International Journal of Engineering & Technology, 7 (2.7) (2018) 786-790

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper



Classification of Imbalanced Malaria Disease Using Naïve Bayesian Algorithm

T.Sajana[1]*, M.R.Narasingarao[2]

[1]Research Scholar, Department of CSE, K L E F, Vaddeswaram, Guntur, A.P.
[2]Professor, Department of CSE, K L E F, Vaddeswaram, Guntur, A.P.
**Corresponding author E-mail:*sajana.cse@kluniversity.in*

Abstract

Malaria disease is one whose presence is rampant in semi urban and non-urban areas especially resource poor developing countries. It is quite evident from the datasets like malaria, dengue, etc., where there is always a possibility of having more negative patients (non-occurrence of the disease) compared to patients suffering from disease (positive cases). Developing a model based decision support system with such unbalanced datasets is a cause of concern and it is indeed necessary to have a model predicting the disease quite accurately. Classification of imbalanced malaria disease data become a crucial task in medical application domain because most of the conventional machine learning algorithms are showing very poor performance to classify whether a patient is affected by malaria disease or not. In imbalanced data, majority (unaffected) class samples are dominates the minority (affected) class samples leading to class imbalance. To overcome the nature of class imbalance problem, balancing the data samples is the best solution which produces the better accuracy in classification of minority samples. The aim of this research is to propose a comparative study on classifying the imbalanced malaria disease data using Naive Bayesian classifier in different environments like weka and using an R-language. We present here**,** clinical descriptive study on 165 patients of different age group people collected at medical wards of Narasaraopet from 2014-17. Synthetic Minority Oversampling Technique (SMOTE) technique has been used to balance the class distribution and then we performed a comparative study on the dataset using Naïve Bayesian algorithm in various platforms. Out of balanced class distribution data, 70% data was given to train the Naive Bayesian algorithm and the rest of the data was used for testing the model for both weka and R programming environments. Experimental results have indicated that, classification of malaria disease data in weka environment has highest accuracy of 88.5% than the Naive Bayesian algorithm accuracy of 87.5% using R programming language. The impact of vector borne disease is very high in medical applications. Prediction of disease like malaria is an hour of the need and this is possible only with a suitable model for a given dataset. Hence, we have developed a model with Naive Bayesian algorithm is used for current research.

Keywords: *Imbalanced Data; Malaria; Naïve Bayesian; Weka; R.*

1. Introduction

Malaria disease is one of the elevating issues among the vector borne diseases in medical domain [1]. It becomes one of the global health problem caused by a mosquito bite [1] [2]. Malaria, which is a vector borne disease, affects the rural population for many years. Even though people maintain with healthy life style with good food habits and with neat surroundings, still due to the climate changes or for any other reason, many people are affected by malaria disease [1][3].

According to 2016 report of World Health Organization (WHO), one million people are dying annually due to the vector borne diseases like malaria [4]. Irrespective of the age factor, Millions of deaths have occurred which is estimated to be 839000 in 2000 (range: 653000-1.1 million) to 438000 in 2015 (range: 236000-635000) i.e., 48% have been recorded. Overall, it is estimated that, the incessant presence of Malaria disease has decreased the world population by 60% [5].

One way to solve this problem is to properly identify the patient with a disease and providing accurate diagnosis. Malaria a vector borne disease, the data of which always creates a class imbalance problem because of the presence of samples of negative class (unaffected patients) dominating the samples of positive class (affected patients) [6] [7]. Handling such a kind of imbalanced data, is a critical task because prediction of a patient with or without a disease becomes an important problem in the medical scenario [6] [7] [8].

Unbalanced data sets are include not only medical domain but also domains like credit card fraud detection[9], detection of problems in software's[9][10],detection of oil spills in satellite radar images[11][12], frauds in tele communications[9], detection of frauds in financial sector[9] etc. Analyzing an imbalanced data set is very pertinent in different diseases like malaria [6]. Handling such a complex nature of data, becomes an issue in Data Mining and Machine Learning domains [6] and it is observed that, most of the traditional machine learning algorithms is very sensitive with imbalanced data[13][14]. In general, the main goal of machine learning algorithms is to achieve good accuracy for classification/prediction. While dealing with imbalanced data, the classifier predicts towards the majority class rather than minority class because the classifier will be trained with many majority class examples rather than minority class examples and hence the prediction of the classifier will be more oriented towards majority class examples. So, accurate classification of minority class samples is important than majority class samples especially in medical diagnosis [6]. For example,

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

misclassification of an affected patient is more dangerous than unaffected patient [13] [14].

