# A Review on Class Imbalance Problem: Analysis and Potential Solutions

**Satyam Maheshwari[1], Dr R C Jain[2] and Dr R S Jadon[3]**

**[1] Department of Computer Applications**
**SATI, Vidisha, India**

**[2] Ex-Director**
**SATI, Vidisha, India**

**[3] Department of Computer Applications**
**MITS, Gwalior, India**

## Abstract

The Imbalanced class problem is a recent challenge in data mining. A dataset is said to be imbalance when their classification categories are not properly mentioned, and the class which has fewer instances as compare to other classes is of more interest from the point of view of the learning task. Here we study about the various factors that influence the datasets and leads to imbalance the dataset like as Features selection, classification of imbalance datasets. Here in this paper, we also discussed about the various sampling methods utilize for dataset balancing and for getting measurable performance.

***Keywords:*** *Imbalanced classification, Preprocessing, Feature selection, Cost-sensitive learning, Ensemble learning.*

## 1. Introduction

The Imbalance data learning issues, attends much interest from industries, academics & research teams, refer as the top most challenging issues in the field of the data mining [1], have high attention committed in scientific publication [2]. These issues have been observed in several fields like as medical diagnosis [3], detection of fraudulent calls [4], risk management [5], text classification [6], modern manufacturing plants [7], detection of oil spills from satellite images [8], fault diagnosis [9], [10], anomaly detection [11], [12], and face recognition [13]. When a model prepared with imbalanced data set, it ultimately gives its inclination towards class, as the classic learning algorithm increased the level of accuracy. Inductive classifiers are built to decrease the faults based on training instances, At the time of the learning algorithm, can overlook classes having less instances [14].

The various techniques have been generated to control such kind of situation, from the general sampling adjustment to highly complicated such as modifying the algorithm.

These imbalance errors have gain much focus from the fields such as Machine Learning & pattern recognition. When the single class is highly advertised as compared to the other class based on the majority. This concept is mostly required in the real world applications, where it becomes expensive for not classifying the examples based on the minority class, like as searching of the fraudulent telephone calls, diagnosis of rare diseases, information retrieval, text categorization and filtering tasks [15].

The many works have been done to resolve such kind of issues, which are being divided into two groups:

1) To generate efficient algorithm or replace the existing ones to detect the problem is defined as internal approaches.
2) Process the data in advance to remove the effect offered by the class imbalance taken as the external work process.

The inner work process has some drawback of dealing with specific algorithm, on the other hand the external work process is separate from the classifier which is being used and more adjustable, due to this the CO2RBFN is applicable for resolving the errors occurred in imbalanced classification [16]. The actual classification of the minority class is more crucial as compare to the majority classes, for instance, in predicting protein interactions, the numbers of non-interacting proteins are higher as compare to the number of interacting proteins. In medical analysis the number of disease causes is lesser as compared to non-diseases cases [17].

The figure 1, describes the distribution of majority class, minority class and noise samples.

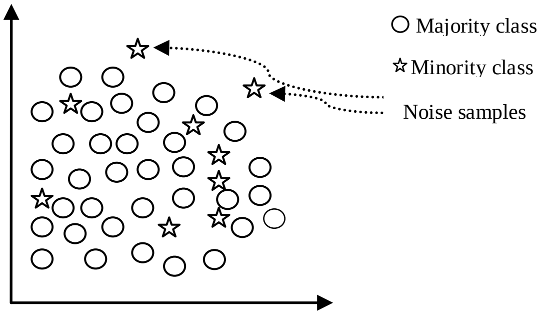

Fig 1. The data set having a between-class imbalance

The class imbalance problem exists in a large number of domains, some of them are Medical diagnosis, Fraud detection, Risk management, Fault diagnosis, detection of oil spills and Face recognition. These are some examples which suffers most due to a class imbalance problem.

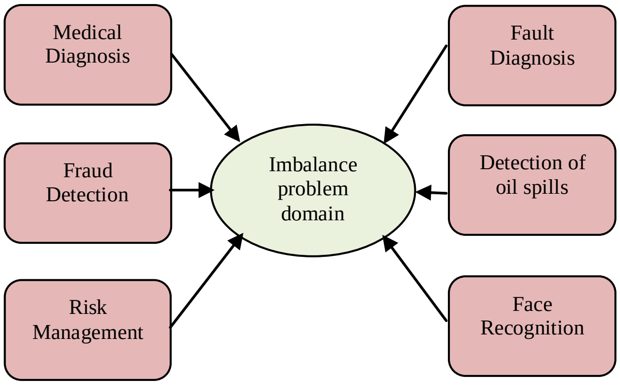

Fig 2. The area which suffers most due to a class imbalance problem

## 2. Learning of Imbalance Dataset

At present the main attention of the class imbalance is on the summation of the imbalance class learning based on the AL (Active Learning) techniques. The association of the latest techniques with the class imbalance learning becomes a famous topic in the field of the imbalance class learning. Few earlier researches of them are as below:-

The brief description of various CIL (class imbalance learning) methods is presented in [18].

