International Journal of Electrical and Computer Engineering (IJECE)
Vol. 7, No. 4, August 2017, pp. 2215~2222
ISSN: 2088-8708, DOI: 10.11591/ijece.v7i4.pp2215-2222

Learning from a Class Imbalanced Public Health Dataset: a Cost-based Comparison of Classifier Performance

Rohini R. Rao[1], Krishnamoorthi Makkithaya[2]
[1]Department of Computer Applications, Manipal Institute of Technology (MIT), Manipal, India
[2]Department of Computer Science & Engineering, Manipal Institute of Technology (MIT), Manipal, India

Article Info

Article history:

Received Jun 14, 2017
Revised May 31, 2017
Accepted Aug 28, 2017

Keyword:

Class imbalance
Classifier accuracy
Cost benefit analysis
Data mining
Healthcare

ABSTRACT

Public health care systems routinely collect health-related data from the population. This data can be analyzed using data mining techniques to find novel, interesting patterns, which could help formulate effective public health policies and interventions. The occurrence of chronic illness is rare in the population and the effect of this class imbalance, on the performance of various classifiers was studied. The objective of this work is to identify the best classifiers for class imbalanced health datasets through a cost-based comparison of classifier performance. The popular, open-source data mining tool WEKA, was used to build a variety of core classifiers as well as classifier ensembles, to evaluate the classifiers' performance. The unequal misclassification costs were represented in a cost matrix, and cost-benefit analysis was also performed. In another experiment, various sampling methods such as under-sampling, over-sampling, and SMOTE was performed to balance the class distribution in the dataset, and the costs were compared. The Bayesian classifiers performed well with a high recall, low number of false negatives and were not affected by the class imbalance. Results confirm that total cost of Bayesian classifiers can be further reduced using cost-sensitive learning methods. Classifiers built using the random under-sampled dataset showed a dramatic drop in costs and high classification accuracy.

Corresponding Author:

Rohini R. Rao,
Department of Computer Applications,
Manipal Institute of Technology, Manipal--576104, India.
Email: rohini.rao@manipal.edu

1. INTRODUCTION

All over the world, public health organizations are currently facing the challenge of tackling chronic diseases. Public health agencies need to respond with cost-effective, evidence-based interventions to promote healthy living and prevent chronic diseases. Public Health Organizations routinely collect data on demographics, socio-economic status, living conditions, and underlying health conditions in the population. Data mining techniques can be applied to this population-based data, to gain new insights into the underlying health problems. In the medical diagnosis domain, classifiers have been built to predict diseases such as breast cancer, insomnia, thyroid disease, Parkinson's disease and even recommend medication [1-6]. Pollettini et al. [7] propose a classifier which automatically classifies patients into surveillance levels based on associations among patient features and health. Classifiers have also been designed to predict the cost of healthcare services, to predict intensive care unit readmission, mortality rate and life expectancy rate [1], [8]. Sensor based, unobtrusive, continuous home monitoring systems have been deployed, and human activity is being assessed using classifiers [3]. In the public health domain, classification techniques can be used to analyze the effect of various social and environmental factors, such as work environment, living conditions, education on the health of the population. The relationship between patient features and diseases could help

formulate effective public health interventions [8]. For instance, Takeda et al. [9] utilized multiple logistic analyses to report the significant associations between mental health and psychosocial stressors like family relationship, pregnancy, and income.

However, these public health datasets often suffer from "rare cases" or "rare classes" problem, which result in imbalanced classes in the training datasets [10]. For example, most heath datasets usually have very few cases of the target disease, when compared to the number of healthy patients in the dataset [11-14]. In the binary classification for medical diagnosis, the rare minority class refers to the positive instances or the target class, whereas the majority class is represented by the negative instances in the dataset. Class imbalance can also occur when the data collection process is limited, resulting in artificial imbalances. To be classified as a class-imbalanced dataset, rarity should be between 0.1 to 10%.

