction process, generally constructed with the prospect of balanced data. Usually classifiers assume that the data is balanced during the training phase. They require balanced representations of all the classes contained in a dataset to perform effectively [15]. In the current Big Data scenario, though the ratio between classes (imbalance level) remains the same as regular data, due to the hugeness of the data, actual instance levels tend to increase manifold. This leads to huge representation levels for major classes and low representation levels for minority classes. This leads to a major issue in terms of classifier overtraining for major classes and under training in terms of the minority classes [42]. The imbalance acceptance levels of classifiers vary between models [43].

The analysis of various classifiers algorithms in terms of imbalance levels, data size and their impact on classifier metrics are Naïve Bayes represents the probability based technique, Decision Tree and several other techniques. Decision Table is a rule based technique and Logistic Regression is for function based techniques. Selected classifiers are applied on the chosen datasets and the results are recorded in the **confusion matrix** also known as an **error matrix**. A confusion matrix is a table that is primarily used to describe the performance of a classification model on a set of test data for which the true values are known. **"classifier"** or **"algorithm"** or **"Learning algorithm"** are other name of classification model**.** The performance of a particular algorithm can be easily seen through the confusion matrix. A sample confusion matrix is shown in Table 1.

Table 1. Confusion Matrix

		Actual Value (As confirmed by experiment)	
		Positives	**Negatives**
Predicted Value (Predicted by the test)	**Positives**	**TP True Positive**	**FP False Positive**
	Nagatives	**FN False Nagative**	**TN True Negative**

The True positives **(TP)** mean those instances that are positives and are classified as positives. A false positive (FP) means those instances that are negatives and are classified as positives. False negative **(FN)** means those instances that are positives and are classified as negatives. True negatives **(TN)** mean those instances that are negatives and are classified as negatives. Standard classifier metrics used for classification analysis are shown in Table 2. All the metrics are derivable using data from the confusion matrix. The current analysis examines all the metrics in terms of imbalance and data hugeness.

There exists well known datasets which have different characteristics with representations namely, large to moderate size, binary/ multi-classes and with varying imbalance levels ranging from low (0) to very high (164091). KDD CUP 99 dataset was obtained from UCI repository [44], E-coli, Iris and Abalone were obtained from KEEL repository [45] and Bank data was obtained from [46]. The datasets used in this contribution serves for pilot study before entering into the deep sense of thought in the area of classification on imbalanced Big Data. Table 3 highlights on the characteristics of the various datasets and the nature of imbalance. Different experimental study has been done using the data sets available and have studied the impact of balanced and imbalanced data classification behavior.

Table 2. Performance Metrics

Metric	Formula
True Positive Rate (TPR)/ Sensitivity/ Recall	$\frac{TP}{TP + FN}$
True Negative Rate/ Specificity (TNR)	$\frac{TN}{FP + TN}$
False Positive Rate (FPR)	$\frac{FP}{FP + TN}$
False Negative Rate (FNR)	$\frac{FN}{FN + TP}$
Precision/ Positive Prediction Rate (PPR)	$\frac{TP}{TP + FP}$
Negative Prediction Rate (NPR)	$\frac{TN}{FN + TN}$
F-Measure	$\frac{2 * Precision * Recall}{Precision + Recall}$
Correct Classification % / Accuracy	$\frac{(TP + TN) * 100}{TP + FP + TN + FN}$
Incorrect Classification %	$\frac{(FP + FN) * 100}{TP + FP + TN + FN}$
Area Under Curve (AUC)	$\frac{1 + TPR - FPR}{2}$
Mathews Correlation Coefficient (MCC)	$\frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + TN) * (TP + FN) * (FP + TN) * (TN + FN)}}$

Table 3. Properties of Data Set

Data set	**Class**	**Nature of imbalance**
KDD CUP 99	Multi-Class (23)	High
E-Coli	Multi-Class (8)	Moderate
Iris	Multi-Class (3)	Balanced
Bank	Binary (2)	Low
Abalone	Binary (2)	High

7. DISCUSSIONS

The paper analyzed on the big data potential and its impact on several domains, tools and methods. Also the impact of existing classifiers on data with varied imbalance levels and evaluated their prediction accuracies using standard classifier metrics has been deeply elaborated in literature review, highlighting on several schemes in the proposed area of study. Ensemble is a form of supervised learning technique, which is a combination of learning algorithms. It process the utility of utilizing multiple algorithms to obtain better predictive performance compared to the usage of single learning techniques [47,48]. Hence they are not bound by the number or type of the individual components being used. Machine learning ensembles, in contrast to statistical ensembles utilizes finite models for building classifiers, however, they allow flexible structures to exist in the mechanism [21].