The upcoming part discuss in details about the external & internal imbalance learning method. The external methods separate from the algorithm which is being used. Also have the dataset to cover them before the classifiers. Various re-sampling methods, like as random & focused oversampling and under-sampling, comes with this stream. Here in random under-sampling all the majority–class is vanished in a random manner till it's doesn't meet a specific class [19]. In case of oversampling, all the minority-class randomly generates the copies, till a class ratio doesn't meet [18].

In [22] the author has introduced various techniques depend on the k-means clustering & genetic algorithm. These algorithms deal with 2 processes, firstly k-means is taken to determine the cluster in the minority set and secondly, it is dealt with high-tech and efficient online samples for the re-sampling process. Classifier permits the method which is taken in handling the errors associate with the imbalance class learning. The author has introduced a new approach for clustering depended on the sampling technique for controlling the problems. The Evaluation algorithms are also taken as a strong source of controlling the class imbalance problems.

In [23] the author has represented the evolutionary techniques general nested exemplar groups, which deals with the Euclidean n-space to save the elements, while calculating the distance to the closest generalized exemplar. This technique deals with the evolutionary algorithm for picking the most appropriate exemplars for the purpose of re-sampling.

In [24] the author has introduced an evolutionary cooperative, competitive planning for designing the radial-basis function networks CO2RBFN with the help of the cooperative competitive methods having a radial - basis function on imbalanced datasets. In CO2RBFN which separately shows the portion of the solution only and managing so that to build the complete RBFN, then they gets better generalization for latest pattern by showing whole information about the errors volume.

In [25] the authors have introduced to work with dataset with the help of the preprocessing step having fuzzy rule depended on the classification system by the action of the adaptive inference system having parametric conjunction operators.

In [26] the authors have introduced the applications of K-nearest Neighbor (k-NN) classifier to analysis the performance. In [27] the authors discuss about the effect of various classifiers with distinct re-sampling plan on imbalanced datasets having distinct imbalance ratios.

https://doi.org/10.20943/01201706.4351

The figure 3 describes the various methods used in dealing with the class imbalance problem. Some of them are:

(i) Data level approach,

(ii) Algorithm level approach,

(iii) Cost-sensitive approach, and

(iv) Ensemble learning.

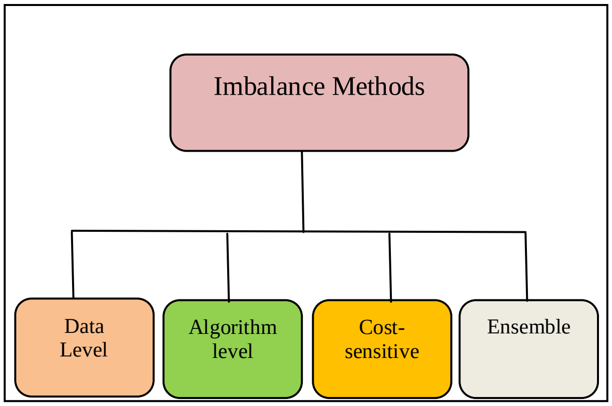

Fig 3. Class imbalance methods

## 2.1 Data Preprocessing

Data preprocessing comprises of two different forms of re-sampling:

Undersampling: Random under-sampling [33] is referred as the non-heuristic technique having a motive to maintain the distribution of the classes with the help of the removal of the majority classes randomly. The main cause is to balance the dataset to remove the idiosyncrasies of that machine learning algorithm.

Oversampling: Random over-sampling [33] referred as a heuristic technique that motive is to balance the distribution of class with the help of random replication of minority class examples. Random over-sampling can enhance the likelihood of getting overfitting, as it gives same copies of minority class examples. Therefore, a symbolic classifier, generates some rules which are accurate, but also gives a replicated example. Moreover oversampling gives an additional computational work when the data set is bigger but imbalanced.

