International Journal of ChemTech Research
CODEN (USA): IJCRGG, ISSN: 0974-4290, ISSN(Online):2455-9555
Vol.10 No.5, pp 590-594, 2017

Resolving Class Imbalance in Meteorological Datasets for Predicting Dengue Outbreak

Vivek Jagadeesan Sharavanan, Arun venkatesh Aapakudal Venkataraman, Muthusaravanan Sivaramakrishnan and Ram Kothandan*

Department of Biotechnology, Kumaraguru College of technology, Coimbatore, Tamilnadu, India.

Abstract : The vector-borne diseases are highly contagious and responsible for high rate of mortality especially in tropical and subtropical region. The rapid urbanization and change in climatic pattern restricts the elimination of vector-related diseases. The weather has a direct effect on outbreak of vector-borne diseases. In this study, an attempt has been made to predict prior occurrence of dengue infection in North Tamilnadu region by constructing a machine learning based model. The training set utilized in this study was collected manually using text mining approaches from various meteorological sources. The major pitfall during the data preparation is the class imbalance problem that exists with the outbreak of disease in the various monsoon seasons. In this present study, we employed cost-sensitive based approach to overcome the class imbalance in the meteorological dataset. Based on the prediction obtained, it was observed that both feature ranking and cost-sensitive method improved the prediction performance. Prediction in performance obtained in the study was based on F-measure, MCC and ROC value.

Keywords : vector-borne diseases, meteorological data, machine learning, class imbalance, cost sensitive.

Introduction:

Dengue is a mosquito borne viral infection, transmitted to a greater extent by the species Aedes aegypti and to a lesser extent by Aedes albopictus. This disease is mostly prevalent in tropical and subtropical regions of the world. It is estimated that around 3.9 billion people in 128 countries are exposed to the risk of infection. The improper drainage system and poor sanitary conditions in developing countries makes it an ideal habitat for the mosquitoes to breed in. For this reason, dengue is quite common in urban areas than in rural areas [1]

India is one of the major country present in the dengue zone. High population density in urban centres, poor sanitary conditions, improper drainage system, tropical climate, poor quality water storage facility makes the situation easier for dengue outbreak to take place and threatens health care system to act on [2] Vector density and distribution determines the outbreak pattern of the disease. Vector population depends on various meteorological features such as rainfall, temperature, humidity, altitude etc. The lifespan of the larvae is increased when it is incubated under optimum tropical climate. The rate of infection with dengue virus is magnified to a great extent by the ability of the vector to breed in stagnant\stored water in and around house. This would provide sufficient information on the region of outbreak and time for the health department to take preventive measures [3,4]

Tamil Nadu is strategically located near the equator and it receives rainfall mostly from the monsoons [5]. This makes it an ideal location for vector breeding. This is evident from the fact that Northern Tamil Nadu has experienced frequent epidemic outbreaks in the past. Chennai being the state capital, with its expanding radius and placed in Northern Tamil Nadu has been chosen as our region of study. The need of the hour is an early warning system which is able to precisely locate the area of outbreak and help in taking preventive measures to control vector breeding.

In order to overcome this problem, we collected the meteorological data for the Chennai region from commercial weather service provider weather underground [6] .Meteorological data contained various features\parameters in which few parameters such as precipitation, temperature, humidity was extracted and pre-processed for obtaining accuracy in predictions. For the purpose of the study, we considered precipitation data as a major feature since incidence of monsoon is the key for dengue outbreak. Pre-processing of data was initially done to overcome missing data and to remove data redundancy. Data imbalance in the dataset occurred due to the very few data represented event of rainfall - which is considered as a minority class when compared to the data representing event of no rainfall (majority class).

The classical classifiers are used for complete/balanced datasets and are not used for incomplete/imbalanced datasets, because the decision making is favoured to majority class. Moreover, the presence of imbalanced dataset reduces the performance of machine learning techniques, as the accuracy and decision making by preconception to the majority class, which results misclassification of minority class samples or furthermore considering them as meaningless data [7].

Class imbalance issue can be dealt in three different approaches. They are algorithmic approach, data-pre-processing and feature selection approach. In data-pre-processing method, the data is resized by increasing or decreasing the data instances. Selection of features that allow the classifier to reach maximum performance is the feature selection approach. Algorithmic method involves the use of modifying algorithm that reduces the class imbalance issue. One among them is cost sensitive classification. Cost sensitive method works by assigning cost to the training set as most of the classifiers tend to neglect the difference between types of misclassification errors. Hence, we have undertaken the cost sensitive approach in our work to obtain more accurate predictions [8 9]. By standardizing the cost of the misclassification, the accuracy of the model obtained was about 97%.