The organization of rest of the paper as follows: Section 2 describes the literature review of various methods of classification of malaria disease data and imbalanced data applications. Section 3 presents the experimental frame work for classification of imbalanced malaria disease data using Naive Bayesian algorithm on different platforms like Weka and R Programming and Section 4 shows the experimental outcomes.

2. Literature Survey

Early stage identification of malaria disease is very important for a disease affected patient; otherwise, sometimes it may even leads to death also. Many researchers developed various methods, which were mainly based on conventional machine learning algorithms that are producing accurate results with balanced data as follows:

- Purnima Pundit et al. [15] analyzed Digital Holographic Interferometric Microscopic (DHM) image for detection of infected erythrocytes using Artificial Neural Networks.
- Francis Bbosa et al. [16] predicted the status of malaria patients based on Rule based classification.
- Chunqing Wu et al. [17] developed a Disease free equilibrium model with Halanay inequalities for discrete time dynamic system of Neural Networks for diagnosing malaria disease.
- Meng-Hsim Tsai et al. [18] proposed a method for detection of malaria parasite using Adaptive histogram, Threshold segmentation and K-means clustering techniques.
- Farah Zakiyah et al. [19] examined the Classification of Plasmodium Vivax parasite using image processing and KNN techniques.
- Kshipra C.Charpe et al. [20] investigated malaria parasite stages using Image processing and classification techniques.
- J.Somasekar et al. [21] examined a method for detection of effected erythrocytes based on Adaptive median filter, edge enhancement and Fuzzy C-Means clustering techniques.
- Bruno B Andrade et al. [8] investigated active and inactive malaria disease patients based on laboratory assessment.

All the above stated survey, deals with balanced malaria disease data only. But, identification of an affected patient (positive class sample) in imbalanced malaria disease data is very important [6] [7]. Because of the skewed distribution of data, classifying the positive class samples from the past few years became a challenging issue to many researchers which leads to class imbalance problem in many fields [9].Connecting to the context of class imbalance problem, a plethora of methods is proposed by many researchers to classify the imbalanced data sets on different medical data sets which are stated below:

- Guo Haixiang et al. presented a plethora of methods for handling imbalanced data sets in the era of many application domains [1].
- Salma Jamal developed a Predictive model of anti-malarial molecules inhibiting apicoplast formation for an imbalanced malaria disease dataset using Cost sensitive Naïve Bayesian, Random Forest and J48 algorithms and found that Random Forest algorithm produces better accuracy[6].
- Bruno B Andrade et al. suggested that severe stage of malaria disease causes to reducing of inflammatory cytokines which is a high level imbalance class problem in medical domain [8].
- Rashmi Dubey developed an ensemble frame work of sampling and feature selection methods for classification of Alzheimer's disease which occurred in elderly patients [22].
- Wing W. Y. Ng et al. proposed a Diversified Sensitivity-Based Under sampling method – DSUS on different datasets and stated that DSUS is an efficient method when compared with other under sampling techniques [23].
- Yazan F et al. Conducted a comparative study on 2080 cardiac patients for classifying the patients status using a set of classifiers like C4.5, logistic regression, discriminant analysis, Support Vector Machine (SVM) and Classification and Regression Tree (CART) and found that CART and SVM are has highest precision in diagnosing the disease[24].
- Jia Pengfei et al. proposed a new sampling method called Distinct –Borderline algorithm for balancing the class distribution of imbalanced data sets and [25]
- N.Poolsawad et al. Investigated comparative study between set of classifiers like Multilayer Perceptron (MLP), RBFN, SVM, Decision Tree and Random forest algorithms on LIFELAB dataset and found that RF algorithm produces better accuracy [26].
- V.Garcia et al. suggested various methods for handling of imbalanced data sets [27].
- Jaree Thongkam et al. developed a model for prediction of breast cancer imbalanced data set using C-Support vector machine algorithm [28].
- Xing-Ming Zhao et al. suggested ensemble classifiers as 1NN (Nearest Neighbor), SVM (Support Vector Machine), C4.5 and Logistic regression algorithms for protein classification and conducted a two ways of comparative study. One with balanced data and another with unbalanced data and individual classifiers with ensemble classifiers and concluded that balanced data is good for accurate prediction of samples while ensemble classifiers are best predictors among all the stated classifiers [29].