## 2.2 Algorithm level methods for handling imbalance

Drummond and Holte [29] give the report, while using the C4. 5's default settings, oversampling is ineffective, mostly gives little or no change in performance at the time of modification of class distribution & misclassification costs. Further, they observe that over-sampling prunes less and hence generalizes less than under-sampling, and that an adjustment of the C4.5's parameter settings to enhance the influence of pruning and other overfitting ignorance factors can rebuild the performance level of over-sampling. Discrimination process based on the internal biasing, a measured distance element is introduced [30] which works with KNN classification. The reason behind this weighted distance is to pay for a training sample of the imbalance dataset in the absence of distracting the class. Hence the weighted are given not in the general k-NN pattern, also not to any prototype separately. In this way the weighting element becomes higher in the majority class as compare to the minority one. The farness of positive minority class prototype becomes less as compare to majority class. This will let the latest pattern to search their closest prototype in the minority class.

One more work process on the imbalance data set is based on the SVM biases algorithm in this the hyper planes are going away from the positive class. This happens due to the reason that to adjust the skew attached with the imbalance dataset, which force the hyper plane to go closer towards the positive class. This type of biasing can be done in several ways. Chang et al [31] proposed the varying kernel function to generate the bias. The veropoulos et al [32] recommend using various penalty constant for several classes, committing faults at the positive instances become expensive as compare to the negative instance.

## 2.3 Cost-Sensitive Methods

It incorporates both data level and algorithm level when the misclassification cost is high. The main feature of this method is that it tries to minimize the total cost of misclassification. In cost-sensitive methods it is more interesting to recognize the positive instances rather than the negative ones. For example, in medical domain the cost of misclassifying a non cancerous patient is limited to additional medical tests, while the cost of misdiagnosis of will be fatal as potentially cancerous patients will be considered healthy. Therefore, the cost associated with a positive instance must be greater than the cost of misclassifying a negative one, i.e. C(+,-)>C(-,+).

## 2.4 Ensemble Methods

Ensemble based classifiers are designed to improve the accuracy of a single classifier by training several classifiers and combining them to output a new classifier that outperforms every one of them. Therefore, ensemble based methods are based on the combination between ensemble learning algorithms and the hybrid approaches such as data and algorithms, or cost-sensitive learning solutions. In algorithm level approaches, instead of modifying the base classifier, ensemble learning algorithms slightly modify the base learner. On the other hand, in the case of data level approaches, the new method preprocess the data before learning each classifier.

### 2.5 Feature Selection for imbalance datasets

Zheng et al [28] developed a process for feature selection which is appropriate for high-dimensional unbalanced data sets. They introduce some feature selection framework, which takes the features for positive and negative classes individually and then specifically merge them. The authors suggest general ways for transforming existing measures in order to consider features for negative and positive classes separately.

## 3. Motivation

The rare events are those events which take place very rapidly, that is their frequency lies between the 5% to 10% based on the application. Classification in the rare events refers as a general issue in various domains such as network intrusion detection, fraudulent transactions, direct marketing, & medical diagnostics. For instance, in the network intrusion detection domain, the number of intrusions is very less amount of overall network traffic. In medical databases, at the time of image classification [36], pixels which are not normal shows only little amount among the overall image.

The behavior of the image needs fast detection rates and permits low error rate as the cost of that misclassifying a cancerous patient is very expensive. Here when all these majority class shows 98-99% of that overall population, a trivial classifier that links all with the class of majority to gain accurate output. It's applicable mostly for the skewed distribution or imbalance dataset. The accuracy of the classification can't take as a measure of performance level. The ROC observation [37] & metrics such as precision, recall and F-value [38, 39] are taken for measuring the performance of learning algorithm on that minority class. The superiority of class imbalance in several fields leads to a surge in research working for the minority classes. The various work process for working with imbalance data set are discussed in [36, 39, 40].

The analysis drawn from the comparative study of each of the following methods is shown in the Table 1:

Table 1: Comparative study

| *Approach* | *Methods & its description* | *Algorithms* | *Advantages* | *Disadvantages* |
|---|---|---|---|---|
| *Data Level approach* | Under sampling -This method randomly removes samples of majority class | Random under sampling [33], One Sided Selection (OSS)[35], Neighborhood Cleaning Rule(NCL)[34], Tomek Links[21] | It is more versatile and independent of classifier selected, therefore data need to be prepared once for classification | It sometimes throws the important data, which may be useful in the induction process. |
| | Oversampling - This method adds new samples in existing class | Random oversampling [33], Synthetic Minority Oversampling Technique(SMOTE)[42] | Generates rules which are accurate and also used to improve the accuracy of classification | The problem of overfitting / overgeneralization occurs |
| | Hybrid - In this method samples from both classes are removed | SMOTE + Tomek links [33], SMOTE + ENN [33] | Overcome the problem of oversampling, but not by cutting down the size of majority classes | A longer training time |
| *Algorithm level approach* | Bagging method - improve the stability and accuracy of ensemble learning algorithms | Decision tree(C4.5)[44], Random Forest [43] | Reduce dissimilarity and better classification performance than individual classifiers | It takes very much time to process, may lead to overfitting |
| | Boosting method | AdaBoost [20], SMOTEBOOST [58] | It boosts the performance of other learning methods | Ignore overall performance of the classifier |
| *Cost-sensitive approach* | It incorporates both data level and algorithm level when the misclassification cost is high- Cost sensitive boosting algorithms | | It tries to minimize the total cost of misclassifications | Cost is not precisely known, have to use approximations or ratios of proportionate |

