ABSTRACT: The rapid growth in technologies and inexpensive internet connection has increased the volume of data generated. The data generated can be used to derive lots of information and patterns. Data sets are an essential part of the Machine Learning (ML) technique. But modern data sets are suffering from class imbalance. ML does not work very well with unbalanced data sets. In this context, this paper aims to provide a systematic literature review of unbalanced data sets for ML. The collected papers on class imbalance problem for ML were 4 major categories like binary class imbalance, multi-class imbalance, binary and multi-class imbalance, and rare events class imbalance. The survey focused on, various issues in class imbalance for ML. The purpose of the present paper is to help the scholars and readers in understanding the impact of the class imbalance for ML. This article contributes to the role of unbalanced data sets and their impact on the predictive systems.

KEYWORDS: Big data, Unbalance data sets, Class imbalance problem, Machine Learning (ML),

INTRODUCTION

Class distribution is the process of categorization of the labelled object into equal parts. This group of labelled objects is generated using the same number of events as presents in the data sets or they may be generated based on similar features or characteristics of the events. The ratio of the labelled object is equal in most of the instances but it is not true for every case. For instance, if a count of any class (majority class) in the data sets is over count than other class (minority class); it means there is a problem of class imbalance [1, 2, 3]. In the environment of data mining, class imbalance problem is very common; but the effects of this problem are very harmful because detection of rare events, abnormal activities, as well as unusual patterns, is very difficult. One of the difficulties is misclassification of rare events which generate a heavy loss in performance and cost also [4, 5]. To minimize the impact of class imbalance, the primary step is to focus on learning of imbalanced data sets.

The focal point of the learning imbalanced data sets is to identify factors of imbalanced data sets such as minority class, rare events or ratio calculated between the difference of majority and minority classes. Some slight points related to data imbalance are, the class imbalance is explained regarding a particular dataset. i.e. if any training distribution follows basic distribution for classification and it is the assumption that there is no issue with this classification then the whole dataset is balanced. But if the situation is not normal, meaning the performance of classified data is not satisfactory then the situation is quite complicated. Here data can be balanced artificially to avoid the problem but this solution cannot destroy the problem [6, 7, 8]. Another point is the relative proportions of the events in the data sets. The proportion of the total number of
events present in the class must be absolute. For instance, the class imbalance problem for a dataset with 50,000 positive examples and 5,000,000 negative examples is quite different from a dataset with 50 positive examples and 5000 negative examples—even though the class proportions are identical. These two problems can be referred to as problems with relative rarity and absolute rarity [9].

Machine learning is the direction towards getting the solution about how to train the computer system automatically using data sets or previous experiences. The journey of machine learning has started from laboratory curiosity to a practical technology in widespread commercial use. Machine learning is used in artificial intelligence (AI) for developing software for computer vision, speech recognition and natural language processing [10].

To handle class imbalance using machine learning techniques is the most challenging task. As per the current scenario, it has been observed, that only unequal ratio of the objects in classes is not the challenging issue but the nature of the data is also challenging [1]. Many machine learning approaches have been developed in the past decade to cope with imbalanced data classification, most of which have been based on sample techniques, cost-sensitive learning, and ensemble methods. Though several published surveys related to imbalanced learning were focused on detailed techniques application literature is neglected. For researchers from management, biology or other domains, rather than sophisticated algorithms, the problems that can be solved using imbalanced learning techniques and the building of imbalanced learning systems with mature yet effective methods may be of more concern [2].

The research questions (RQs) for the study are as follows:

**RQ1:** What is the role of the unbalanced dataset for machine learning?

**RQ2:** What are the different categories and sub-categories of the unbalanced dataset?

To address above RQs, a systematic literature survey was carried out with keywords ‘big data’, ‘unbalanced data set’, ‘class imbalanced problem’, and ‘machine learning’.

TAK (Title, Abstract, and Keywords) principal was used to shortlist the collected papers. Final shortlisted papers were categorized into different categories for better understanding of the research area. The research objectives (ROs) for the paper are as follows:

**RO1:** To identify the issues of the unbalanced dataset for machine learning.

**RO2:** To classify unbalanced dataset for understanding themes in current research.

The remaining structure of the paper is as follows; Section 2 describes the literature review part. Conclusion and future scope of research is described in Section 3.