3. Experimental Framework

Dataset description: we have **c**ollected 165 patients data from the medical wards of Narasaraopet in which it consists of various attributes as listed below:

Attributes	Actual range
Age	1-20 yrs
Haemoglobin	M13.0-18.0/F11.0-16.5 grms%
RBC	3.80-5.80 millions/cumm
Hct	35.0-50.0 %
Mcv	80-97 fl
Mch	26.5-33.5 pg/cells
Mchc	31.5-35.0 %
Platelets	1.50-3.90 lakhs/cumm
WBC	3,500-10,000 cells/cumm
Granuls	43.0-76.0 %
Lymphocytes	17.0-48.0 %
Monocytes	4.0-10.0 %
Malaria	Positive/Negative

Consider the class Imbalance Distribution of Malaria Disease Dataset as follows:

Total No of instances	165
No of Attributes	13
Minority Class samples	5
Majority Class samples	160
Minority Class	Positive
Majority Class	Negative
% of Class(Minority, Majority)	3.1 , 96.9
Imbalance Ratio	1: 32

The best method for classifying the samples of imbalanced data is a two-step process as follows: Initially it is necessary to balance the class distribution and then apply the classifier to perform classification of samples.

Step-1: Balancing the class distribution: - In order to balance the class distribution of the data, researchers have compared under sampling method with over sampling and found that it is always a better to. It is an oversampling technique in which minority class samples are oversampled in a synthetic manner.

Consider the algorithm of SMOTE:

Algorithm: SMOTE (T, N, k)

Input: Number of minority class samples T; Amount of SMOTE N%; Number of nearest neighbor's k.

Output: (N/100)* T synthetic minority class samples.

Method: (∗If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd. ∗)

If N<100

Then randomize the T minority class samples
T = (N/100) ∗T
N = 100
End if
N = (int) (N/100) (∗ the amount of SMOTE is assumed to be in integral multiples of 100. ∗)
k = Number of nearest neighbors, numattrs = Number of attributes, Sample [ ] [ ]: array for original minority class samples, new index: keeps a count of number of synthetic samples generated, initialized to '0', Synthetic [ ] [ ]: array for synthetic samples. (∗ Compute k nearest neighbors for each minority class sample only. ∗)
For i ←1 to T
Compute k nearest neighbors for i, and save the indices in the narray
Generate (N, i, narray)
End for
Generate (N, i, narray) (∗ Function to generate the synthetic samples. ∗)
While N =0 (Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.)
For attr ←1 to numattrs
Compute:
Dif= Sample [narray [nn]] [attr] −Sample[i] [attr]
Compute: gap = random number between 0 and 1
Synthetic [new index] [attr] =Sample[i] [attr] +gap∗dif
End for
New index++, N = N −1
Endwhile
Return (∗ End of generate ∗)
End of algorithm.

Step-2: Classification – Unlike traditional statistical methods, the machine learning models provide dynamic output as further data is fed into it. It doesn't assume the skills and insight needed to analyses the disease using statistical techniques.

It is recognized that translational research is essential to extract knowledge in diagnosing the disease. In fact, the development of various machine learning techniques/models are going to be a practical tool to predict disease like malaria, which is very much rampant in nonurban areas/remote villages. The best part of machine learning techniques is that, as more data is generated, the model improves in precision and can be widely employed in different clinical settings.

For the classification of samples, we proposed machine learning technique called Naïve Bayesian algorithm which uses class conditional probabilities for classification of samples and conducted comparative study in both Weka and R programming environments.