## 4. Classification of Various Methods

The various methods have been proposed, some of them are:

### 4.1 Sampling Methods

It is taken as a simple data level technique helpful for balancing the classes made up of re-sampling the actual data set, or by oversampling the minority class or by under-sampling the majority class, till the classes are not become approximately equally represented. Both strategies are applicable in all sorts of learning system, as they work as a preprocessing phase, permitting the learning system to get the training instances as if they associate with a well-balanced data set. Hence, any bias of a system for majority class because of various proportion per instance, then every class become likely to suppress. The Hulse et al. [41] recommends the use of the re-sampling techniques based on various factors, having the ratio in between the negative and positive examples, and some other specialties of the data & behavior of the classifier, but somehow the re-sampling techniques have some crucial errors. The under sampling may throw away important data, on the other hand, them over sampling enhance the size of that dataset, increased the computational burden of the algorithm.

### 4.1.1 Oversampling

The easiest technique to enhance the size the minority class relates with random over-sampling, which is referred as a non heuristic technique which equalize the distribution of class with the help of randomly repetition of positive examples, still this technique replicates the running examples, the possibilities of overfitting occurrence is high. Chawla suggests an approach that is a Synthetic Minority Over-sampling method [42] where the minority class is referred as the over sampled by generating synthetic example instead of replacement based over sampling. The process of over-sample of minority happens by putting in the synthetic examples, also with the line segments linking with all\any minority class closest neighbor. Based on the value of over sampling needs, the entire links up elements in the k nearest are selected in a random manner.

Based on SMOTE algorithm, various modifications have suggested in the literature. The SMOTE work process, not able to control its dataset function. It was being generalized to control mixed datasets of nominal features and continuous feature, Chawla suggested SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous) and SMOTE-N (Synthetic Minority Oversampling Technique Nominal), the SMOTE can be forwarded towards nominal functions.

SMOTE (Synthetic Minority Oversampling Technique) was introduced to calculate the effect of data set in the minority class [43]. SMOTE generates synthetic instances of the minority class through working on the “feature space” instead on “data space”. By synthetically creation high instances of the minority classes, the inductive learners like as decision trees (e.g. C4.5 [44]) or rule-learners (e.g. RIPPER [45]), capable of making the decision area broad. We work on the continuous & discrete properties distinctly in SMOTE. For the closest neighbor, we deal with Euclidean distance for the purpose continuous feature & Value Distance Metric (with the Euclidean assumption) as in nominal feature [46, 47]. The latest minority samples are as below:

- For the continuous features.
- Take the difference between a feature vector (minority class sample) and one of its k nearest neighbors (minority class samples).
- Multiply this difference by a random number between 0 and 1.
- Add this difference to the feature value of the original feature vector, thus creating a new feature vector.
- For the nominal features.
- Take majority vote between the feature vector under consideration and its k nearest neighbors for the nominal feature value. In the case of a tie, choose at random.
- Assign that value to the new synthetic minority class sample.

With the help of these methods, a latest minority sample of class is generated next to that of the class sample. These neighbors are used based on the value of SMOTE. Therefore, applying the SMOTE, high normal regions are being studied in that minority class, permitting classifier to predict the unknown example related to the minority classes. The pairing of the SMOTE and under sampling makes optimal classifier from the SMOTE in the majority and that under-sampling found at the convex hull of ROC curve [42, 48].

### 4.1.2 Undersampling

Under sampling is an effective technique for imbalance learning classing. This process applies the subset of majority class for classifier training. As several majority class are overlooked, this leads to make the training set equalize, and speed up the process. The general method such as Random majority under-sampling (RUS), in RUS, the major disadvantage of under-sampling is the ignorance of the important information present in the examples. There are various ways for improving its performance, like Tomek links, the Condensed Nearest Neighbor Rule and One-sided selection, etc. One-sided selection (OSS) is suggested by Rule Kubat and Matwin taken to practically under-sample the majority class by eradicating majority class examples that refer as redundant or 'noisy'. Over-sampling is a technique for better minority class recognition; duplication of minority data randomly with some increment in latest information is also dominants to over fitting.