**2. LITERATURE REVIEW**

Phases of research methodology are given in Figure 1. Phase 1 focused on the relation between ML and data sets. Various issues of unbalanced data sets for ML are discussed in phase 2. Finally, phase 3 discussed various categories and subcategories of unbalanced data sets for ML.

![Figure 1: Research Methodology](image-url)
The study considers four significant areas as i) Binary Class Imbalance, ii) Multi-class Imbalance, iii) Binary and Multi-class Imbalance, and iv) Rare events detection in binary and multi-class imbalance, followed by research gap identification.

2.1 Binary Class Imbalance (BCI)

In binary class imbalance, majority and minority classes play the main role. If the result of the prediction system depends on the majority class then there are no issues and accuracy is also better. However, if the result of the prediction system depends on minority class then the predicted results are not accurate or we cannot rely on such a system. The article was focused on BCI issues such as minority, majority class and rare class found in real-life problems. Most standard classifiers learning algorithms such as back-propagation neural network, decision tree, Bayesian network are also confronted difficulties while experienced unbalanced data set [2]. One of the aspects of BCI was within a class (imbalance in a single class). The random re-sampled method was the solution of such type of class imbalance [11]. To handle binary class imbalance for the number of events present in the minority as well as majority class (Data Level approach), ‘Synthetic Minority over-sampling Technique’ was proposed. This technique was a combination of under-sampling and oversampling [12]. To address class imbalance, we have four basic approaches data level approach (external), Algorithmic level approaches (internal), Cost-sensitive learning approaches, and Ensemble-based approaches [4]. For proper analysis of imbalance of data set, pre-processing is required in which we have to focus on noise which is present in events. According to Seiffertetal et al. [13], if noise reduction was noticeable on sampling technique performance and it was not often positive. In this situation, Random Under-sampling, Naive Bayes (NB) as well as Support Vector Machine (SVM) are performed better at all noise levels.

One of the addressing methods of binary class imbalance is bagging. Using the bagging approach with under-sampling can perform better on the binary class imbalance. Apart from the Data Level approach, for predicting accurate results it should be better class imbalanced could be handle with features of data sets. Feature extraction performs better to accurate prediction. In the field medical diagnosis this approach is most beneficial; such as to detect breast cancer malignancy where authors were used, an ensemble classifier EUSBoost to handle class imbalance [9,14].In the world of Machine Learning (ML) binary class imbalance problem is a relative problem and it is based on the degree of class imbalance, the size of training data, the complexity of the data sets and the classifiers which are used for classification. As per behaviour of different methods of oversampling and under-sampling used for balancing data in ML, authors were proposed two methods Smote + Tomek and Smote + ENN which worked on few positive examples whereas for larger examples authors were suggested Random over-sampling. To get an accurate prediction for real-time data with the cost-effective approach, Lin (2018) was proposed an online algorithm ‘Make Balanced-Cost Bound’ (Mb-Cb) which works on unpredictable domains [15, 16, 17]

2.2 Multi-class Imbalances

Uneven distribution among multi-class classification is the Multiclass Imbalance problem. To address the multi-class imbalanced problems, authors were divided addressing techniques into two categories: model-oriented solutions and data-oriented techniques. The proposed technique is based on Mahalanobis distance and named as Mahalanobis Distance-based Over-sampling technique (MDO). MDO has reduced the risk of overlapping between the classes regions in multi-class imbalance [18]. To face a high false alarm rate, Gao et al. [19] were proposed two-stage adaptive weighted extreme learning machine (AWELM) method for eyeglass and watch wearable. In the first stage, WELM classifier detected the suspected fall and in stage two it was refining the result detected in the first stage. SMOTE has a drawback that it has been balanced data set artificially; to overcome this drawback, authors were proposed a weighted kernel-based SMOTE (WK-SMOTE) for nonlinear problems where oversampling is used in the feature space of support vector machine (SVM) classifier. Finally, the authors conclude that the combination of WK-SMOTE and hierarchical SVM classifier can balance the multi-class imbalance problem [20].
2.3 Binary and Multi-class Imbalance