We have also observed that the study that exists in imbalance at several levels are examined empirically and observed making data specific algorithm with fine-tuning an impossible factor. Also the current study highlighted on the significance focusing on the identity of generic model that effectively handles imbalance data at all levels without the need for data specific fine-tuning. It is also observed that datasets with low level to moderate imbalance level is more effective on the study. When the datasets are operated on the high imbalance level, the model is observed to be slightly biased towards the majority classes. Also the observations recorded for Big Data applications doesn't considers data processing time during the study illustrations.

8. CONCLUSIONS

The paper has presented a detailed review on data imbalance techniques, tools and techniques for real-time data in big data applications prone to imbalanced data. The implications performed with such implications and observations which were experimented using several types of data are generally observed and the effectiveness of point of reliability is recorded. Also a detailed analysis of classifiers indicating the inefficiency of incorporating hugeness and imbalance levels also reviewed in detail. From the review outcomes, it is noted that the ensembles are proposed as probable solutions to the issue. The paper also has noted on the validity of handling imbalance, data hugeness, theoretical analysis and operational nature for data classification. The review of the work that exists shows the need of an improved algorithm design and design model to overcome on the several drawbacks and other design considerations with focus on classifying imbalance data for Big data applications.

ACKNOWLEDGEMENTS

The authors would like to extend sincere thanks to the management for providing us support and environment for carrying out the research and to other fellow colleagues for their support.

REFERENCES

[1] Ankita Karale, Bharathi Patil, "A Survey on Big Data", International Journal on Computer Science and Information Technology, Vol 2, Issue 4, August 2015.

[2] Hu H, Wen.Y, Chua.T.S and Li.X, "Toward scalable systems for big data analytics: A technology tutorial", IEEE Access, Vol. 2, pp. 652 – 687, July 2014.

[3] Navya Francis and Sheena Kurian.K, "Data Processing for Big Data Applications Using Hadoop Framework", International Journal of Advanced Research in Computer and Communication Engineering, Vol.4, Issue 3, March 2015.

[4] Narayana Bhagavatulal.V.S, Srinadh Raju.S, Sudhir Varma.S and Jose Moses.G, "A Survey Of Hadoop Ecosystem as a Handler of Bigdata", International Journal of Advanced Technology in Engineering and Science, Vol.4, Issue 08, August 2016.

[5] Fernández.V., García.A., Palade.S.V and Herrera.F, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics", Journal on Information Sciences, volume 250, pp. 113 – 141, 2013.

[6] Phua, C, Damminda A and Vincent L, "Minority report in fraud detection: classification of skewed data", ACM sig kdd explorations newsletter, Vol. 6, No.1, pp 50 – 59, 2004.

[7] Mazurowski.M.A., Habas.P.A., Zurada.J.M., Lo.J.Y., Baker.J.A, and Tourassi.G.D, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance", Neural networks, Vol. 2, No. 21, pp: 427 – 436, 2008.

[8] Chun Wei Tsai, Chin Feng Lai, Han Chieh Chao, Athanasios V. Vasilakos,"Big data Analytics: A Survey", Journal of Big Data, Vol. 2, No. 1, pp. 21, 2015.

[9] Min Chen , Shiwen Mao, Yunhao Liu, "Big Data: A Survey", Springer Science and Business Media, New York, 2014.

[10] Sjaak Wolfert, Lan Ge, Cor Verdouw and Marc-Jeroen Bogaardt, "Big Data in Smart Farming – A Review", Journal of Agricultural Systems, Vol. 153, pp. 69 – 80, 2017.