### 4.2 Feature Selection Methods

The various feature selection methods are:

#### 4.2.1 Correlation coefficient

The correlation coefficient refers as a statistical exam that calculates the power and sort of the connection between two variables. Correlation coefficients lie from -1 to 1. The exact value of the coefficient shows the strength of the relationship; absolute values near to 1 show a tough relationship. The symbol of the coefficient shows the way of the relationship: a positive sign indicates that the two variables increase or decrease with each other and a negative sign indicates that one variable increases as the other gets decreases.

The correlation coefficient is used to analyze the accuracy of the machine learning problems, and then the other functions are based on the ranking level [49]. In the errors where the covariance (Xi, Y) lies in between the Feature that is (Xi) & target such as (Y), the variances suppose as (var (Xi)) & target (var (Y)) are familiar, then the correlation would compute directly [50].

#### 4.2.2 Chi-square

Chi is as the process of measuring the separate functions of the class statistically. It is referred as a dual side matrix. This approach response erratically at the time of the low counts of its features, this factor is common between the imbalance data set [51]. On the other hand the chi-square test is well applicable to the nominal data, its drops out at the time of continuous data flow [52].

#### 4.2.3 Odds Ratio

Odd Ratio watches the odds of a function occurring in its positive class generalize by the odds of the feature occurring in the negative class.
The standard odds ratio refers as the single-sided metric. If the zero count is present at the denominator, we exchange the denominator by 1. This shows the consistency how Forman calculate the dual-sided odds ratio [53]. A pilot research found the single-sided algorithm offers better on our data, but other studies, like as Forman [53] have worked with the dual-sided algorithm. This metric is built up to work on solely having binary data sets [30].

#### 4.2.4 Signal-to-noise Correlation Coefficient

S2N calculates the ratio of few needed signal such as the class labels to the surrounding noise in a feature. At that time this ratio is referred as an electrical engineering factor, here in machine learning community, it is being pertain to leukemia classification with tough results [43]. It is a single sided metric [52].

#### 4.2.5 Information Gain

IG calculates the variation between class label‘s entropy and that conditional entropy of the class label's feature. This measure is taken as dual-sided. Such as the chi-square test, it generalizes for nominal data, but not able to control, continuous data well due to the same reasons [52].

#### 4.2.6 Relief

This feature taking matrix depends on the closest neighbor through Kira & Rendell [54]. It describes the features depends on the level of distinguishing themselves. The time relief picks any particular instance, it finds any two closest neighbors.One of their personal class and the other one from the neighbor class, this proves that the instances of distinct classes possess very distinct values, whereas the instance of the same class possess same values, as the actual probability is hard to compute. This happens by computing the random instances & its closest hits & misses distances. The distance is "0" for the discrete variables & "1" for the continuous variables, we deals with standard Euclidean distance, We can choose the number to set of the instances, and selecting more number shows good approximation[55, 50]. The suggestive feature selection method :

For recommending the feature selection the probability of feature distributed function is appropriate for the information about the report of the sample distribution is various classes, based on this report the significance & privilege of every feature is found out. The prediction based techniques of the distribution function are further separated into 2 categories such as parametric & non-parametric [56].

The parametric technique is taken as a crucial process of distribution, so that the problems present in the probability distribution can be figured out on the basis of the distribution parameters. A problem present in the parametric refers as they are not having any pre-defined structure, which can be taken as the model for the distribution of the data. And all the classic parametric distributions are single exponential that mean that they are having single maximum point. Whereas maximum data sets of the actual world are referring as the Multi-exponential. The distribution element which is assumed with parametric techniques for that kind of data set is not appropriate [57].

The non-parametric technique have nothing to do with the structure and figure of the distribution function, samples, they compute the distribution function directly from the sample because of this reason non-technique is taken higher than that of parametric method [57]. Due to this reason the non-parametric technique used for figuring the probability distribution features function for various classes. The formula for the non-parametric probability distribution estimation function is given as

$$P(x) \cong \frac{K}{N \times V} \quad (1)$$

P(x) shows the appraisal amount of probability distribution function for sample x, v is the volume around the sample, N refer as a whole number of samples and also K is taken as the number of samples within a volume V. These concepts have been shown in the following picture based on situation of parameters K and V, the nonparametric estimation techniques are divided into two general groups: for that group of techniques where the amount of K is taken constant, all samples are taken as X and V is finds in the way that it surely includes K sample K methods known as nearest neighbor estimation. That group of methods which suppose V a constant amount and obtain the number of points existed in volume V in order to estimate the probability distribution function are called distribution estimation methods based on kernel [57].

