IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Volume-5, Issue-12 E-ISSN: 2347-2693

A Survey on knowledge extraction Approaches from Big Data and Rectifying Misclassification strategies

Jyoti Arora[1*], Ambica Sood[2]

[1*]Computer Science Engineering, Chitkara University, Kalu Jhanda, India
[2] Computer Science Engineering, Chandigarh University, Gharuan, India

[*]*Corresponding Author:* jyoti@chitkarauniversity.edu.in

Available online at: www.ijcseonline.org

Received: 30/Nov/2017, Revised: 06/Dec/2017, Accepted: 20/Dec/2017, Published: 31/Dec/2017

Abstract— The amount of data is increasing now a days due to usage of portable resources like smart phones, tablets and many more for accessing social sites. The requirement to analyze such big data to extract meaningful data came into existence. Traditional methods have been explored by number of researchers to analyze such data. These methods removed faulty data, uncertain data or misclassified data for better analyses. But this leads to loss of data. There is need to take into consideration the rectification of uncertainty in aspect of big datasets also. So, In this paper we survey big data, some traditional methods for data analyses, advance methods for data analyses, issues related to these methods, misclassification concept, the survey of rectification techniques for high accuracy followed by bearer future scope.

***Keywords*—**Big Data, Misclassification, Machine Learning, Knowledge, Discovery, Mining

I. INTRODUCTION

Now a days, Information Technology is keep on growing. Digital data is generating through devices, in addition to growing through world wide web with higher speed. The social sites as twitter, Facebook, WhatsApp and many more are the key sources of unstructured data production. The high volume produced data is known as big data. This kind of enormous data has excellent influence on modern culture as we are able to stream popular music, search directions on GPS, to send out photos, videos to our buddies and purchase new clothes at the same time. Big Data has various characteristics like volume, large number of sources, high speed, complex and different kinds of data. Based on these characteristics, discovering some correlation among this huge data is a challenging task. As creation of data through gadgets is much easier than extracting useful information. The process of examining number of records from relational databases and excerpting useful information is data mining. Mining of big data is not an easy task as recent time validates that firstly there is a need to collect data generated from various devices and then process, analyze and store this data for further use. The rise of generated data is because of surfing of social networking sites. These sites grant users to create different data in large amount and with high speed, So, mining is requires to extract useful from such unmanaged data. The first book for introduction of mining was published in 1998 by Indrukya and Weiss[1] that was based on extraction of meaningful information from data. Various activities like classification, estimation, prediction, visualization, pattern extraction and rules etc. can be performed on datasets to obtain a valuable information that can be used as useful knowledge. According to [3]traditional machine learning methods like supervised, unsupervised and reinforcement learning approaches are good enough for small sized datasets. But it is not applicable for big data because of data size, its generation speed and different variety.

These traditional learning methods are based on type of data used and type of output requires. But these traditional learning methods are not valuable for huge data analysis[19]. Weka[2] is a one of the popular and easy to use for knowledge analysis. There are some decision making theories given by Molotov [4] for mining like soft set theory and bijective soft set theory etc. These theories are applicable for small datasets. Most of the researchers have worked on these approaches and theories. Extracting knowledge using these theories from huge soft sets is still a challenge for researchers. With the Advanced machine learning approaches like active learning, parallel learning came into existence. Hadoop, Map reduce, canva, Jupyter and many more are existing big data analysis based on new machine learning methods like deep learning, active learning and kernel learning etc. These learning methods are tendering these days for massive data analysis. But before choosing learning methods there is problem of missing values, difficult to categorize data, noisy data and mislabeling in data generation, data acquisition, data storage and data analysis. Researchers have worked on these difficulties using traditional methods and leads to loss of data. In huge datasets also, wrong categorization may lead to loss of important information. This survey paper mainly focus on this wrong classification concepts and their rectification techniques.

Misclassification is defined as wrong classification. When two instances of same class is classified in two different classes that leads to misclassification. and it effects prediction accuracy. For that, a complete remonstration from the aspect of extraction of information and misclassification is still required. In this paper, brief review on the extraction of information based on misclassified data to have a basic idea to use data mining algorithms was discussed. Further techniques to extract uncertain data and to tolerate that misclassification big data perspective were discussed.

Figure 1 shows the outline of this paper, and its different subsections organized as following. Section I give a brief introduction to the big data, creation, significance and need to process big data. Section II explains Machine Learning that deals with discussion of some traditional mining algorithms and recent big data mining methods and tools. Section III deals with misclassification concepts, uncertain data and data that is hard to classify and some ways to compute misclassification. In section IV Rectification methods are discussed. Finally the paper is concluded in section V along with some future challenges in context of misclassification in big data soft set environment.

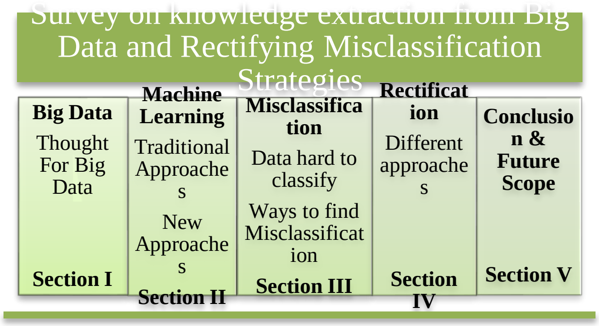


Figure 1. Outline of Paper.

A. Big Data

Currently buzz is around large data. In data storm period, data is generated from distinct sources and collected together within the database. The priority is not to expel the data rather hoard it for later use. This results in the formation of big data. Different researchers and organizations have their own views about big data. IDC characterizes Big Data advancements as another era of advancements and models, intended to monetarily separate an incentive from extensive volumes of a wide assortment of information by empowering high-speed catch, and additionally evaluation. There are three fundamental attributes of Big Data: the information itself, the investigation of the information, and the introduction of the consequences of the consideration [8]. According to Fisher et al. In this era, the data which is too huge and not handled by one machine and which is unable to process and manage using present data systems is termed as Big data[9]. IDC[10]indicates that the trading of big data is about $16100 million in 2014. According to [11] it will rise up to 32400 million and $114000 million by 2017 and 2018, respectively. In 2001 according to Meta Group Laney, big data has three dimensions like amount of data, speed of data and kind of data i.e.3Vs[13] But in 2012 with these characteristics high, huge or big keyword is added and termed as big data. Then for better explanation of big data more Vs are given i.e. Veracity, Variability and Value[12]. Most of the researchers explained big data is related with 5Vs Volume, Velocity, Variety, Value and Veracity[7] as shown in fig2. Gartner firstly explained 3V's of Big Data in [16] like Volume refers to the amount of data that is being handled and utilized in order to get the desired results. Velocity is all about the data travels from one point to another due to high requests that end users have for streamed data over numerous devices. Variety represents different kind of data that is stored, investigated and utilized. Further, Value in [17] is all about the quality of data that is stored and the further use of it. Veracity[17] [13] deals with consistency of big data. These 5Vs give rise to huge complex data. Microsoft also defined complex huge data as: “Now a day's huge data term is used that requires high computational power to process this, as requires in machine learning and for artificial intelligence for complex data."[14][138]

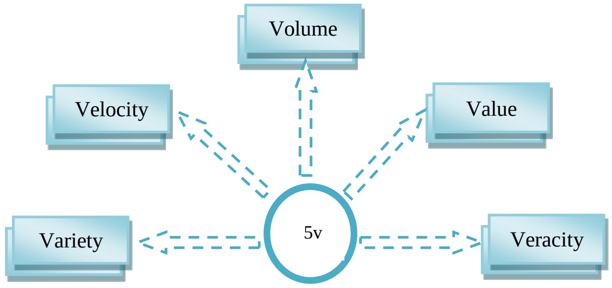
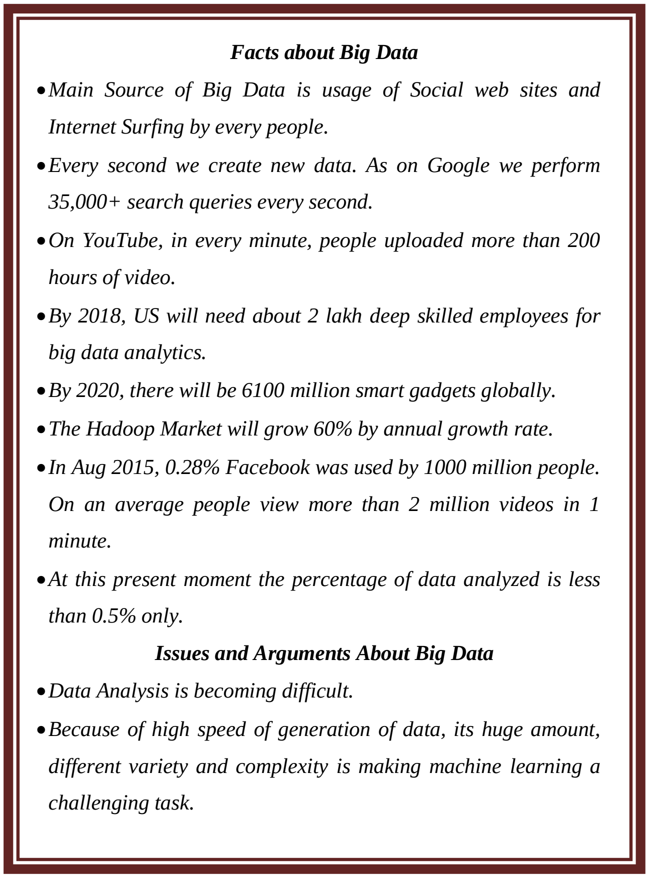


Figure 2. 5Vs of Big Data.