Consider the Naive Bayesian classification algorithm:
Algorithm: classification of tuples for a given dataset D.
Input: A dataset D consists of training tuples.
Output: prediction of a class for a given tuple.
Method:
D: A training set of tuples
X = (x_1, x_2, x_3…x_m) is an n-dimensional vector each vector tuple of D defined over 'n' attributes of A_1,A_2,A_3,……A_n.
C_1, C2, Cm: set of classes
For each tuple of 'X'
Predict the highest posterior probability that a tuple 'X' be longs to class 'C_i' if and only if,

$$P(C_i|X) > P(C_j|X) \text{ for } 1 \le j \le m, j! = i$$

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

By Bayesian theorem calculate maximized posterior hypothesis,

Where P(X) is constant for all classes.
Find the maximize P (X| C_i). P (C_i)
If the class prior probabilities are not known,
Then P (C_1) = P (C_2) =……P (C_m) and maximize the P(X/C_i).
Else
Maximize P(X C_i) P (C_i) where P (C_i) = |C_ad|/|D|
[|C_i, D|] is the number of training tuples of class C_i in D]

If the dataset 'D' contains many number of attributes,
Then find the class-conditional independence to compute P (X|C_i).
i.e, P (X|C_i) =
P (x_1|C_i)* P (x_2|C_i)* P (x_3|C_i)* ……*P (x_n |C_i)
Where x_k is a value of attribute A_K for tuple X.
If A_K is categorical, then
P (x_k C_i) = [|C_i, D| having the value x_k for A_K] / |C_i, D|.
If A_K is continuous valued, then
P (x_k, C_i) =g (x_k, μc_i, σc_i) where g(x, μ, σ) is defined with

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

[μc_i or μ is mean or average and σc_i or σ is the standard deviation of the values of attribute A_K]
Predict the class of X based on class C_i if and only
If P (X | C_i)*P (C_i) > P(X C_j) P (C_j) for 1≤j≤m, j≠i.
End of algorithm.

4. Results

Conducted experiments on collected imbalanced malaria disease data set i.e., initially performed class balance distribution using SMOTE algorithm later performed a comparative study of Naive Bayesian Classifier on different platforms. Consider the Imbalanced Malaria data set as shown in figure 1:

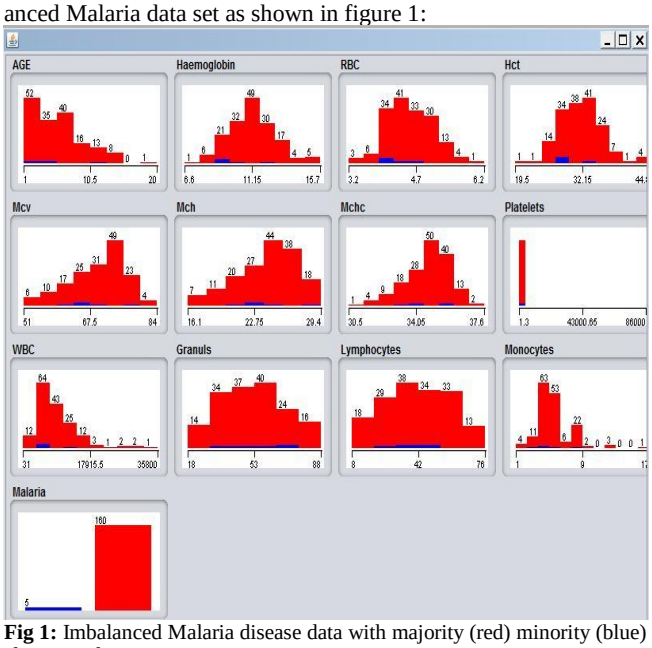


Fig 1: Imbalanced Malaria disease data with majority (red) minority (blue) class samples.

Later applied SMOTE algorithm to obtain balanced Malaria dataset as shown in Figure 2.