## 5. Evaluation in Imbalanced Domain

The evaluation criteria are a key factor in assessing the classification performance and guiding the classifier modeling. In a two-class problem, the confusion matrix shown in Table 2, records the results correctly and incorrectly recognized examples of each class.

$$\text{Acc}= \frac{TP+TN}{TP+FN+FP+TN} \quad (2)$$

Table 2. Confusion matrix for a two-class problem

| | ***Predicted Positive ‘1’*** | ***Predicted Negative ‘0’*** |
|---|---|---|
| Actual Positive ‘1’ | True Positive (TP) | False Negative (FN) |
| Actual Negative ‘0’ | False Positive (FP) | True Negative (TN) |

Traditionally, the accuracy rate Eq. (2) has been the most commonly used empirical measures. However, in the framework of imbalanced data sets, accuracy is no longer a proper measure, since it does not distinguish between the numbers of correctly classified examples of different classes. Hence, it may lead to erroneous conclusions, i.e., a classifier achieving an accuracy of 90% in a dataset with an IR value of 9 is not accurate if it classifies all examples as negatives.

For this reason, when working in imbalanced domains, there are more appropriate metrics to be considered instead of accuracy. Specifically, we can obtain four metrics from Table 2 used to measure the classification performance of both, positive and negative classes independently.

1) *True positive rate* $TP_{rate} = \frac{TP}{TP+FN}$is the percentage of positive instances correctly classified.

2) *True negative rate* $TN_{rate} = \frac{TN}{FP+TN}$is the percentage of negative instances correctly classified.

3) *False positive rate* $FP_{rate} = \frac{FP}{FP+TN}$is the percentage of negative instances misclassified.

4) *False negative rate* $FN_{rate} = \frac{FN}{TP+FN}$is the percentage of positive instances misclassified.

Clearly, since classification intends to achieve good quality results for both classes, none of these measures alone are adequate by itself.

## 6. Conclusion

In this paper, we have discussed about the main characteristics of the imbalance dataset such as the feature selection, classification of the imbalance dataset, sampling techniques and its various algorithms. Here we have also focused on the several sorts of the solution regarding the misbalancing issue of the data sets. Finally, we conclude that by the application of the various techniques, which we have already discussed in this paper, will be helpful in fighting with the problems related to the imbalance datasets because the data level methods offer efficient results with the help of over sampling algorithms for processing and balancing the dataset.

To summarize, each method offers advantages and disadvantages which varied and depend on the context and type of data. Therefore, in IDL, no solution is optimal solution and the best solution depends on the context of learned data.

### Acknowledgments

I am thankful to Dr R C Jain sir, Dr R S Jadon sir and Dr Sunil Joshi, who have contributed towards the development of this paper.

## References

[1] Bhoj Raj Sharma, Daljeet Kaur and Manju, "A Review on Data Mining: Its Challenges, Issues and Applications", Vol. 3, No. 2, 25 June 2013.

[2] Yale Song, Louis-Philippe Morenci, and Randall Davis, "Distribution-Sensitive Learning for Imbalanced Datasets", in IEEE International Conference on Automatic Face and Gesture Recognition, 2013.

[3] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J.A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance", Neural Network, Vol. 21, No. 2–3, 2008, pp. 427–436.

[4] T. E. Fawcett, and F. Provost, "Adaptive fraud detection", in Data Mining and Knowledge Discovery, 1997, pp. 291–316.

[5] K. Ezawa, M. Singh, and S. W. Norton, "Learning goal oriented bayesian networks for telecommunications risk management", in Proc. of thirteenth Int. Conference on Machine Learning, 1996.

[6] C. Cardie, and N. Howe, "Improving minority class prediction using case-specific feature weights", in Proc. of Fourteenth International Conference on Machine Learning, Nashville, TN, 1997.

[7] P. Riddle, R. Segal, and O. Etzioni, "Representation design and brute-force induction in a Boeing manufactoring domain", Appli. Artif. Intell., 1991, pp. 125–147.

[8] M. Kubat, R. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images", Machine Learning, 1998, pp. 195–215.

[9] Z. Yang, W. Tang, A. Shintemirov, and Q. Wu, "Association rule mining based dissolved gas analysis for fault diagnosis of power transformers", IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., Vol. 39, No. 6, 2009, pp. 597–610.

[10] Z.-B. Zhu and Z.-H. Song, "Fault diagnosis based on imbalance modified kernel fisher discriminant analysis", Chem. Eng. Res. Des., Vol. 88, No. 8, 2010, pp. 936–951.