This paper considers a health based dataset which records the presence of chronic diseases i.e. diabetes, heart disease and hypertension in the population. Patient demographics, living conditions, socio-economic status are also recorded in the dataset. The authors attempt to build classifiers to predict the occurrence of any of the three chronic diseases in the population. Literature review indicates that there is no single classifier method which yields the best result for all types of class imbalanced training datasets [1]. The amount of the class-imbalance bias depends on factors such as the classification method, the number of attributes in the dataset and the sample size [15]. The motivation for this work is to study the effect of class imbalance on the various classifiers for the health dataset. The objective of this work is to identify the best classifiers for the class imbalanced health dataset from a cost-based comparison of classifier performance. This work is relevant to public health policy makers, who can use the classifiers to augment medical prognosis in the population and also identify the underlying patient features that are correlated with chronic diseases.

1.1. Class Imbalance Problem.

This class imbalance problem is a challenge to classifier techniques because a normal classifier aims to improve overall classifier accuracy. Consider a two-class binary classifier for a health dataset, in which the outcomes are labeled as positive (P) or negative (N). The classifier accuracy can be computed by applying the classifier to a test dataset and comparing the classifier result with actual class labels. There are four possible outcomes, if the predicted value is P and the actual value is also P, then it is a true positive (TP). If the actual value is N and it is predicted as P, then it results in a false positive (FP). Conversely, when both the prediction outcome and the actual value are N, it indicates a true negative (TN). False negative (FN) occurs when the prediction outcome is N while the actual value is P. The class imbalance problem affects different classifiers in a variety of ways, for instance in decision tree induction algorithms, it results in smaller disjuncts [11]. The classifiers trained on class-imbalanced data are usually biased towards the majority negative class, and the accuracy of predictions for the minority target class is very poor. The class-imbalance phenomenon often produces classifiers that have a poor predictive recall, particularly when the positive label is the minority target class [4]. The problem of imbalanced data is also associated with asymmetric costs of misclassifying elements of different classes. For example, consider a binary classifier built for medical diagnosis, the cost of misdiagnosing a health patient as having a health condition (false positive) is less than the cost of falsely diagnosing a sick patient as a healthy person (false negative). The FP case could lead to more diagnostic tests until the patient is diagnosed as healthy. The cost of FN error could result in delayed diagnosis, and ultimately, the loss of life. Therefore, in medical diagnosis based classifiers, the cost of FN is more significant that the FP cost. While the FP cost can be calculated as the expenses incurred in further testing, the cost of FN's hard to quantify.

1.2. Learning from class Imbalanced Datasets

There are two broad approaches to finding effective classifiers in the class-imbalanced datasets: the algorithm specific approach and the data pre-processing approach. In the algorithm specific approach, the classifier methods that are known to work effectively in the class-imbalanced datasets can be used with no modification of datasets. For example, Weiss [12] advocates the use of instance-based learning methods like k-Nearest Neighbors or Support Vector Machines, to predict the minority class effectively. They found that independent of the training size, linearly separable domains are not sensitive to imbalance. He also concludes that non-greedy search techniques used in the Genetic algorithm make it more suitable for dealing with the class imbalance. In the decision tree algorithm, he suggests that splitting rules can be modified to ensure that both classes are addressed. Japkowicz [13] concluded that the MultiLayer Perceptron based classifiers are not sensitive to class imbalance. Kernel-based support vector machine classifiers, clustering and utilizing densities to estimate target class membership are known to work well in the class-imbalanced datasets [16]. Simulation studies show that the class-imbalance and high dimensionality impact the performance of classifiers like a k-nearest neighbor, diagonal linear discriminant analysis, random forests and support vector

machines with linear kernel basis [15]. Cost-sensitive learning methods modify existing algorithms using the cost information. For instance, in a tree based classifier, the cost can be used to choose the splitting attribute or manipulate the weight of training records [17].