Methodology

Dataset Preparation:

The dataset for our work was extracted from weather underground [6], a commercial weather service providing company, based in US. Complete data set containing all meteorological features such as rainfall, temperature, humidity etc. was obtained for the region of study. The dataset contains complete meteorological data for about 19 years from 1997 to 2015. The total instances collected from the wunderground were 6820. Model construction and data analysis was carried out in Waikato Environment for Knowledge Analysis (WEKA) [10]. Dataset was pre-processed prior to model construction to avoid missing values and redundancy in the dataset.

Feature Selection:

A series of 8 attributes were extracted from the present in meteorological data and were employed in the model construction. The eight attributes are minimum temperature, mean temperature, max temperature, minimum humidity, mean humidity, maximum humidity and precipitation (mm) [4]. In this present study, dengue outbreak was considered to be our main objective; hence, to predict the desired outcome, we segregated our dataset into days with rainfall and days with no rainfall recorded in a year. The days with rainfall were considered to be positive class and the day without rainfall were considered as negative class.

Learning Algorithm:

The number of positive instances is just 270 and the number of negative instances is 6550, thus there is a huge difference margin or imbalance between the datasets. When processed, these values gave rise to

erroneous results and also damaged the performance of the classifier, making it a biased one. To overcome class imbalance, resampling the dataset or the algorithmic approach can be employed. In this study, a comparison of the oversampling and algorithmic approach was also done to obtain a suitable model. Under sampling method was neglected considering the size of the dataset. In oversampling technique, we employed Synthetic Minority Oversampling Techniques (SMOTE). The number of instances of positive data were increased by applying SMOTE by keeping the number of nearest neighbour as 5. In SMOTE the samples were not replaced but new synthetic samples were created. For example, if the number of instances is to be oversampled to 200% then for each 5 nearest neighbour 2 are considered and one new sample is created in each direction. Thus, it makes the decision region of the positive class to be more general and comparable to negative class and allows the classifier to improve its accuracy. The samples were oversampled to 100%, 200%, 300%, 400%, 500% and were subjected to 10-fold cross validation. The number of positive instances were increased to 4320. Thus, reducing the class imbalance greatly [7,11].

On the other hand, in algorithmic approach cost sensitive classification algorithm was used. The main aim of this method is to reduce the misclassification error by modifying the cost in the cost sensitive matrix [8]. We constructed a 2×2 cost matrix to reweight the instances. The Cost for the correctly classified instances are assumed zero (i.e., the cost associated with the True Positive (TP) and True Negative (TN) is zero). Cost sensitive classifier is bagged to various base classifier such as LibSVM, J48 (C4.5), Random Forest and Random Tree. We reweighted the training instances by altering the total costs assigned to each class. The costs in the matrix were subjected to standardisation to get optimised results. Cost sensitive classification was processed in default parameter settings.

$$\text{Cost} = \text{FN rate} \times C(0, 1) + \text{FP rate} \times C(1, 0) \qquad (1)$$

In feature selection method, we selected the feature with low dimensionality of data. The feature precipitation was pre-processed in such a way that it had only two values either positive or negative. Thus, reducing the dimensionality of the data greatly[9].

Table 1: Cost sensitive classifier prediction results.

	F-Measure	**MCC**	**ROC**
LibSVM	1	1	1
J48(C4.5)	0.997	0.955	1
Random Forest	0.988	0.838	0.999
Random Tree	0.976	0.68	0.906

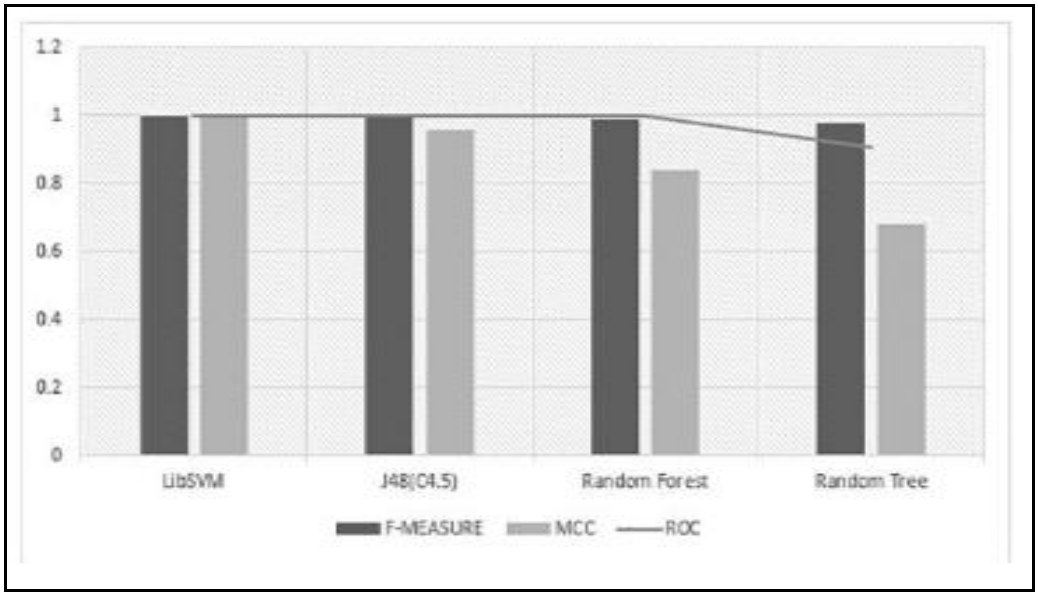

Figure 1: Performance evaluation of algorithm comparison. Performance evaluation was done in terms of F-measure, MCC and ROC.