Some of the solutions proposed by various authors were addressed for binary and multi-class imbalanced. Sun and his friends were discussed about the impact of class imbalance on various application domains and the nature of the class imbalance in those domains. Their discussion was concluded that there is very less research on multi-class imbalance problem [2]. Krawczyk [21] studied about class imbalance problem with various aspects like classification, mining data streams, clustering, regression, and big data analytics. The study concluded that every aspect needs different strategies and solutions as per situation; he was also worked on future research of the above aspects. One of the addressing modes towards the class imbalance is the ensemble. To improve the performance of the ensemble, the authors were proposed ordering-based ensemble pruning metrics. The proposed metrics were focused on best behaving bagging and boosting-based ensembles [22]. In the field of software defect detection, it is necessary to explore the basic characteristics of class imbalance and the effect of imbalance learning. To detect software defect, the authors evaluated various data sets with different classifiers and various input metrics. Authors have also used The Matthews correlation coefficient (MCC) as an unbiased performance measure [23].

2.4 Rare events detection in binary and multi-class imbalance

In the world of imbalance learning, one of the issues is the detection of rare events. Some data sets harmed society such as intrusion detection or cyber-crime etc. To focus on various problems and its solutions of rare events in imbalanced data sets, concerning data mining authors were surveyed various papers and discussed various problems and solutions with their mapping; like Absolute Rarity can be addressed by over-sampling, Relative Rarity has a solution like sampling or cost-sensitive learning. Haixiang et al. [5] studied various papers technically as well as practically and made conclusion that rare events and imbalanced learning are related to a wide range of research area; and every area was required different addressing techniques to detect rare events like for chemical and biomedical engineering areas, re-sampling based ensemble classifiers are suitable while in IT sector online learning was better as real-time data was handled for data imbalance [24].

The summary of literature survey is tabulated below.

**Table 1: Different studies on Unbalanced Data sets**

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Publication</th>
<th>Country</th>
<th>Forms of class imbalanced</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chen et. al. [27]</td>
<td>China</td>
<td>Binary class imbalanced</td>
<td>Binary class problem in highly imbalanced data sets.</td>
</tr>
<tr>
<td>2</td>
<td>Luo et. al. [25]</td>
<td>China</td>
<td>Binary class imbalanced</td>
<td>Anomaly detection in fraud detection, disease detection</td>
</tr>
<tr>
<td>3</td>
<td>Buda et. al. ([26]</td>
<td>USA</td>
<td>Step imbalance and Linear imbalance</td>
<td>Consequences of class imbalance on classification performance and impact of imbalance on classification performances.</td>
</tr>
<tr>
<td>4</td>
<td>Lin et al.[17]</td>
<td>India</td>
<td>Binary class imbalance</td>
<td>Class imbalance due to minority classes</td>
</tr>
<tr>
<td>5</td>
<td>Mathew et al.[20]</td>
<td>Singapore</td>
<td>Multi-class imbalance</td>
<td>Class distribution in imbalance data sets</td>
</tr>
<tr>
<td>6</td>
<td>Song[23]</td>
<td>China</td>
<td>Binary and multi-class imbalance</td>
<td>Effect of imbalance learning,</td>
</tr>
<tr>
<td>7</td>
<td>Haixiang et al.[5]</td>
<td>China</td>
<td>Rare events class imbalance</td>
<td>Detection of rare events in data mining and machine learning.</td>
</tr>
<tr>
<td>8</td>
<td>SUN et al [2]</td>
<td>Canada</td>
<td>Binary and multi-class imbalance</td>
<td>Data imbalanced due to binary and multi-class distribution</td>
</tr>
<tr>
<td>9</td>
<td>Batista et al [28]</td>
<td>Brazil</td>
<td>Binary class imbalance</td>
<td>Imbalanced data sets and its addressing techniques</td>
</tr>
<tr>
<td>10</td>
<td>Japkowicz [11]</td>
<td>Canada</td>
<td>Binary class imbalance</td>
<td>Class imbalance in data sets</td>
</tr>
</tbody>
</table>
Japkowicz et al. [15] | Canada | Binary class imbalance | Class imbalance problem—with three dimension complexity, 2. The size of the training set, 3. The level of imbalance between the two classes