B. Thought For Big Data

This section discusses the question: why there is need to think about big data. There is no doubt that we all are living in the information deluge time, proved by the trend that tremendous quantity of information have constantly generated at unrivalled scales. Huge soft sets are gathered and analyzed in various areas like research, business, security and many more[20].Digital information produced from different electronic devices are expanding day by day. By 2011, The electric information has exploded near about nine time in amount within five years[21] and this will increase to approximately more than 40 trillion GB with in 2021[22]. So, there is a need to think about the word 'Big Data'.

In fig 3. Facts and arguments about big data in accordance with [5][6][18] have been discussed. These days if one will say data, that means big data actually and it will increase day by day.



Facts about Big Data

- *Main Source of Big Data is usage of Social web sites and Internet Surfing by every people.*
- *Every second we create new data. As on Google we perform 35,000+ search queries every second.*
- *On YouTube, in every minute, people uploaded more than 200 hours of video.*
- *By 2018, US will need about 2 lakh deep skilled employees for big data analytics.*
- *By 2020, there will be 6100 million smart gadgets globally.*
- *The Hadoop Market will grow 60% by annual growth rate.*
- *In Aug 2015, 0.28% Facebook was used by 1000 million people. On an average people view more than 2 million videos in 1 minute.*
- *At this present moment the percentage of data analyzed is less than 0.5% only.*

Issues and Arguments About Big Data

- *Data Analysis is becoming difficult.*
- *Because of high speed of generation of data, its huge amount, different variety and complexity is making machine learning a challenging task.*

Figure 3. Facts and Issues about Big Data.

II. Machine Learning

It is obvious that at this present moment massive information is considered, in all the fields of research. Though most of this enormous information is unquestionably important and producing valuable information involves new approach like machine learning. In last few years, machine learning is opted to solve the different challenges like analysis, mining, discovery and exploration in the field of medical, biotechnology, engineering and many more.

Machine Learning is an area of study which technically concentrates on thoughts, efficiency and qualities of examining components. It is clear from the above arguments of big data, that it is an challenging task. By taking some earlier experiences, this learning endeavours us to express the way to quickly look for the best predicator. It was related to the statistics subject but it became a different part of study during 1991's [33]. As this topic has importance in every field of technology for example, information technology, computer science, AI(Artificial Intelligence), research, robotics, biotechnology and many more [23][24][25][26]. Due to its inclusion in all fields, it has great influence on modern world[27]. Mostly, in these days, machine learning is famous for solving challenges like information exploration of huge data, artificial intelligence, identification systems and many more[28].

A. *Traditional Approaches of Machine Learning*

Traditionally, machine learning was mainly categorized into three subparts, according to [29] these are as follows:(i) Supervised Machine Learning Approaches (ii) Unsupervised Machine Learning Approaches and (iii) Reinforcement Machine Learning Approaches.

Supervised ML Approaches: This type of learning is taken as important it is suitable for space time trade off. Here, we have input variable (p) and output variable (q) and we use a classifier to mapping from input data to output data.

$$q = f(p).$$

The main aim of this learning to decide mapping function and with that we can predict output values using input values. Different approaches are used for training the data and testing the data. Like Random forest, Naive bayes, Decision Trees and many more.

Unsupervised ML Approaches: As opposed to supervised ML approaches, in this learning we have input variable (p) for that we don't have output variable. That's why known as unsupervised ML approaches. The main aim of unsupervised ML is to model the structure of data to get more detail of data. Various examples are clustering of objects and neural network are examples of unsupervised ML.

Reinforcement ML Approaches: This learning is efficient for decision making like in artificial intelligence, robotics etc. For an example: air conditioner will work on which temperature in summer season is an example of reinforcement learning. Here, a correct set of input and output is not always introduced. This learning is different from both supervised and unsupervised ML approaches.

Table 1. Knowledge of Supervised, Unsupervised and Reinforcement Machine Learning approaches

Parameters	**Machine Learning Approaches**		
	Supervised [40]	***Unsupervised [62]***	***Reinforcement [33]***
Definition	The learning approach, which takes labelled input for training and provides the desired output. Here, output is predefined.	It is a learning approach in which there is no need of labelled inputs for training. It takes inputs from the environment. Here, output is not predefined.	The learning approach which works on feedback taken from the environment is known as Reinforcement learning approach.
Basic Principle	Task Driven	Data Driven	Algorithm Driven (react to the environment)
Functions Performed	Classification/ Estimation/	Clustering	Decision making

Parameters	**Machine Learning Approaches**		
	Supervised [40]	***Unsupervised [62]***	***Reinforcement [33]***
Definition	The learning approach, which takes labelled input for training and provides the desired output. Here, output is predefined.	It is a learning approach in which there is no need of labelled inputs for training. It takes inputs from the environment. Here, output is not predefined.	The learning approach which works on feedback taken from the environment is known as Reinforcement learning approach.
	Regression		
When Touse	When one wants to classify or sort the data.	When no information about method of classification is provided.	When there is no idea to classify but if correctly classified then appreciable.
Basic Building	Algorithm based on input and output both.	Algorithm based on input.	Algorithm based on state dependent
Requirment	Actual or synthetic data	Actual or synthetic data	Real data
Examples	Naive Bayes, SVM, Linear regression, Random Forest and many more	K-means and X-means etc.	Q-learning and TD-learning etc.

According to these three categories, for handling datasets there are different classifiers and solutions are given in [30][31][32].As Google [30] was also using these learning approaches, for the data collected from voice recognition, search engines, Google maps, translators, web and image search engine etc. At last it can be considered that, for pre processing, data is generally analyzed using supervised and unsupervised learning approaches and for decision making reinforcement learning is applied. Table2 describes some features of techniques and classifiers opted for these three machine learning approaches:

Table 2. Features of Machine Learning approaches

Machine Learning Approaches	**Classification Techniques**	***Classification***	***Estimation***	***Regression***	***Clustering***	***Decision Making***	***Data Analysis***	***Prediction***
Supervised Machine Learning	NB [23]	✓	✓	✓	x	✓	✓	✓
	SVM [34] [35]	✓	✓	✓	x	✓	✓	✓
	NN [33] [36]	✓	✓	✓	x	✓	✓	✓
Unsupervised Learning	K-means [37]	✓	x	x	✓	✓	✓	✓
	X-	✓			✓	✓	✓	✓
Supervised Machine Learning	NB [23]	✓	✓	✓	x	✓	✓	✓
	means [37]		x	x				
Reinforcement Learning	Q-learning [33] [38]	x	x	✓	x	✓	x	✓
	TD-learning [33] [39]	x	x	✓	x	✓	x	✓

In addition to these learning approaches and their features, we can say these traditional approaches are intended using suppositions like the gathered information will be stored in memory for analyzing. But the data is growing rapidly and these approaches are facing complications for handling this high volume data. For this, we require modern strategies to handle large amount of data. The advanced approaches for data analysis are discussed in next part.

B. *Modern Approaches of Machine Learning*

While noticing the requirement of advance approaches of machine learning, in this section we discusses new approaches that are offering or we can say required learning approaches for handling huge datasets. The main feature of these new learning approaches is to concentrate on the way of training and testing of data. There are following modern learning methods for handling and processing big data as shown in fig4.