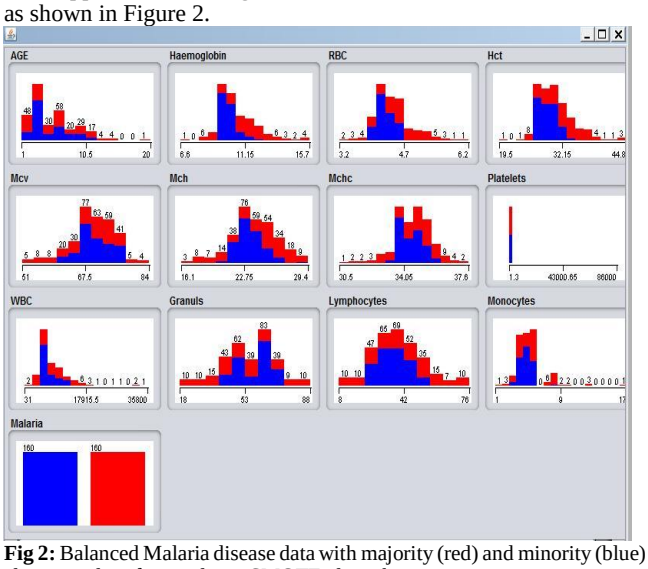


Fig 2: Balanced Malaria disease data with majority (red) and minority (blue) class samples after applying SMOTE algorithm

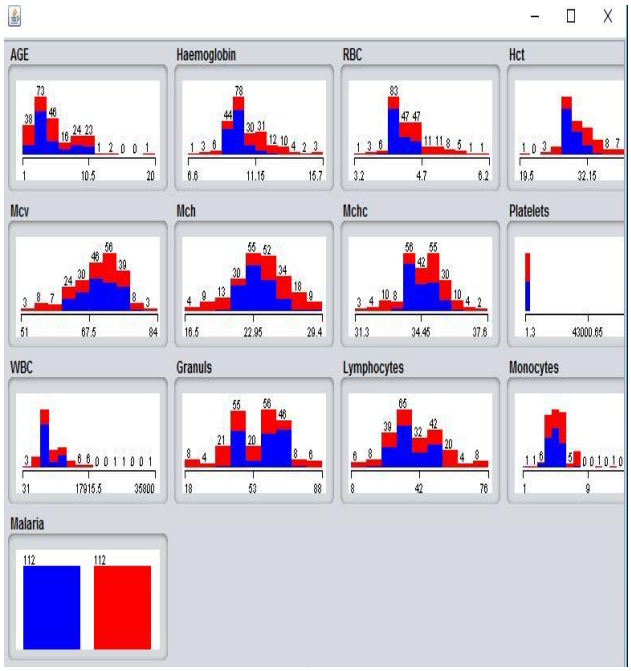


Fig 3: Training data of Naïve Bayesian algorithm

Trained data & Test data after balancing: Among the balanced class distribution 70 % - 112 patient's data is taken to train the Naive Bayesian algorithm as shown in Fig 3 and 30% - 48 patients' data is considered to test the Naive Bayesian classifier performance as shown in Fig 4 as follows:

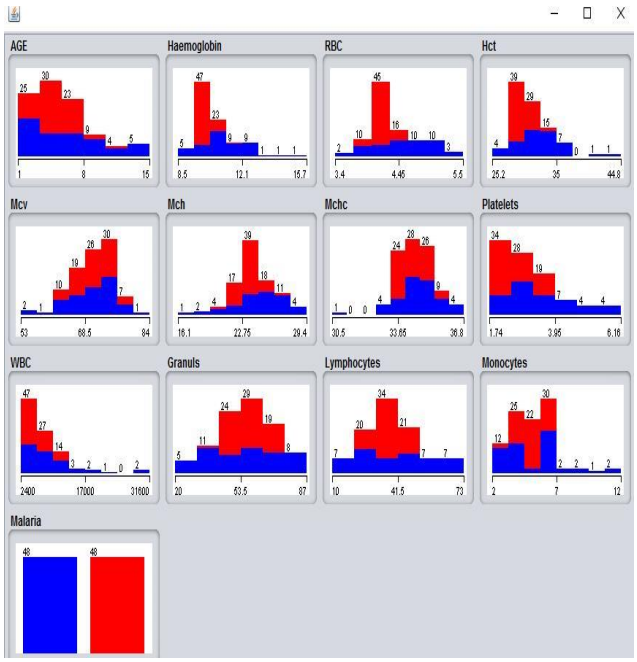


Fig 4: Test data of Naïve Bayesian algorithm

Performance of Naive Bayesian classifier on classification of malaria disease data on different platforms as follows:

Table 1: Comparison study of Naive Bayesian classifier performance on malaria disease data on different platforms

	R– programming	Weka tool
Correctly Classified samples	84	85
Incorrectly Classified samples	12	11
Accuracy %	87.5	88.5
Misclassification %	12.5	11.5

Where

Accuracy = [TP + TN] / [TP + FN + FP + TN]

Which is defined from the Confusion matrix as follows:

Table 2: Confusion Matrix

	Predicted Positive Class	Predicted Negative Class
Actual Positive Class	TP	FN
Actual Negative Class	FP	TN

Consider the Classification performance of Naïve Bayesian algorithm on malaria disease data on different platforms as shown below:

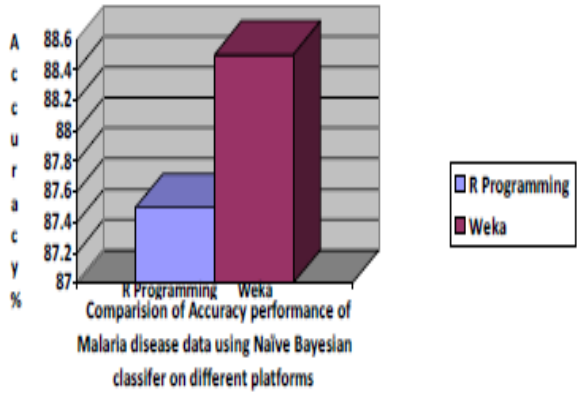


Fig 5: Naïve Bayesian Classifier performance on Malaria disease data set using R programming and Weka environments.

From the Figure-5, Naive Bayesian algorithm classifies the samples with an accuracy of 88.5% in Weka environment where as in R programming the Naive Bayesian algorithms performs 87.5% accuracy for classifying the malaria disease data.

5. Discussions

We conducted a prospective study on clinically collected 165 patient's data which is a skewed distribution data. Applied SMOTE algorithm to balance the class distribution and then conducted a comparison study of Naive Bayesian classifier in both Weka and R programming environments. Experimental results stated that Naive Bayesian classifier in weka environment performed very well for the classification of imbalanced malaria disease data when compared with R programming environment.

6. Conclusion

Malaria disease becomes one of the class imbalance problems in medical diagnosis. If we are examined in a nutshell, classification of such imbalanced data especially minority class samples is a key thing. So accurate prediction of effected patient and diagnosing within time is very important otherwise it may lead to death also. Many conventional classifiers are voted for classification of majority class samples only. So handling minority class samples and its classification is becoming a burning issue in medical field. We present the balanced class distribution of imbalanced malaria disease data by using SMOTE algorithm and then performed a comparative study Naive Bayesian Classifier on different platforms for better prediction system.

References

[1] Thanh Quang Bui and Hai Minh Pham (2016). Web- based GIS for spatial pattern detection: application to malaria incidence in Vietnam. Bui and Pham Springer plus 5: 1014: 1-14.

[2] S.T. Khot and R.K. Prasad (2015). Optimal Computer Based Analysis for Detecting Malarial Parasites. Proc. of the 3rd Int.