[11] Li.J., Fong.S., Wong.R.K, and Chu.V.W, "Adaptive multi-objective swarm fusion for imbalanced data classification", Journal of Information Fusion, Vol. 39, pp.1 – 24, 2018.
[12] Han Kyu Lee , Seoung Bum Kim , "An Overlap – Sensitive Margin Classifier for Imbalanced and overlapping Data", Expert Systems With Applications, DOI: 10.1016/j.eswa.2018.01.008, 2018.
[13] Zhang.X, Song.Q, Wang.G, Zhang.K, He.L, and Jia.X, "A dissimilarity-based imbalance data classification algorithm", Journal of Applied Intelligence, Vol.42, No. 3, pp. 544 – 565, 2015.
[14] Du, Jie, et al. "Post-boosting of classification boundary for imbalanced data using geometric mean", Journal of Neural Networks, Vol. 96, pp. 101 – 114, 2017.
[15] Gong.J. and Kim.H., "RHSBoost: Improving classification performance in imbalance data", Journal of Computational Statistics and Data Analysis, vol. 111, pp.1 – 13, 2017.
[16] Ryan Hoens.T and Nitesh V.Chawla,"Imbalanced Learning: Foundations, Algorithms, and Applications", John Wiley & Sons, Inc, 2013.
[17] Nura Muhammad Baba, Mokhairi Makhtar, Syed Abdullah Fadzli and Mohd Khalid wang, "Current Issues in Ensemble Methods and its Applications", Journal of Theoretical and Applied Information Technology, Vol. 81, No. 2, 2015.
[18] Zhongbin Sun, Qinbao Song, Xiaoyan Zhu, Heli Sun, BaowenXu,Yuming Zhou, "A novel ensemble method for classifying imbalanced data", Pattern Recognition, Vol. 48, pp. 1623 – 1637, 2015.
[19] Uma R.S, Suresh and N.Mali, "Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach", Proceedings of Computer Science, Vol. 85, pp. 725 – 732, 2016.
[20] Ramaswami.M, "Validating Predictive Performance of Classifier Models for Multiclass Problem in Educational Data Mining", International Journal of Computer Science Issues, Vol.11, Issue 5, No.2, 2014.
[21] Galar.M., Fernandez.A., Barrenechea.E., Bustince.H and Herrera.F.,"A review on ensembles for the class imbalance problem: bagging, boosting, and hybrid-based approaches", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 42, Issue. 4, pp. 463 – 484, 2012.
[22] Nikita Joshi and Shweta Srivastava, "Improving Classification Accuracy Using Ensemble Learning Technique (Using Different Decision Trees)", International Journal of Computer Science and Mobile Computing, Vol. 3, No. 5, pp. 727 – 732, 2014.
[23] Yongjun Piao, Hyun Woo Park, Cheng Hao Jin, Keun Ho Ryu, "Ensemble Methods for Classification of High Dimensional Data", International Conference on Big Data and Smart Computing (BIGCOMP), IEEE 2014.
[24] Akhlaqur Rahman, SumairaTasnim, "Ensemble Classifiers and their Applications : A Review", International Journal of Computer Trends and Technology(IJCTT), Vol. 10, No. 1, 2014.
[25] Nan-Chen Hsieh and Lun-Ping Hung, "A data driven ensemble classifier for credit scoring analysis", Expert Systems with Applications, Elsevier, Vol. 37, No. 1, pp. 534 – 545, 2010.
[26] Hina Anwar, Usman Qamar and Abdul Wahab Muzaffar Qureshi, "Global Optimization Ensemble Model for Classification Methods", Hindawi Publishing Corporation, The Scientific World Journal, Pages 1–9, 2014.
[27] Asma Gul, Aris Perperoglou, Zardad Khan, Osama Mahmoud Miftahuddin Miftahuddin, Werner Adler and Berthold Lausen, "Ensemble of a subset of KNN classifiers", Advances in Data Analysis and Classification, Pages 1–14, Springer.
[28] Alireza Osareh and Bita Shadgar, "An Efficient Ensemble Learning Method for Gene Microarray Classification", Hindawi Publishing Corporation, International Journal of Bio Medical Research, Volume 2013, pp. 1 – 10, July 2013.
[29] Yongjun Piao, Minghao Piao, Cheng Hao Jin, Ho Sun Shon, Ji-Moon Chung, Buhyun Hwang, and Keun Ho Ryu, "A New Ensemble Method with Feature Space Partitioning for High-Dimensional Data Classification", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Vol. 2015, pp. 1 – 12, January 2015.
[30] Thomas Dietterich.G , "Ensemble Methods in Machine Learning", Oregon State University, Corvallis, USA.
[31] Vaishali Ganganwar, "An Overview of classification algorithms for imbalanced datasets", International Journal of Emerging Technology and Advanced Engineering, Vol. 2, No. 4, April 2012.
[32] Chawla.N.V, et al."SMOTE: Synthetic Minority Over-sampling Technique", Research Journal of Artificial Intelligence, Vol.16, pp. 321–357, 2002 .
[33] Yu, L, "A DBN-based re-sampling SVM ensemble learning paradigm for credit classification with imbalanced data", Journal of Applied Soft Computing, Vol. 69, pp. 192 – 202, 2018.