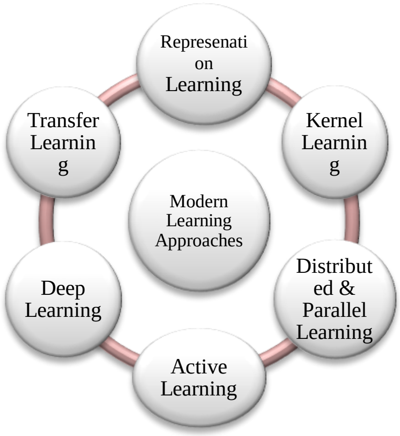


Figure 4. Approaches of Machine Learning

.

The modern learning has been worked on some new principles. Based on these modern learning there are some tools that are considered as learning tools for big data.

Table3 shows the tools used for learning. This part of the paper describes the detail of these approaches:

Representation learning: For high dimensionality data, representation learning is one of the solution to extract the useful information from unstructured data. This learning method gives high accuracy and performance as compared to classical machine learning methods. This learning is mainly designed to accomplish the enhancement in performance based on computational and mathematical calculations. As it takes possible input designs and gives a fairly sized mastered manifestation.[63][64]. Due to its popularity in high dimensionality mining, it has been enhanced in different fields these years as in the field of automatic biased and many more [64][65][66][67]. Representation learning provides real life applications like NLP(natural language processing), modern vehicles and many more discussed in [68][69][70]. The detailed information of this learning approach is discussed in [107].The plugin based on representation learning is: MatDeepRep[130]

a) MatDeepRep : This is a plugin used in MATLAB for representation learning. It is used for the classification of images based on low level, middle level and high level features. Most of researchers used MatDeepRep for recognition of large amount of images. Now a days, usage of social networking sites is increasing, there uploading pictures, downloading pictures give rise to the importance of MatDeepRep.

Kernel Learning: Few years back, kernel learning was considered itself as extremely effective. Because kernel based learning has the ways to enhance the performance and efficiency based on computational using some non linear approaches. For kernel based learning, the kernel function is used for computation. As this method maps original inputs to high dimension outputs. To enhance the performance and accuracy, there is only need to change the kernel methods but the way to choose kernel function is a future challenge for researchers[98] [99]. Real life applications of kernel learning are parameter estimation and many more are discussed in[100][101][102][103] [104]. And the explanatory detail of kernel learning has been discussed in [105]. Kernel learning is based on support vector machine, logistic regression but kernel functions make the difference. In short, a learning where kernel of operating system is used is known as kernel learning. The frameworks based on kernel learning are Spider[131], Kernel PCA[132].

a) Spider: Spider plugin is used in MATLAB for kernel learning for large amount and large sized images. In this learning Basic concepts of svm are used for training the data and testing the image data.

b) Kernel PCA: Kernel learning used as principal component analysis, for feature extraction and feature selection is popular in Kernel PCA framework.

Distributed and parallel learning: Due to large amount of data, extracting useful information is a challenging task. To understand the complex data, simple way is to distribute and parallelize it. In this learning method, when data is allocated to different data centres then learning is taken as distributed learning method similar in parallel learning, learning data is trained in parallel manner. [84]These learning methods avoids to take whole amount of data as single because of distributed and parallel manner as compared to another learning methods.[79][80]. Real life applications of distributed and parallel learning are meta learning and many more can be seen in [81][82][83][85].Frameworks or tools based on distributed and parallel learning are: Map reduce[15] and Apache hadoop[119], Apache spark[118], Strom[128], pig [129]etc. are all based on Distributed and Parallel learning.

a) Map reduce: This framework is the most famous and mostly used for big data processing. Mapreduce is based on two major components known as mapper and reducer. The main function of this framework is to produce a key-value pair for the inputs by mapper and then it will be combined by using reduce function. Now a days, Mapreduce is the most used framework.

b) Hadoop: This framework is not only used for processing but also used for the storage purpose of big data. Processing of data is done by map reducer and storage is in Hadoop distributed file system. This framework works well in case of any failure, because it has three replicas in built. Here storage is in distributed manner and mapping is performed in parallel manner. The main advantage of this framework is that this helps to store the large data in less space.

c) Spark: This framework is used for various purposes like for stream data processing, graph data processing, structure data processing and other type of data. Here, machine learning libraries like Mlib and others according to requirement are need to be install for getting machine learning features. Spark has a great features of abstraction because of presence of resilient distributed datasets(RDD) in it. These datasets have the feature of read mode only. Spark is based on cluster environment, for this, it is dependent on YARN(yet another resource navigator).

d) Apache Strom: This framework also works on distributed computation for large datasets. It is based on boss-employee architecture. The cluster has some boss nodes and employee nodes. These nodes works together in a distributed and parallel manner.

e) Pig: Apache pig is one of the engine that works on Apache Hadoop. It has enhanced features like processing the data, reading the data and writing the data by using some operations of Pig Latin.

Active Learning: Active Learning is mostly used when there is large amount of data but that data is not labelled. This learning works better in the complicated situation and gives high accuracy as compared to another algorithms that deals with unlabeled data.[94][95]. Different methods of this learning are mentioned in [95][96][97]. Real life applications of Active learning are classification of images and many more in [96][97].The tool based on active learning is: Canva

a) Canva: Canva is one of the tool for active learning, it has different features like creating contents, papers, effective posts, blogs, books, documents and advertisement for sharing.

Deep Learning: These days, due to large amount of data, Deep Learning is the most thrust topic in machine learning approaches. This learning is the enhancement of traditional learning approaches as it's a combination of supervised and unsupervised learning approaches. Whenever only one of them is used in deep architecture then also it is considered as deep learning for tree like representation learning.[71][72][73]. Various types of deep learning has been discussed in [74][75]. Real life applications of deep learning is advancement of graphic processors and many more in [76][77][78]. Frameworks or tools based on deep learning are: Apache Singa[114], Torch[117], Caffe[115], and Purine[116] etc.

a) Apache Singa : This project is licensed by Apache, is providing a deep learning environment. It is based on easy programming and gives a flexible and scalable architecture for huge amount of data. This can handle more number of attributes and here, stochastic gradient algorithm is generally used for training the data. Training the huge data is based totally on high programming like python, java etc.

b) Torch: Torch project is also based on deep learning model. This framework works on just in time complier known as JIT(just in time) compiler and based on Lua (like javascript). Torch project is based on routines of linear algebra and numeric. It supports the GPU (graphics processing unit) efficiently. Here, one dimensional array considered as tensor and two dimensional array considered as matrices are used for performing deep learning in huge data.

c) Caffe: Caffe is convolution architecture for feature extraction is one of the deep learning framework tool. This tool works faster but sometimes slower in case of huge data. This tool is used for research work in the field of big data learning. Here, n- dimensional array is used for training data. This framework is based on convolution neural network (CNN). It works with GPU(graphics processing unit) compiler.

d) Purine: The framework based on bigraph mechanism is known as purine framework. This framework works parallel in deep learning and make the coding easier for researchers. The scheme followed in purine is the combination of GPU and central processing unit.

Transfer Learning: Whenever training data and testing data lies in same feature space then transfer learning is used. On the other hand, with the increase of data, variety of data has also been increased and it damages hypothesis. For solving this complication, transfer learning came into context for extracting knowledge from large amount of data. The main feature of this learning is that it works faster as compared to another learning algorithms for training and testing. Transfer learning is explained in [106] precisely. Different ways to apply transfer learning has been discussed in[86][87]. This learning is based on principle "what data we have to transfer". Various approaches of transfer learning are instance based learning, parameter based learning and many more are discussed in [88][89][90]. Real life applications of transfer learning is classification of text and others are mentioned in[91][92][93].The application based on transfer learning is: Google Drive.

a) Google Drive: This application is the mostly used by modern society. It provides free storage on cloud for sharing and storage of data. This application is accessible anywhere from all the electronic devices like smart phones, pc's etc.

From the above discussion we can conclude that these learning are better than traditional learning approaches as these learning approaches handle huge amount of data which is generated with high speed in efficient manner. These learning methods will provide high accuracy as compared to classical learning methods. The basic functionality of each learning and the tools that can be used for particular learning has named in Table 3 as follows and is described above.