Weiss [24] | US | Rare events class imbalance | Rare class, rare cases

López et al [4] | Spain | Binary class imbalance | Data intrinsic characteristics and their problems

Seiffertet al. [29] | USA | Binary class imbalance | Impact of class imbalance on software modules

Bartosz Krawczyk [21] | Poland | Binary and multi-class imbalance | Different aspects of imbalanced learning such as classification, clustering, regression, mining data streams and big data analytics, providing a thorough guide to emerging issues in these domains

Chawla et al. [12] | USA | Binary class imbalance | One of the solution of binary class problem

Abdi et al. [18] | Iran | Multi-class imbalance | Addressed multi-class imbalance

Azaria [30] | USA | Multi-class imbalance | Behavioral analysis of insider threat (BAIT) framework

Blaszczyński et al. [14] | Poland | Binary class imbalance | Minority class imbalanced

Krawczyk et al. [9] | Poland | Binary class imbalance | Binary class classification problem

Galar et al. [22] | Spain | Binary and multi-class imbalance | Imbalanced classes

Gao et al. [19] | China | Multi-class imbalance | Abnormal activity recognition

Year wise and country wise publication of these 22 papers is shown in Figure 1 and 2 respectively.

![Year wise no. of papers](image)

**Figure 1:** Year wise no. of papers

Figure 1 shows the year wise published papers on unbalanced data sets. From year 2016 and 2018 has shown the highest number of papers which shows the growing nature of this field.
2.5 Research Gaps

As per the literature review discussed in the above section, it is clear that class imbalance problem is very common and effects of class imbalance are very harmful specifically for healthcare and security domains. Following are the summarized findings which give new directions towards research.

- The focal point of the class imbalance is minority class. It is important to work on the structure and features of the minority class for detailed analysis.
- To study about imbalanced regression problem for detailed analysis of rare events.
- To examine-in detail nature of class imbalance from binary class to multi-class distribution.
- Introduce novel approaches for ensemble classifiers based on a combination of oversampling, bagging and data level solutions to produce efficient learners.
- Evaluate class imbalance of data streams for different combinations of class distributions.
- In terms of big data, object types and noise are the sources of class imbalance. Examine exact object and noise level to evaluate class imbalance, detailed analysis is required.

In recent studies, Lakhe et al. [31] used regression analysis using data mining, Sivasankaran et al. [32] reviewed AI for industrial applications, and Narwane et al. [33] proposed smart manufacturing case study. However, these studies didn’t address the machine learning aspect. Thus this study bridges this gap.

3. CONCLUSION AND FUTURE SCOPE

In the world of digitization, the focus of attention is data sets. Most of the Machine Learning systems are dependent on quality data sets. Machine Learning algorithms are efficient and also perform well if data sets are balanced. But practically the situation is exactly opposite which results, the performance of the systems is very poor specifically healthcare systems. To improve the performance of the system, data sets need to be balanced. The paper discussed various issues of binary class imbalanced in the data sets and also focused research gap of the same for Machine Learning. To tackle this issue the better solution is that one should focus on detailed structure and features on the data sets rather than the number of events of the class which are present in the data set.

Limitations of the study are as follows: The papers were collected from selected peer-reviewed journals. The papers are from English journals and which are focused on unbalanced data sets and their impact on ML. However, this study provides crucial insight into the use of unbalanced datasets for ML. The future study of the class imbalanced focuses on multi-class imbalance and various novel approaches of classifiers for Machine Learning. The current study can be extended for digital forensic, software defect analysis, natural language processing, etc.

REFERENCES


Swati V. Narwane, Research Scholar, Department of Computer Engineering, Datta Meghe College of Engineering, Navi Mumbai, India, Email Id : svnarwane@gmail.com

Sudhir D. Sawarkar, Principal, Datta Meghe College of Engineering, Navi Mumbai, India Email Id: sudhir_sawarkar@yahoo.com