[11] W. Khreich, E. Granger, A. Miri, and R. Sabourin, "Iterative Boolean combination of classifiers in the roc space: An application to anomaly detection with hmms", Pattern Recogn., Vol. 43, No. 8, 2010, pp. 2732–2752.

[12] M. Tavallaee, N. Stakhanova, and A. Ghorbani, "Toward credible evaluation of anomaly-based intrusion-detection methods", IEEE Trans. Syst., Man, Cybern. C, Appl. Rev, Vol. 40, No. 5, Sep. 2010, pp. 516–524.

[13] Y.-H. Liu and Y.-T. Chen, "Total margin-based adaptive fuzzy support vector machines for multiview face recognition", in Proc. IEEE Int. Conf. Syst., Man Cybern., Vol. 2, 2005, pp. 1704–1711.

[14] Dr. D. Ramyachitra, P. Manikandan, "Imbalanced dataset classification and solutions: A review", Vol. 5, No. 4, 2014.

[15] V. García, J.S. Sánchez, R.A. Mollineda, R. Alejo, J.M. Sotoca, "The class imbalance problem in pattern classification and learning", Pattern Analysis and Learning Group, Dept.de Llenguatjes I Sistemes Informàtics, Universitat Jaume I.

[16] María Dolores Pérez-Godoy, Alberto Fernández, Antonio Jesús Rivera, María José del Jesus, "Analysis of an evolutionary RBFN design algorithm, CO2RBFN, for imbalanced data sets", Pattern Recognition Letters, Vol. 31, 2010, pp. 2375–2388.

[17] Putthiporn Thanathamathee, Chidchanok Lursinsap, "Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques", Pattern Recognition Letters, Vol. 34, 2013, pp. 1339–1347.

[18] Vaishali Ganganwar, "An overview of classification algorithms for imbalanced datasets", Vol. 2, No. 4, 2012, pp. 2250 – 2459.

[19] G. Weiss, "Mining with rarity: A unifying framework", SIGKDD Explor. Newsletter, Vol. 6, No. 1, 2004, pp. 7–19.

[20] Y. Freund, and R.E. Schapire, "Experiments with a new boosting algorithm", 13th International Conference on Machine Learning, 1996.

[21] I. Tomek, "Two modifications of CNN", IEEE Transaction System Man Communication, Vol. 6, 1976, pp. 769-772.

[22] Yang Yong, "The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm", Energy Proc., Vol. 17, 2012, pp. 164–170.

[23] J. Carmona and Francisco Herrera, "Evolutionary-based selection of generalized instances for imbalanced classification", Knowledge-Based Systems, Vol. 25, 2012, pp. 3–12.

[24] María Dolores Pérez-Godoy, Alberto Fernández, Antonio Jesús Rivera, María José del Jesus, "Analysis of an evolutionary RBFN design algorithm, CO2RBFN, for imbalanced data sets", Pattern Recognition, Vol. 31, 2010, pp. 2375–2388.

[25] Alberto Fernández, María José del Jesus, Francisco Herrera, "On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets", Expert Systems with Applications, Vol. 36, 2009, pp. 9805–9812.

[26] Jordan M. Malof, Maciej A. Mazurowski, Georgia D. Tourassi, "The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support", Neural Networks, Vol. 25, 2012, pp. 141–145.

[27] V. Garcia, J.S. Sanchez, R.A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance", Knowledge-Based Systems, Vol. 25, 2012, pp. 13–21.

[28] Zheng Z., Wu. X., Shihari R., "Feature Selection for text categorization on imbalanced data", Vol. 6, No. 2, pp. 80-89.

[29] C. Drummond and R.C. Holte, "C4.5, Class Imbalance and Cost Sensitivity: Why Under-sampling beats Over-

sampling", in Workshop on Learning from Imbalanced Data Sets, Vol. 2, 2003.

[30] Matías Di Martino, Alicia Fernández, Pablo Iturralde, Federico Lecumberry, "Novel classifier scheme for imbalanced problems", Pattern Recognition Letters, Vol. 34, 2013, pp. 1146–1151.

[31] G. Wu & E. Chang, "Class-Boundary Alignment for Imbalanced Dataset Learning", ICML Workshop on Learning from Imbalanced Data Set, Vol. 2, 2003.

[32] R. Muscat, M. Mahfouf, A. Zughrat, Y.Y. Yang, S. Thornton, A. V. Khondabi and S. Sortanos, "Hierarchical Fuzzy Support Vector Machine (SVM) for Rail Data Classification", The International Federation of Automatic Control Cape Town, South Africa, 2014, pp. 24-29.

[33] G.E.A.P.A. Batista, R. C. Prati, M.C. Monard, "A study of the behavior of several methods for balancing machine learning training data", SIGKDD Explor. Newsl. Vol. 6, No. 1, 2004, pp. 20-29.