Conf. on Front. of Intell. Comput. (FICTA) Advances in Intelligent Systems and Computing 327 vol 1: 69-80.
[3] Md Z Rahman, Leon id Roytman et al. (2015). Environmental Data Analysis and Remote Sensing for Early Detection of Dengue and Malaria. Proc. of SPIE Vol. 9112: 1-9.
[4] WHO Malaria Report (2016)
http://www.who.int/mediacentre/factsheets/fs387/en/
[5] World Malaria Report (2015) Pages-x, xi.
http://apps.who.int/iris/bitstream/ 10665/200018/ 1/ 9789241565158 _eng.pdf
[6] Salma Jamal, Vinita Periwal et al. (2013).Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. BMC Bioinformatics 2013, 14:55, 1-8.
[7] Tsige Ketema and Ketema Bacha. (2013). Plasmodium vivax associated severe malaria complications among children in some malaria endemic areas of Ethiopia. BMC Public Health 2013, 13:637, 1-7.
[8] Bruno B Andrade, Antonio Reis-Filho et al. (2010). Severe Plasmodium vivax malaria exhibits marked inflammatory imbalance. Malaria Journal 2010, 9:13, 1-8.
[9] Guo Haixiang, Li Yijing, et al. (2016). Learning from class-imbalanced data: Review of methods and applications. Expert systems with applications: 1-49.
[10] Bartosz Krawczyk (2016). Learning from imbalanced data: open challenges and future directions. Prog Artif Intell: 1-12.
[11] Xiaoheng Deng, Weijian Zhong et al. (2016). An Imbalanced Data Classification Method Based On Automatic Clustering Under-Sampling. IEEE transaction: 1-8.
[12] Aida Ali, Siti Mariyam Shamsuddin et al. (2013). Classification with class imbalance problem: a review. International journal of Advances in Soft Computing and its Applications 5(3): 1-30.
[13] N. Poolsawad, C. Kambhampati et al. (2014). Balancing Class for Performance of Classification with a Clinical Dataset. Proceedings of the World Congress on engineering vol.1: 1-6.
[14] M. Mostafizur Rahman and D. N. Davis (2013). Addressing the Class Imbalance Problem in Medical Datasets. International Journal of Machine Learning and Computing 3(2): 224-228.
[15] Purnima Pandit, A. Anand. Artificial Neural Networks for Detection of Malaria in RBCs. 2016 AUG; arXiv: 1608.06627.
[16] Francis Bbosa, Ronald Wesonga et al. (2016). Clinical malaria diagnosis: rule-based Classification statistical prototype.Springer Plus. 5:939.
[17] Chunqing Wu and Patricia JY Wong (2016). Multi-dimensional discrete Halanay inequalities and the global stability of the disease free equilibrium of a discrete delayed malaria model. Advances in Difference Equations. 2016:113.
[18] Meng-Hsiun Tsai, Shyr-Shen Yu et.al. (2015). Blood Smear Image Based Malaria Parasite and Infected-Erythrocyte Detection and Segmentation. TRANSACTIONAL PROCESSING SYSTEMS. J Med Syst 39: 118. DOI 10.1007/s10916-015-0280-9.
[19] Farah Zakiyah Rahmanti, Sutojo et al. (2015). Plasmodium Vivax Classification from Digitalization Microscopic Thick Blood Film Using Combination of Second Order Statistical Feature Extraction and K-Nearest Neighbour (K-NN) Classifier Method. IEEE 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME) Bandung.2015 November.2-3.
[20] Kshipra C. Charpe, Dr. V. K. Bairagi et al. (2015). Automated Malaria Parasite and there Stage Detection in Microscopic Blood Images. IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO).
[21] J. Somasekar, B. Eswara Reddy. (2015). Segmentation of erythrocytes infected with malaria parasites for the diagnosis using microscopy imaging. Elsevier - Computers and Electrical Engineering. 336–51.
[22] Rashmi Dubey, Jiayu Zhou et al. (2014). Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study. Elsevier Neuro Image 87: 220–241.
[23] Wing W. Y. Ng, Junjie Hu et al. (2015). Diversified Sensitivity-Based Under sampling for Imbalance Classification Problems. IEEE TRANSACTIONS ON CYBERNETICS: 1-11.
[24] Yazan F, Roumani et al. (2013). Classifying highly imbalanced ICU data. Health care Manag Sci 16:119- 128.
[25] Jia Pengfei, Zhang Chunkai et al. (2014). A New Sampling Approach for classification of Imbalanced Data sets with High Density. IEEE transaction: 217-222.
[26] N. Poolsawad, C. Kambhampati et al. (2014). Balancing Class for Performance of Classification with a Clinical Dataset. Proceedings of the World Congress on engineering vol 1: 1-6.
[27] V.Garcia, J.S.Sanchez et al. (2012). On the Effectiveness of preprocessing methods when dealing with different levels of class imbalance. Elsevier –Knowledge Based Systems.13-21.
[28] Jaree Thongkam, Gauandong Xu et al. (2009). Toward breast cancer survivability prediction model through improving training space. Elsevier – Expert systems with Applications.12200-12209.
[29] Xing-Ming Zhao, Xin Li et al. (2007). Protein classification with imbalanced data. Wiley InterScience.1125-1132.