[34] Borsos, Zalán, Camelia Lemnaru, and Rodica Potolea. ―Dealing with overlap and imbalance: A new metric and approach", Journal of Pattern Analysis and Applications, pp. 1 – 15.2016.
[35] Cao.P, Zhao.D, and Zaiane.O,"An optimized cost-sensitive SVM for imbalanced data learning,‖ In Lecture Notes in Computer Science including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics", Vol. 7819, pp. 280–292, 2013.
[36] Wang.H, Gao.Y, Shi.Y, and Wang.H, "A fast distributed classification algorithm for large-scale imbalanced data", Proceedings of IEEE International Conference on Data Mining, ICDM pp. 1251–256, 2017.
[37] Napierała.K, and Stefanowski.J, "Addressing imbalanced data with argument based rule learning", Journal of Expert Systems with Applications, Vol.42, No. 24, pp. 9468 – 9481, 2015.
[38] Madasamy.K and Ramaswamy.M, "Data Imbalance and Classifiers: Impact and Solutions from a Big Data Perspective", International Journal of Computational Intelligence Research (IJCIR), Vol.13, No. 9, pp.2267 – 2281, 2017.
[39] Alberto Fernandez, Sara del Rio, Nitesh V.Chawla, Francisco Herrera, "An Insight into imbalanced Big data classification : Outcomes and challenges", Journal of Complex Intelligent Systems, 2017, Springer.
[40] Akila.S, and Srinivasulu Reddy.U. "Risk based bagged ensemble (RBE) for credit card fraud detection", International Conference on Inventive Computing and Informatics (ICICI), 2017.
[41] Liu, Zhenbing, et al."Cost-Sensitive Collaborative Representation Based Classification via Probability Estimation Addressing the Class Imbalance Problem", Journal of Artificial Intelligence and Robotics, pp. 287 – 294, 2018, Springer.
[42] López.V., Fernández.A. and Herrera.F.,"On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed", Journal on Information Sciences, Vol. 257, pp.1 – 13, 2014.
[43] Maurya.C.K., Toshniwal.D and Venkoparao.G.V., "Online sparse class imbalance learning on big data", Journal of Neuro Computing, Vol. 216, pp.250 – 260, 2016.
[44] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (last accessed 12 March 2021)
[45] http://sci2s.ugr.es/keel/datasets.php (last accessed 12 March 2021)
[46] Halvaiee, Neda Soltani, and Mohammad Kazem Akbari ,"A novel model for credit card fraud detection using Artificial Immune Systems", Journal of Applied Soft Computing, Vol. 24, pp. 40 – 49, 2014.
[47] Polikar.R.,"Ensemble based systems in decision making", IEEE transactions on Circuits and systems magazine, Vol. 6, No. 3, pp. 21 – 45, 2006.
[48] Rokach.L.,"Ensemble-based classifiers", Journal of Artificial Intelligence Review, Vol. 33, No. 1-2, pp.1 – 39, 2010.
[49] Characteristics of Big data. Available online at https://www.edureka.co/blog/big-data-characteristics/ (last accessed 12 March 2021)
[50] Dorin Moldovan, Adrian Olosutean, Viorica Chifu, Cristina Pop, Tudor Cioara, Ionut Anghel, Ioan Salomie, "Big Data Analytics for the Daily Living Activitiesof the People with Dementia" Proceeding of the 14th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), pp.175–181, 2018.
[51] https://hadoop.apache.org (last accessed 13 May 2021)
[52] https://data-flair.training/blogs/hadoop-ecosystem-components (last accessed 13 May 2021)
[53] Awad M and Alabdallah A, "Addressing Imbalanced Classes Problem of Intrusion Detection System Using Weighted Extreme Learning Machine", International Journal of Computer Networks and Communications, Vol.11, No. 5, pp.39-58, September 2019.

AUTHORS

Dr.P. Ramasubramaniam received his Ph.D from Madurai Kammaraj University, Madurai, India in 2012. Currently he is working as Professor and Head in Department of Computer Science and Engineering, Shadan Women's College of Engineering and Technology, Hyderabad, India. His research areas are Data mining, Image Processing, and Text analysis. He has to his credit several papers in referred journals and conferences. He is also a member of several membership societies and has 30 years of teaching experience in various engineering colleges.

Dr.S. Hariharan received his B.E degree specialized in Computer Science and Engineering from Madurai Kammaraj University, Madurai, India in 2002, M.E degree specialized in the field of Computer Science and Engineering from Anna University, Chennai, India in 2004 and Ph.D degree in the area of Information Retrieval from Anna University, Chennai, India in the year 2010. He is a member of IAENG, IACSIT, ISTE, CSTA and has 17 years of experience in teaching. Currently he is working as Professor in Department of Computer Science and Engineering, Shadan Women's College of Engineering and Technology, Hyderabad, India. His research interests include Information Retrieval, Data mining, Opinion Mining, Web mining. He has to his credit several papers in referred journals and conferences. He also serves as editorial board member and as program committee member for several international journals and conferences.