Table 3. Describes the basic functionality of modern approach of learning

Machine Learning Approaches For Big Data	**References**	**Basic Functions**	**Tools Used**
Deep Learning	[71-78]	Supervised and/or Unsupervised Learning Approaches	Apache Singa, Torch, Purine and Caffe etc.
Distributed and Parallel Learning	[79-85]	Distributed and Parallel Manner	Apache Hadoop, Mapreduce,Apache Pig and Storm etc.
Transfer Learning	[86-93] [106]	Instance Based, Feature Based and Parameter Based	Google Drive etc.
Active Learning	[94-97]	Query Strategy on Unlabeled Data	Canva etc.
Kernel Learning	[98-103]	Kernel Methods and Kernel Functions	Kernel PCA, Spider
Representation Learning	[63-70]	Feature Selection, Feature Extraction and distance metric Learning	MatDeepRep

Problem With these modern approaches and tools:
At this point, majority of the study in the field of big data and machine learning has aimed at handling huge data and its algorithms inclusion etc. Most of the approaches and tools have brushed aside the enhancement of pre processing approaches. Traditional learning approaches have some problems for analyzing data even researchers are facing problems with advanced approaches also. As these approaches haven't proposed on inefficient data. There is a need to remonstrate these various kind of inefficient data for getting high accuracy.

Mislabelled Data: With the growth of data, mislabelled data is also growing. In case of millions of records, it is difficult to classify unlabeled data. Whenever machine learning algorithm will be used for training using this type of data will results low accuracy. So there should be a solution to label this data before the classification process takes place [108]. For an example: Label of one image is that image contains two people. But without this label it will be difficult to classify that in which category it is to be classify.

Noisy Data: The disturbed data or incorrect data is termed as noisy data. As data has been gathered from devices but that data is not the correct record. This type of inefficiency can be resolved by clustering etc.[110] For an example: Image captured or collected is in blurred form.
Missing Values: Missing Value means that, the data collected doesn't have any value. While collecting data, value has been missed. This data also leads to less accuracy similar to another inefficient data types. This problem can be solved easily by using some mean values or by using some clustering and filtering processes. For this inefficient data, researchers have given the solution using some computation methods [109].

Imbalance Data: This is defined as imbalance data in the classes as there are more number of records in one class as compared to other. These uncertainty also leads to less accuracy. This problem can be solved by using sampling ways. [111]

Hard to classify data: This describes that some of the records in data sets are difficult to categorize that type of records are termed as hard to classify. Whenever there is no relationship between instances then it lead to a problem. So there is a need to think why these records are difficult to classify and how it leads to misclassification[127]. The solution for this inefficient data is boundary value analysis.
Because of above problems, most of time, data is wrong classified. Researchers have used various tools for coping up with these types of problems. But still it's an open challenge[112][113]. For that reason misclassification concept is to be remonstrated.

III. MISCLASSIFICATION

Misclassification is the situation of data errors. Massive data packages through web methods are often untrustworthy, susceptible to outages in addition to losses, this interruptions usually are increasing in case of utilization of multiple soft sets. Social researchers wanted the answer of the questions related to the collection of data and trying to account for any misclassification inside their data. The data set taken may perhaps have some of scores of fecal material data, however for many people it is actually arbitrary. To generate statistical results about a data set, we all have to know the particular weak points in the data. Additionally, scientists need to have the ability to account for the particular misclassification inside the datasets. To achieve this necessitates each one has their own perspective about misclassification. As Healy et. al in [126] describes that Whenever we have number of items, that can be divided into some different categories. But in which the item is classified is not a right category for that item, then that item will create some misclassification. That time, it is required to find the right value or right category of the item.

Misclassification concept requires to be consider while extracting valuable information from the data. Here, there are several theories those have taken into consider for knowledge extraction but haven't computed misclassification. Table4 gives the name of that theories and their faulty data consideration and rectification feature.

Table 4. Techniques those have considered faulty data

Extracting knowledge Techniques	**Faulty Data Considered**	**Misclassification Computed**	**Rectified Misclassification**
Fuzzy theory[48][50]	✔	×	×
Rough set theory[46]	✔	×	×
Cloud theory[42][52][55]	✔	×	×
Frequent item set[41][43][47][49]	✔	×	×
Clustering[44][45][51][53][54]	✔	×	×

From the above table, we can see, there is need to survey about misclassification concept and computation of misclassification. There are various ways to compute misclassification, are given as follows:

A. Ways to Compute Misclassification

Computation of misclassification is a statistical and mathematical method. While classifying data into categories, the data which is not classified, or sometimes classified to wrong class is coined to misclassification. For an example, There are two classes: one class named as Animal Class and another class named as Birds Class.

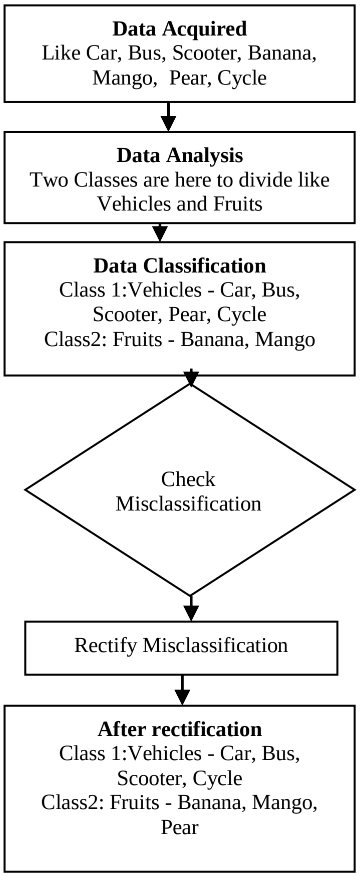

Data Acquired
Like Car, Bus, Scooter, Banana, Mango, Pear, Cycle
Data Analysis
Two Classes are here to divide like Vehicles and Fruits
Data Classification
Class 1:Vehicles - Car, Bus, Scooter, Pear, Cycle
Class2: Fruits - Banana, Mango
Check Misclassification
Rectify Misclassification
After rectification
Class 1:Vehicles - Car, Bus, Scooter, Cycle
Class2: Fruits - Banana, Mango, Pear

Figure 5. Flow Chart For Rectification of Misclassification

Problem is that to classify the data into their particular class. Following are the various ways:

I) Whenever numeric/ string data is classified based on confusion matrix. The confusion matrix defined using four terms: **True Positive:** This term describes the records that follows that condition correctly in the presence of condition. **False Positive:** It describes the records, that follows the condition in the absence of the given condition. **False Negative:** describes the records, that doesn't follows the particular condition in the presence of that particular condition. **True Negative:** It describes the records, that doesn't follow the condition in the absence of condition. These four terms are represented in matrix in figure5. below named as confusion matrix.

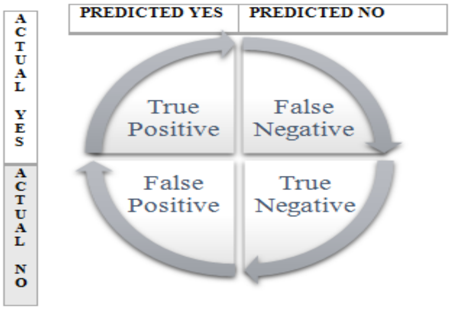


Figure 6. Confusion Matrix

Misclassification can be computed using following equation. Most of the mining tools follows this basic concept for computation.

Misclassification = $\frac{FP+FN}{TP+TN+FP+FN}$
Misclassification = $100 - Accuracy$
Where Accuracy =$\frac{TP+TN}{TP+TN+FP+FN}$

II)Whenever we have dataset divided into ranges or in sets form then misclassification computation is as follows using set theory:
As we have U set of elements and two classes named as X and Y. By mistake, the elements of X set classified into Y set and X and Y both are the subsets of universal set U then misclassification degree is computed as follows:

Misclassification = $1 - \frac{|X \cap Y|}{|X|}$
For an example:

U={x_1,x_2,x_3,x_4,x_5,x_6,x_7} are set of houses.
(F,E) is soft set over U, E=e_1,e_2,e_3,e_4 are the parameters of the dataset. These are very cheap, cheap, costly and very costly respectively.
Let mapping of these parameters are shown in table5:
F(e_1)=F(very cheap) = {x_1,x_2,x_3}
F(e_2)=(cheap) = {x_4,x_5,x_6}
F(e_3)=F(costly) = {x_7}
F(e_4)=F(very costly) = {x_4,x_5,x_6,x_7}
F(e_1,e_2,e_3) and F(e_1,e_4) is BSS(bijective soft set) as it follows the properties of bss theory given in [56]
where F(e_1,e_2) F(e_1,e_3) F(e_2,e_4) are not BSS.