[34] D. Wilson, "Asymptotic properties of nearest neighbor rules using edited data", IEEE Trans. System Man Commun., Vol. 2, No. 3, 1972, pp. 408-421.

[35] M. Kubat and S. Matwin., "Addressing the curse of imbalanced training sets: One sided selection", in Proceedings of the Fourteenth International Conference on Machine Learning, 1997, pp. 179-186.

[36] William A. Rivera and Amit Goel and J. Peter Kincaid, "Blending Propensity Score Matching and Synthetic Minority Over-Sampling Technique for Imbalanced Classification", Winter Simulation Conference, 2014.

[37] T. Fawatt, "An Introduction to ROC Analysis", pattern recognition letters, Vol. 27, No. 8, pp. 861-874, 2006.

[38] M. Hossin, M. N. Sulaiman, N. Mustpaha and R. W. Rahmat, "Improving Accuracy Metric With Precision And Recall Metrics For Optimizing Stochastic Classifier ", ICOCI, 2011.

[39] M. Joshi, V. Kumar, R. Agarwal, "Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements", in First IEEE International Conference on Data Mining, 2001.

[40] M. Joshi, R. Agarwal, V. Kumar, "Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong?", Proceedings of Eighth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002.

[41] J. Hulse, T. Khoshgoftaar, Napolitano, "A Experimental perspectives on learning from imbalanced data", Proceedings of the 24th International Conference on Machine learning, 2007, pp. 935–942.

[42] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Research, Vol. 16, 2002, pp. 321 –357.

[43] L. Breiman, "Random Forest", Machine Learning, Vol. 45, 2001, pp. 5-32.

[44] J. Quinlan, "C4.5: Programs for Machine Learning", San Mateo, CA: Morgan Kaufman, 1992.

[45] W. Cohen, "Fast Effective Rule Induction", Proceedings of the 12th International Conference on Machine Learning, 1995, pp. 115-123.

[46] C. Stanfill, D. Waltz, "Toward Memory-based Reasoning", Communications of the ACM, Vol. 29, No. 12, 1986, pp. 1213-1228.

[47] S. Cost, S. Salzberg, "A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features", Machine Learning, Vol. 10, No. 1, 1993, pp. 57-78.

[48] F. Provost, T. Fawcett, "Robust Classification for Imprecise Environments", Machine Learning, Vol. 42, 2001, pp. 203-231.

[49] I. Guyon, A Elisseeff, "An introduction to variable and feature selection", JMRL special Issue on variable and Feature Selection, Vol. 3, 2003, pp. 1157-1182.

[50] X. Chen and M. Wasikowski, "FAST: a roc-based feature selection metric for small samples and imbalanced data classification problems", In Proceeding of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, 2008, pp. 124-132.

[51] G. Forman, "An extensive empirical study of feature selection metrics for text classification", Journal of Machine Learning Research, Vol. 3, 2003, pp. 1289–1305.

[52] M. Wasikowski and X. Chen, "Combating the small sample class imbalance problem using feature selection", IEEE Transactions on knowledge and data engineering, 2009.

[53] G. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction", Journal of Artificial Intelligence Research, Vol. 19, 2003, pp. 315–354.

[54] K. Kira, and, L. Rendell, "The feature selection problem: Traditional methods and new algorithms", Proc. of the 9th International Conference on Machine Learning, 1992, pp. 249-256.

[55] I. Kononenko, "Estimating attributes: Analysis and extension of RELIEF", In Proc. of the 7th European Conference on Machine Learning, 1994, pp. 171-182.

[56] A.R. Webb, Statistical Pattern Recognition, Second Edition, Wiely, 2002.

[57] R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification, Second Edition, Wiley, 1997.

[58] N. V. Chawla, A. Lazarevic, L. O. Hall, K. W. Bowyer, "SmoteBoost: Improving predicition of the minority class in boosting", In 7th European Conference on Principles and Practice of Knowledge Discovery in Database, 2003, pp. 107-119.

Satyam Maheshwari is M.Tech. and have 12 years of experience as an Assistant professor at SATI, Vidisha. He is a Ph.D. scholar at RGPV, Bhopal. He is a life member of ISTE and has more than 6 research papers in the field of computer science.

Dr R C Jain is Ex-Director of SATI, Vidisha. He has experience of more than 40 years and guided more than 27 Ph.D. in the field of computer science.

Dr R S Jadon is a Professor and Head of Computer applications Dept. of MITS, Gwalior. He has experience of more than 25 years and guided more than 10 Ph.D. in the field of computer science.