Performance evaluation:

The Precision and Recall for finding F-measure are obtained from the equations,

Precision= TP/TP+FP (2)
Recall=TP/TP+FN (3)

Results and Discussions:

The main aim of this work is to construct a model to predict prior outbreak of dengue in the Northern Tamil Nadu region particularly Chennai. The work also provides solution to tackle class imbalance in the dataset. The performance of classifiers was usually evaluated using various parameters such as precision, recall, and Mathew's Correlation Coefficient (MCC) and Area under the Receiver Operating Curve (ROC) values**.** However, in the predictive performance of the disease related events (like our work), the parameter precision is best used because it gives the exact measure of the model. Also, a single false prediction in this scenario would lead to tragic events [12].In pre-processing approach, the oversampling of the samples to over 5 times help in bringing the positive instances from just 270 to 4320, which balanced the data to a great extent. However, it was observed that the dataset tends to over fit the model constructed. This was quite noticeable with the prediction performance calculated. The reason could be overgeneralization (i.e.) generation of synthetic samples increases the occurrence of overlapping between classes [13]. In algorithmic approach, cost sensitive classifier were bagged to various base classifiers and resulting data instances yielded an accuracy of over 97%. The value of MCC, F-measure and ROC obtained in all the bagged base classifiers were above 0.5, thus indicating good predictions in result. When comparing both the SMOTE and cost sensitive classifier, performance was good in both the cases, but SMOTE was avoided because of its nature of over fitness. SMOTE overcrowds a particular region instead of increasing the overall instances. This is because new instances are synthesized based on the number of nearest neighbors and it is also due to the number of new instances required per repetition. This is the most possible reason for overfitting nature of the SMOTE [12]

Conclusion:

In the development of this prediction model, we analysed an empirical comparison between two methods-SMOTE and cost sensitive method to overcome the class imbalance problem in the prediction of dengue outbreak. Dealing class imbalance at the data level for predicting disease related outbreak would yield oversampled synthesized instances while cost sensitive method deals with algorithm level and provides steady performance and are more effective than the oversampling method in improving the accuracy of the prediction model. The performance of the cost sensitive classifier can be further strengthened by employing the ranker method where the individual features are evaluated by attribute evaluators. Thus, we conclude that for the prediction for the dengue outbreak with high imbalance in the dataset, cost sensitive method performs better than the oversampling method (SMOTE).

References:

1. WHO, "Global Strategy for Dengue Prevention and Control 2012–2020," *World Heal. Organiszation*, p. 43, 2012.
2. Kashinkunti MD *et al*, "A study of clinical profile of dengue fever in a tertiary care teaching hospital," *Sch. J. Appl. Med. Sci.*, vol. 1, no. 4, pp. 280–282, 2013.
3. Chandran R and Azeez PA, "Outbreak of dengue in Tamil Nadu, India," *Curr. Sci.*, vol. 109, no. 1, pp. 171–176, 2015.
4. Shweta *et al*, "Prediction of Dengue Outbreak using Environmental Factors," pp. 716–719, 2016.
5. IMD, "District wise rainfall information," no. mm, 2013.
6. www.wunderground.com.
7. Tsagalidis E and Evangelidis G, "Meteorological Data Mining: Exploiting Domain Expertise in the Class Imbalance Problem," *7th Int. Conf. Adv. Data Min. Appl.*, no. October, pp. 1–13, 2011.
8. Ling CX and Sheng VS, "Cost-Sensitive Learning and the Class Imbalance Problem," 2008.
9. Longadge R *et al*, "Class imbalance problem in data mining: review," *Int. J. Comput. Sci. Netw.*, vol. 2, no. 1, pp. 83–87, 2013.

10. www.cs-waikato.ac.nz/ml/weka.
11. Chawla NV *et al*, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
12. Kothandan R (2015). Handling class imbalance problem in miRNA dataset associated with cancer. *Bioinformation*, *11*(1), 6–10. https://doi.org/10.6026/97320630011006
13. He H *et al* . ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proc. Int. Jt. Conf. Neural Networks* 1322–1328 (2008). doi:10.1109/IJCNN.2008.4633969

*****