Table 5. Mapping of each parameter of dataset

H vs P	e1(very cheap)	e2(cheap)	e3(costly)	e4(Very costly)
x1	1	0	0	0
x2	1	0	0	0
x3	1	0	0	0
x4	0	1	0	1
x5	0	1	0	1
x6	0	1	0	1
x7	0	0	1	1

As set is not a bijective soft set. so elements are misclassified the computation of misclassification for parameter e_2 and e_4.
For $F(e_2,e_4)$
As $C(e_2,e_4) = 1-|e_2 \cap e_4|/ |e_4|$
$1-3/4=1/4 =25\%$.
As 25% elements are misclassified here. Here are total 7 elements and approx 1 or 2 elements are misclassified. Some researchers have followed another way out for computation, that are named in the following table.

Table.6 gives the detail of articles, those have followed some method for computation of misclassification rate or error. These have given the detailed percentage of misclassification. but not any way out of rectification of misclassification based on huge datasets.

Table 6. Survey of Misclassification concepts and methods in soft sets

Ref.no. Year	Description	Faulty Data Considered	Method for Computing misclassification	%age of Misclassification	Rectification Method Followed
[126] 1981	Discussed about the effect of misclassification error.	Y	Changing the design of the classifier	NM	N
[121] 1999	Worked on mislabeled data	Y	Supervised Learning	40%	N
[123] 2004	Epidemiology dataset has been considered	Y	Log Linear Model	NM	N
[120] 2005	Classification of dataset has been performed	Y	Logit Model	70%	N
[125] 2005	Focused on Improving the classification Accuracy	Y	Changing Cost Ratio	NM	N
[124] 2009	In this labeling of samples has been performed	Y	Reflect And Correct Method	NM	N
[122] 2009	Bio-informative dataset has been considered	Y	Eliminated the misclassified elements	NM	N
[56] 2016	Shoreline Resources dataset has been considered	Y	Bijective Soft Set Theory	25%	N

Where NM stands for Not Mentioned, Y- Yes, N-No

IV. RECTIFICATION

Rectification means the correction. In this section we are discussing about rectification of misclassified data. As after repairing the misclassified data, we can get 100% accuracy and 0%. misclassification.

Need of rectification:

As elements that are not diagnosed properly resulted not classified properly, that leads to misclassification, and it should be rectified. Misclassification is an issue, which may be quite expensive because it is linked to fixable methods.

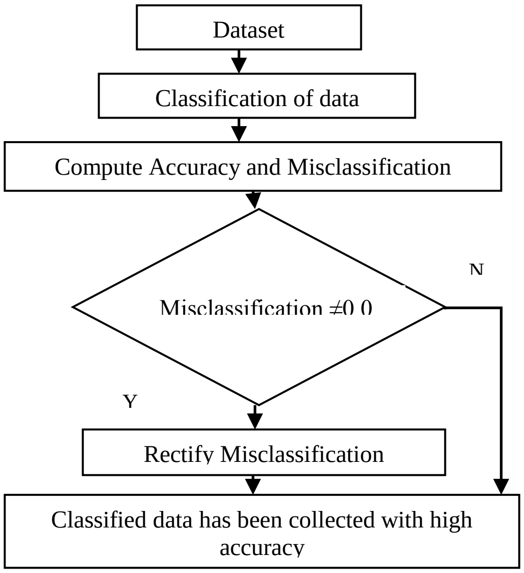

Dataset
Classification of data
Compute Accuracy and Misclassification
Misclassification ≠0.0
N
Y
Rectify Misclassification
Classified data has been collected with high accuracy

Fig.7

Rectification Methods: These days, misclassification is a research topic in the field of data mining in big data. As every problems has some solution. Various methods has been developed for rectifying problem. some of them are given below. But there are few methods that are rectifying misclassification based on parameters. In this section already existing methods has been discussed.
Some methods are: Naive methods, Regression Calibration, Pooled Estimation, Multiple imputation and Corrected Score Estimation. These methods are applicable on small sized datasets. The rectification of misclassification in case of large data is still a challenging task.

Regression Calibration: Regression Calibration termed as RC is really a conventional method to solve the problem occurred due to misclassification. Regression calibration is a easy and common approach.
This method is mostly used method for regression type problems. The main working of this method is that it replaces the variable Y with the regression function of Y that can be any variable F. This F variable is further a approximate value and helps in data analysis and this will provide a high accuracy. For computing and rectifying misclassification error in small dataset this method is popular method. It is used for handling discrete data and covariant data. It is not giving an accurate value but just giving an estimation of uncertain or infected value.
For computation Rosner et. al.[138] has given a formula, first it has described the relationship between the input correct value C and the output estimated near value E using some linear relationship:

$$E = \beta 0 + \beta 1C + \varepsilon$$

this equation is similar to y=mx + c where β0 is intercept and β1 is slope and ϵ is misclassified value. now this β1 is a value to find that show the proper relationship between E and C. But we don't have direct value of C so another formula

$$E = A0 + A1M + \epsilon$$

where A0 is equal to ˜ β0 and A1 is equals to ˜ β1 and ϵ is equals to ˜ ϵ. where M is wrong value. Now by relating C and M we can find out the difference between the β1 and ˜ β1. Hence M = C + R where R is random error.

Pooled Estimation: This pooled estimation method is based on some estimation, it combines the estimator of regression calibration and estimator of validated data. This method can be extended for getting high performance. But to handle large datasets is not too much appropriate for this method. The working principle is based on variance computation and in this validated estimator of true analysis will ignore the misclassified elements.

Multiple Imputation: This method is based on regression model known as logistic regression method. This method is for rectifying misclassification occurred using some missing values. This method's main working is to handle missing values as well as misclassification that's why known as multiple imputation. In this method missing values are filled repeatedly and at last combined the result. The detailed information about multiple imputation is in [134].

Corrected Score Estimation: This method is proposed by [135] and it is the enhanced method of [136] and [137]. Whenever there is no misclassification then this method defines the score function.

$$\text{score}(\gamma) = \frac{1}{n}\sum_{j=1}^{n} \delta\, j \left[Yj - \frac{\frac{1}{n}\sum_{k=1}^{n} Yj(t)\ exp(\alpha Yi)}{\frac{1}{n}\sum_{j=1}^{n} Yj(t)\ exp(\alpha Yi)} \right]$$

In this the values of Y is replaced by the observed wrong value. as observed wrong value is V and it is replaced with V = U f(V) where U is misclassification function matrix and f is also a function. It works better in case of less misclassification. this method is different from RC and MI.[133]

V. Conclusion and Future Scope

In this paper, survey of big data, their extracting knowledge techniques and misclassification computation has been done. As big data is in growing stage and by using various techniques, knowledge has been extracted. As machine learning is the topic that is adopted these days as a research topic for extracting information. But this task leads to some problem. In this paper, we discussed the knowledge of that repairing techniques in context of big data. In future these techniques can be enhanced to get better results in easy manner. In future, this concept can contribute in real life applications like analysis the data of social networking sites like Facebook, Google drive etc. , analysis of health data and political data and many more.

References

[1] Weiss, Sholom M., and Nitin Indurkhya. *Predictive data mining: a practical guide*. Morgan Kaufmann, 1998.

[2] Basu, Sugato, and Prem Melville. "Weka Tutorial." *ht-tp://www. cs. utexas. edu/users/ml/tutorials/Weka-tut*.

[3] Fisher, Danyel, Rob DeLine, Mary Czerwinski, and Steven Drucker. "Interactions with big data analytics." *interactions* 19, no. 3 (2012): 50-59.

[4] Molodtsov, Dmitriy. "Soft set theory—first results." *Computers & Mathematics with Applications* 37, no. 4-5 (1999): 19-31.

[5] Marr, Bernard. "Big Data: 20 Mind-Boggling Facts Everyone Must Read." *Forbes Magazine* (2015).

[6] https://www.modeln.com/blog/high-tech/2016/10-interesting-facts-big-data/

[7] Tole, Alexandru Adrian. "Big data challenges." *Database Systems Journal* 4, no. 3 (2013): 31-40.

[8] Gantz, John, and David Reinsel. "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east." *IDC iView: IDC Analyze the future* 2007, no. 2012 (2012): 1-16.

[9] Fisher, Danyel, Rob DeLine, Mary Czerwinski, and Steven Drucker. "Interactions with big data analytics." *interactions* 19, no. 3 (2012): 50-59.

[10] Press, Gil. "$16.1 billion big data market: 2014 predictions from IDC and IIA." *Forbes. com* (2013).

[11] Big data and analytics—an IDC four pillar research area, IDC, Tech. Rep. 2013.

[12] Laney, Doug. "3D data management: Controlling data volume, velocity and variety." *META Group Research Note* 6 (2001): 70.

[13] Sagiroglu, Seref, and Duygu Sinanc. "Big data: A review." In *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, pp. 42-47. IEEE, 2013.

[14] The Big Bang: How the Big Data Explosion Is Changing the World - Microsoft UK Enterprise Insights Blog - Site Home - MSDN Blogs.

[15] Landset, Sara, Taghi M. Khoshgoftaar, Aaron N. Richter, and Tawfiq Hasanin. "A survey of open source tools for machine learning with big data in the Hadoop ecosystem." *Journal of Big Data* 2, no. 1 (2015): 24.

[16] Beyer, Mark A., and Douglas Laney. "The Importance of "Big Data": A Definition. Gartner." (2012).

[17] Hashem, Ibrahim Abaker Targio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. "The rise of "big data" on cloud computing: Review and open research issues." *Information Systems* 47 (2015): 98-115.

[18] Khan, Nawsher, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani. "Big data: survey, technologies, opportunities, and challenges." *The Scientific World Journal* 2014 (2014).

[19] Kaisler, Stephen, Frank Armour, J. Alberto Espinosa, and William Money. "Big data: Issues and challenges moving forward." In *System sciences (HICSS), 2013 46th Hawaii international conference on*, pp. 995-1004. IEEE, 2013.

[20] Sandryhaila, Aliaksei, and Jose MF Moura. "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure." *IEEE Signal Processing Magazine* 31, no. 5 (2014): 80-90.

[21] Gantz, J., and D. Reinsel. "Extracting value from chaos technical report white paper." *International Data Corporation (IDC) Sponsored by EMC Corporation* (2011).

[22] Gantz, John, and David Reinsel. "The digital universe decade-are you ready." *IDC White Paper* (2010): 1-16.

[23] Mitchell, Tom M. "Machine learning. WCB." (1997).

[24] Russell, Stuart, Peter Norvig, and Artificial Intelligence. "A modern approach." *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* 25 (1995): 27.

[25] Cherkassky, Vladimir, and Filip M. Mulier. *Learning from data: concepts, theory, and methods*. John Wiley & Sons, 2007.

[26] Mitchell, Tom Michael. *The discipline of machine learning*. Vol. 9. Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.

[27] Rudin, Cynthia, and Kiri L. Wagstaff. "Machine learning for science and society." *Machine Learning* 95, no. 1 (2014): 1-9.

[28] Bishop, Christopher M. "Pattern recognition." *Machine Learning* 128 (2006): 1-58.

[29] Adam, Bernard, and Ian F. Smith. "Reinforcement learning for structural control." *Journal of Computing in Civil Engineering* 22, no. 2 (2008): 133-139.

[30] Jones, Nicola. "The learning machines." *Nature* 505, no. 7482 (2014): 146.

[31] Langford, John. "Tutorial on practical prediction theory for classification." *Journal of machine learning research* 6, no. Mar (2005): 273-306.

[32] Bekkerman, Ron, Ran El-Yaniv, Naftali Tishby, and Yoad Winter. "Distributional word clusters vs. words for text categorization." *Journal of Machine Learning Research* 3, no. Mar (2003): 1183-1208.

[33] Burch, Carl. "A survey of machine learning." *Tech. report, Pennsylvania Governor's School for the Sciences* (2001).

[34] Zheng, Jun, Furao Shen, Hongjun Fan, and Jinxi Zhao. "An online incremental learning support vector machine for large-scale data." *Neural Computing and Applications* 22, no. 5 (2013): 1023-1035.

[35] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20, no. 3 (1995): 273-297.

[36] Dong, Xu, Ying Li, Chun Wu, and Yueming Cai. "A learner based on neural network for cognitive radio." In *Communication Technology (ICCT), 2010 12th IEEE International Conference on*, pp. 893-896. IEEE, 2010.

[37] Safatly, Lise, Mario Bkassiny, Mohammed Al-Husseini, and Ali El-Hajj. "Cognitive radio transceivers: RF, spectrum sensing, and learning algorithms review." *International Journal of Antennas and Propagation* 2014 (2014).

[38] Galindo-Serrano, Ana, and Lorenza Giupponi. "Distributed Q-learning for aggregated interference control in cognitive radio networks." *IEEE Transactions on Vehicular Technology* 59, no. 4 (2010): 1823-1834.

[39] Sutton, Richard S. "Learning to predict by the methods of temporal differences." *Machine learning* 3, no. 1 (1988): 9-44.

[40] O. Okun, G. Valentini, (Eds.), Supervised and Unsupervised Ensemble Methods and their Applications Studies in Computational Intelligence, vol. 126, Springer, Heidelberg, 2008.

[41] Abaei, Golnoush, Ali Selamat, and Hamido Fujita. "An empirical study based on semi-supervised hybrid self-organizing map for software fault prediction." *Knowledge-Based Systems* 74 (2015): 28-39.

[42] Abdi, Amir, Suvi Heinonen, Christopher Juhlin, and Tuomo Karinen. "Constraints on the geometry of the Suasselkä post-glacial fault, northern Finland, based on reflection seismic imaging." *Tectonophysics* 649 (2015): 130-138.

[43] Audet, Patrick, Bradley D. Pinno, and Evelyne Thiffault. "Reclamation of boreal forest after oil sands mining: anticipating novel challenges in novel environments." *Canadian Journal of Forest Research* 45, no. 3 (2014): 364-371.

[44] Bissig, Thomas, Alan H. Clark, Amelia Rainbow, and Allan Montgomery. "Physiographic and tectonic settings of high-sulfidation epithermal gold–silver deposits of the Andes and their controls on mineralizing processes." *Ore Geology Reviews* 65 (2015): 327-364.

[45] Botros, N. S. "The role of the granite emplacement and structural setting on the genesis of gold mineralization in Egypt." *Ore Geology Reviews* 70 (2015): 173-187.

[46] Karapetrou, S., M. Manakou, D. Bindi, B. Petrovic, and K. Pitilakis. "“Time-building specific” seismic vulnerability assessment of a hospital RC building using field monitoring data." *Engineering Structures* 112 (2016): 114-132.

[47] Khan, Salman H., M. Ali Akbar, Farrukh Shahzad, Mudassar Farooq, and Zeashan Khan. "Secure biometric template generation for multi-factor authentication." *Pattern Recognition* 48, no. 2 (2015): 458-472.

[48] Moss, S., J. Melia, J. Sutton, C. Mathews, and M. Kirby. "Prostate-specific antigen testing rates and referral patterns from general practice data in England." *International journal of clinical practice* 70, no. 4 (2016): 312-318.

[49] Naoi, Makoto, Masao Nakatani, Kenshiro Otsuki, Yasuo Yabe, Thabang Kgarume, Osamu Murakami, Thabang Masakale et al. "Steady activity of microfractures on geological faults loaded by mining stress." *Tectonophysics* 649 (2015): 100-114.

[50] Pavel, Ana B., Dmitriy Sonkin, and Anupama Reddy. "Integrative modeling of multi-omics data to identify cancer drivers and infer patient-specific gene activity." *BMC systems biology* 10, no. 1 (2016): 16.

[51] Sang, Jitao, Yue Gao, Bing-kun Bao, Cees Snoek, and Qionghai Dai. "Recent advances in social multimedia big data mining and applications." *Multimedia Systems* 22, no. 1 (2016): 1-3.

[52] Tosun, Suleyman, Vahid B. Ajabshir, Ozge Mercanoglu, and Ozcan Ozturk. "Fault-tolerant topology generation method for application-specific network-on-chips." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34, no. 9 (2015): 1495-1508.

[53] Yang, Chunsheng, Yanni Zou, Pinhua Lai, and Nan Jiang. "Data mining-based methods for fault isolation with validated FMEA model ranking." *Applied Intelligence* 43, no. 4 (2015): 913-923.

[54] Zhang, Yongshuang, Changbao Guo, Hengxing Lan, Nengjuan Zhou, and Xin Yao. "Reactivation mechanism of ancient giant landslides in the tectonically active zone: a case study in Southwest China." *Environmental Earth Sciences* 74, no. 2 (2015): 1719-1729.

[55] Zimek, Arthur, and Jilles Vreeken. "The blind men and the elephant: On meeting the problem of multiple truths in data from clustering and pattern mining perspectives." *Machine Learning* 98, no. 1-2 (2015): 121-155.

[56] Gong, Ke, Panpan Wang, and Yi Peng. "Fault-tolerant enhanced bijective soft set with applications." *Applied Soft Computing* (2016).

[57] Haseena, Hassan H., Paul K. Joseph, and Abraham T. Mathew. "Classification of arrhythmia using hybrid networks." *Journal of medical systems* 35, no. 6 (2011): 1617-1630.

[58] Kumar, S. Udhaya, H. Hannah Inbarani, and S. Senthil Kumar. "Improved bijective-soft-set-based classification for gene expression data." In *Computational Intelligence, Cyber Security and Computational Models*, pp. 127-132. Springer India, 2014.

[59] Senthilkumar, S., H. Hannah Inbarani, and S. Udhayakumar. "Modified soft rough set for multiclass classification." In *Computational Intelligence, Cyber Security and Computational Models*, pp. 379-384. Springer India, 2014.

[60] Wei, Song, Hani Hagras, and Daniyal Alghazzawi. "A cloud computing based Big-Bang Big-Crunch fuzzy logic multi classifier system for Soccer video scenes classification." *Memetic Computing* 8, no. 4 (2016): 307-323.

[61] Fernández, Alberto, Sara del Río, Abdullah Bawakid, and Francisco Herrera. "Fuzzy rule based classification systems for big data with MapReduce: granularity analysis." *Advances in Data Analysis and Classification* (2016): 1-20.

[62] Nelles, Oliver. "Unsupervised Learning Techniques." In *Nonlinear System Identification*, pp. 137-155. Springer Berlin Heidelberg, 2001.

[63] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35, no. 8 (2013): 1798-1828.

[64] Huang, Fei, and Alexander Yates. "Biased representation learning for domain adaptation." In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1313-1323. Association for Computational Linguistics, 2012.

[65] Tu, Wenting, and Shiliang Sun. "Cross-domain representation-learning framework with combination of class-separate and domain-merge objectives." In *Proceedings of the 1st International*

Workshop on Cross Domain Knowledge Discovery in Web and Social Network Mining, pp. 18-25. ACM, 2012.

[66] Li, Shou-Shan, Chu-Ren Huang, and Cheng-Qing Zong. "Multi-domain sentiment classification with classifier combination." *Journal of Computer Science and Technology* 26, no. 1 (2011): 25-33.

[67] F Huang, E Yates, Exploring representation-learning approaches to domain adaptation, in Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (Uppsala, 2010), pp. 23–30

[68] A Bordes, X Glorot, JWAY Bengio, Joint learning of words and meaning representations for open-text semantic parsing, in Proceedings of 15th International Conference on Artificial Intelligence and Statistics (La Palma, 2012), pp. 127–135

[69] N. Boulanger-Lewandowski, Y. Bengio, P. Vincent, Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription. arXiv preprint (2012). arXiv:1206.6392

[70] K Dwivedi, K Biswaranjan, A Sethi, Drowsy driver detection using representation learning, in Proceedings of the IEEE International Advance Computing Conference (Gurgaon, 2014), pp. 995–999

[71] D Yu, L Deng, Deep learning and its applications to signal and information processing. IEEE Signal Proc Mag 28(1), 145–154 (2011)

[72] I Arel, DC Rose, TP Karnowski, Deep machine learning-a new frontier in artificial intelligence research. IEEE Comput Intell Mag 5(4), 13–18 (2010)

[73] Y Bengio, Learning deep architectures for AI. Foundations Trends Mach Learn 2(1), 1–127 (2009)

[74] R Collobert, J Weston, L Bottou, M Karlen, K Kavukcuoglu, P Kuksa, Natural language processing (almost) from scratch. J Mach Learn Res 12, 2493–2537 (2011)

[75] P Le Callet, C Viard-Gaudin, D Barba, A convolutional neural network approach for objective video quality assessment. IEEE Trans Neural Networ 17(5), 1316–1327 (2006)

[76] GE Dahl, D Yu, L Deng, A Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Proc 20(1), 30–42 (2012)

[77] G Hinton, L Deng, Y Dong, GE Dahl, A Mohamed, N Jaitly, A Senior, V Vanhoucke, P Nguyen, TN Sainath, B Kingsbury, Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Proc Mag 29(6), 82–97 (2012)

[78] DC Ciresan, U Meier, LM Gambardella, J Schmidhuber, Deep, big, simple neural nets for handwritten digit recognition. Neural Comput 22(12), 3207–3220(2010)

[79] D Peteiro-Barral, B Guijarro-Berdiñas, A survey of methods for distributed machine learning. Progress in Artificial Intelligence 2(1), 1–11 (2012)

[80] H Zheng, SR Kulkarni, HV Poor, Attribute-distributed learning: models, limits,and algorithms. IEEE Trans Signal Process 59(1), 386–398 (2011)

[81] H Chen, T Li, C Luo, SJ Horng, G Wang, A rough set-based method for updating decision rules on attribute values' coarsening and refining. IEEE Trans Knowl Data Eng 26(12), 2886–2899 (2014)

[82] J Chen, C Wang, R Wang, Using stacked generalization to combine SVMs in magnitude and shape feature spaces for classification of hyperspectral data. IEEE Trans Geosci Remote 47(7), 2193–2205 (2009)

[83] E Leyva, A González, R Pérez, A set of complexity measures designed for applying meta-learning to instance selection. IEEE Trans Knowl Data Eng 27(2), 354–367 (2014)

[84] M Sarnovsky, M Vronc, Distributed boosting algorithm for classification of text documents, in Proceedings of the 12th IEEE International Symposium on Applied Machine Intelligence and Informatics (SAMI) (Herl'any, 2014), pp. 217–220

[85] SR Upadhyaya, Parallel approaches to machine learning—a comprehensive survey. J Parallel Distr Com 73(3), 284–292 (2013)R Bekkerman, M Bilenko, J Langford, Scaling up machine learning: parallel and distributed approaches (Cambridge University Press, Oxford, 2011)

[86] EW Xiang, B Cao, DH Hu, Q Yang, Bridging domains using world wide knowledge for transfer learning. IEEE Trans Knowl Data Eng 22(6), 770–783 (2010)

[87] SJ Pan, Q Yang, A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10), 1345–1359 (2010)

[88] W Fan, I Davidson, B Zadrozny, PS Yu, An improved categorization of classifier's sensitivity on sample selection bias, in Proceedings of the 5th IEEE International Conference on Data Mining (ICDM) (Brussels, 2012), pp. 605–608

[89] J Gao, W Fan, J Jiang, J Han, Knowledge transfer via multiple model local structure mapping, in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Las Vegas, 2008), pp. 283-291

[90] C Wang, S Mahadevan, Manifold alignment using procrustes analysis, in Proceedings of the 25th International Conference on Machine Learning (ICML) (Helsinki, 2008), pp. 1120–1127

[91] X Ling, W Dai, GR Xue, Q Yang, Y Yu, Spectral domain-transfer learning, in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Las Vegas, 2008), pp. 488–496

[92] R Raina, AY Ng, D Koller, 2006, Constructing informative priors using transfer learning, in Proceedings of the 23rd International Conference on Machine Learning (ICML) (Pittsburgh, 2006), pp. 713–720

[93] J Zhang, Deep transfer learning via restricted Boltzmann machine for document classification, in Proceedings of the 10th International Conference on Machine Learning and Applications and Workshops (ICMLA) (Honolulu,2011), pp. 323–326

[94] Y Fu, B Li, X Zhu, C Zhang, Active learning without knowing individual instance labels: a pairwise label homogeneity query approach. IEEE Trans Knowl Data Eng 26(4), 808–822 (2014)

[95] B Settles, Active learning literature survey (University of Wisconsin, Madison, 2010) MM Crawford, D Tuia, HL Yang, Active learning: any value for classification of remotely sensed data? P IEEE 101(3), 593–608 (2013)

[96] MM Haque, LB Holder, MK Skinner, DJ Cook, Generalized query-based active learning to identify differentially methylated regions in DNA. IEEE ACM Trans Comput Bi 10(3), 632–644 (2013)

[97] D Tuia, M Volpi, L Copa, M Kanevski, J Munoz-Mari, A survey of active learning algorithms for supervised remote sensing image classification. IEEE J Sel Top Sign Proces 5(3), 606–617 (2011)

[98] G Ding, Q Wu, YD Yao, J Wang, Y Chen, Kernel-based learning for statistical signal processing in cognitive radio networks. IEEE Signal Proc Mag 30(4), 126–136 (2013)

[99] C Li, M Georgiopoulos, GC Anagnostopoulos, A unifying framework for typical multitask multiple kernel learning problems. IEEE Trans Neur Net Lear Syst 25(7), 1287–1297 (2014)

[100] G Montavon, M Braun, T Krueger, KR Muller, Analyzing local structure in kernel-based learning: explanation, complexity, and reliability assessment. IEEE Signal Proc Mag 30(4), 62–74 (2013)

[101] K Slavakis, S Theodoridis, I Yamada, Online kernel-based classification using adaptive projection algorithms. IEEE Trans Signal Process 56(7), 2781–2796 (2008)

[102] S Theodoridis, K Slavakis, I Yamada, Adaptive learning in a world of projections. IEEE Signal Proc Mag 28(1), 97–123 (2011)

[103] K Slavakis, S Theodoridis, I Yamada, Adaptive constrained learning in reproducing kernel Hilbert spaces: the robust beamforming case. IEEE Trans Signal Process 57(12), 4744–4764 (2009)

[104] K Slavakis, P Bouboulis, S Theodoridis, Adaptive multiregression in reproducing kernel Hilbert spaces: the multiaccess MIMO channel case. IEEE Trans Neural Netw Learn Syst 23(2), 260–276 (2012)

[105] KR Müller, S Mika, G Rätsch, K Tsuda, B Schölkopf, An introduction to kernel based learning algorithms. IEEE Trans Neural Networ 12(2), 181–201 (2001)

[106] Kocaguneli, Ekrem, Tim Menzies, and Emilia Mendes. "Transfer learning in effort estimation." *Empirical Software Engineering* 20, no. 3 (2015): 813-843.

[107] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35, no. 8 (2013): 1798-1828.

[108] Van Hulse J, Khoshgoftaar T. Knowledge discovery from imbalanced and noisy data. Data Knowl Eng. 2009;68(12):1513–42.

[109] Khoshgoftaar TM, Hulse JV. Imputation techniques for multivariate missingness in software measurement data.Software Quality J. 16(4):563–600; 2008.

[110] Khoshgoftaar TM, Van Hulse J, Napolitano A. Comparing boosting and bagging techniques with noisy and imbalanced data. Syst Man Cybern Part A Syst Hum IEEE Trans. 2011;41(3):552–68.

[111] Van Hulse J, Khoshgoftaar TM, Napolitano A. Experimental perspectives on learning from imbalanced data. In:Proceedings of the 24th International Conference on Machine Learning; 2007. pp. 935–42.

[112] Hogan JM, Peut T. Large Scale Read Classification for Next Generation Sequencing. Procedia Comput Sci.2014;29:2003–12.

[113] Sun K, Miao W, Zhang X, Rao R. An Improvement to Feature Selection of Random Forests on Spark. In: 2014 IEEE 17th International Conference on Computational Science and Engineering (CSE); 2014. pp. 774–9.

[114] Ooi, Beng Chin, Kian-Lee Tan, Sheng Wang, Wei Wang, Qingchao Cai, Gang Chen, Jinyang Gao et al. "SINGA: A distributed deep learning platform." In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 685-688. ACM, 2015.

[115] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell.Ca_e: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014.

[116] M. Lin, S. Li, X. Luo, and S. Yan. Purine: A bi-graph based deep learning framework. CoRR, abs/1412.6249,2014.

[117] N. Vasilache, J. Johnson, M. Mathieu, S. Chintala, S. Piantino, and Y. LeCun. Fast convolutional nets with fb_t: A GPU performance evaluation. CoRR, abs/1412.7580, 2014.

[118] Zaharia, Matei, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. "Spark: Cluster Computing with Working Sets." *HotCloud* 10, no. 10-10 (2010): 95.

[119] T. White, Hadoop: The Definitive Guide, O'Reilly Media, 2009.

[120] Caudill, Steven B., and Franklin G. Mixon. "Analysing misleading discrete responses: A logit model based on misclassified data." *Oxford Bulletin of Economics and Statistics* 67, no. 1 (2005): 105-113.

[121] Brodley, Carla E., and Mark A. Friedl. "Identifying mislabeled training data." *Journal of Artificial Intelligence Research* 11 (1999): 131-167.

[122] Miranda, André LB, Luís Paulo F. Garcia, André CPLF Carvalho, and Ana C. Lorena. "Use of classification algorithms in noise detection and elimination." In *International Conference on Hybrid Artificial Intelligence Systems*, pp. 417-424. Springer Berlin Heidelberg, 2009.

[123] Van den Hout, Ardo, and Peter GM Van der Heijden. "The analysis of multivariate misclassified data with special attention to randomized response data." *Sociological Methods & Research* 32, no. 3 (2004): 384-410.

[124] Bilgic, Mustafa, and Lise Getoor. "Reflect and correct: A misclassification prediction approach to active inference." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3, no. 4 (2009): 20.

[125] Ciraco, Michelle, Michael Rogalewski, and Gary Weiss. "Improving classifier utility by altering the misclassification cost ratio." In *Proceedings of the 1st international workshop on Utility-based data mining*, pp. 46-52. ACM, 2005.

[126] Healy, J. D. "The effects of misclassification error on the estimation of several population proportions." *Bell System Technical Journal* 60, no. 5 (1981): 697-705.

[127] Smith, Michael R., Tony Martinez, and Christophe Giraud-Carrier. "An instance level analysis of data complexity." *Machine learning* 95, no. 2 (2014): 225-256.

[128] Evans, Robert. "Apache storm, a hands on tutorial." In *Cloud Engineering (IC2E), 2015 IEEE International Conference on*, pp. 2-2. IEEE, 2015.

[129] Olston, Christopher, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. "Pig latin: a not-so-foreign language for data processing." In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1099-1110. ACM, 2008.

[130] Presutti, Valentina, Francesco Draicchio, and Aldo Gangemi. "Knowledge extraction based on discourse representation theory and linguistic frames." In *International Conference on Knowledge Engineering and Knowledge Management*, pp. 114-129. Springer Berlin Heidelberg, 2012.

[131] Brown, Samuel DJ, Rupert A. Collins, Stephane Boyer, MARIE-CAROLINE LEFORT, J. A. G. O. B. A. MALUMBRES-OLARTE, Cor J. Vink, and Robert H. Cruickshank. "Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding." *Molecular Ecology Resources* 12, no. 3 (2012): 562-565.

[132] Mika, Sebastian, Bernhard Schölkopf, Alexander J. Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. "Kernel PCA and De-noising in feature spaces." In *NIPS*, vol. 11, pp. 536-542. 1998.

[133] Zucker, David M., and Donna Spiegelman. "Corrected score estimation in the proportional hazards model with misclassified discrete covariates." *Statistics in medicine* 27, no. 11 (2008): 1911-1933.

[134] Yuan, Yang C. "Multiple imputation for missing data: Concepts and new development (Version 9.0)." *SAS Institute Inc, Rockville, MD* 49 (2010): 1-11.

[135] Zucker, David M., and Donna Spiegelman. "Corrected score estimation in the proportional hazards model with misclassified discrete covariates." *Statistics in medicine* 27, no. 11 (2008): 1911-1933.Akazawa K, Kinukawa N, Nakamura T. A note on the corrected score function corrected formisclassification. Journal of the Japan Statistical Society. 1998; 28:115–123.

[136] Nakamura T. Corrected score function of errors-in-variables models: methodology and application to generalized linear models. Biometrika. 1990; 77:127–137.

[137] Spiegelman, Donna, Aidan McDermott, and Bernard Rosner. "Regression calibration method for correcting measurement-error bias in nutritional epidemiology." *The American journal of clinical nutrition* 65, no. 4 (1997): 1179S-1186S.

[138] Gaurav Jain, Kunal Gupta, Arpit Kushwah, Abhishek Agrawal, "*A Survey on Various Issues Big Data in Cloud Computing*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.9, pp.131-134, 2017.

Authors Profile

Miss Jyoti Arora pursed Master's of Technology in Computer Science Engineering from Guru Nanak Dev University Amritsar *in* year 2017. She pursed Bachelor of Technology (CSE) from LLRIET, Moga in year 2014. She is currently working as Assistant Professor in Department of Chitkara School of Engineering & Technology. She has 11months experience of Sacred Heart School, Moga as Maths Teacher.Her research interest is Cloud Computing, Big Data, Data Mining.

Miss Ambica Sood pursed Master's of Technology in Computer Science Engineering from Guru Nanak Dev University Amritsar *in* year 2017. She pursed Bachelor in Information Technology (IT) from GIMET, Amritsar in year 2015. She is currently working as Teaching Assistant in Computer Science Engineering Department of Chandigarh University.