There are also specific algorithms such as Two phase rule induction, CREDOS or the one-class classifiers which are proved to be useful in classifying rare cases in training datasets [11], [16], [18-20]. The main idea of these classification methods is that the algorithm should concentrate on the instances that are difficult to learn. [11], [12]. Hempstalk et al. [16] use a combined density estimation with class probability estimation for the purpose of one-class classification. The PNrule algorithm uses a two-phase rule induction method. While the first phase focuses on recall, in the second phase precision is optimized [18]. CREDOS effectively utilizes "ripple down rules" to learn comparable or better models for a variety of rare classes [20]. In most cases, the cost of misclassification errors are not equal, and a cost-sensitive learning approach is required [19]. In cost sensitive learning methods such as AdaCost and MetaCost, the cost is represented in a matrix, and it is utilized to generate a model with lower cost. Empirically, it is often reported that cost-sensitive learning outperforms random re-sampling [11].

The literature reviewed also indicates that using multiple classifiers in ensembles and aggregating the predictions of multiple classifiers, tend to be more accurate than the core classifiers [19]. Wang et al. [4] discuss an implementation of an ensemble of learning algorithms to recommend medication to diabetic patients. Ensemble methods include bagging, boosting and random forests [17]. Bagging used a majority vote to make more accurate classifications using multiple classifiers. Boosting, on the other hand, uses an adaptive sampling of instances, based on the weights of the instances to improve the performance of the classifiers. Boosting methods like SMOTEBoost and AdaBoost have been found to be effective in the rare case scenario [19]. Blending is an ensemble method where multiple algorithms are prepared on the training data. Meta classifiers combine the predictions of multiple classifiers to make accurate predictions on unseen data.

The data pre-processing approach to the class imbalance problem would be to modify the training dataset itself using various sampling techniques [4], [11-13], [21]. Basic sampling methods include under-sampling to reduce the majority class instances or over-sampling wherein the minority class instances are increased to match the number of majority class instances. The Synthetic Minority Over-sampling Technique (SMOTE) is widely used in the class imbalance problem. SMOTE is an over-sampling approach that creates synthetic minority class samples to match the number of majority class instances. SMOTE is reported to perform better than simple over-sampling. SMOTE is also computationally expensive to implement when compared to sampling methods like random under-sampling [21]. However, other experiments have proved that simple under-sampling tends to outperform SMOTE in most situations [22]. The performance of classifiers implementing SMOTE has been found to vary based on the number of dimensions in the training dataset [22]. Smart re-sampling can be deployed instead of cost-sensitive learning as they can provide new information or eliminate redundant information for the learning algorithm [11]. The disadvantages of sampling are, the random undersampling method can potentially remove certain critical instances, and random over-sampling can lead to over-fitting [11], [12]. The threshold-moving approach to the class imbalance problem does not involve any sampling. Certain classifiers like the Bayesian or decision tree induction, return a probability value along with the class label which can be used to compute a new threshold. In class, balanced datasets the probability threshold is 0.5. In case of class imbalance, the results of the classifier can also be weighted based on costs. In general, threshold moving moves the threshold, so that the rare class tuples are easier to classify. Threshold moving technique is known to reduces the costly FN errors in classifiers used for medical diagnosis.

1.3. Evaluating Classifier Performance

Traditional classification accuracy measures such as accuracy or misclassification rate are not good indicators of classifier accuracy in class-imbalanced datasets [11], [12]. If the target class is very rare, say 0.5%, correctly predicting all instances of the majority class can achieve a very high accuracy level of 99.5%. The accuracy measure of precision and recall are more relevant in the case of class-imbalanced datasets [17]. Precision denotes the fraction of instances that are TPs in the set of all instances predicted as P (TP+FP). Recall measures the fraction of TPs correctly predicted in the set of all actual P instances (TP+FN). Classifiers with high recall have less number of FNs. Hence for rare classes, the classifier should be evaluated based on how it performs on both recall and precision. Usually, in class-imbalanced datasets, the target class has much lower precision and recall than the majority class. Many practitioners have observed that for skewed class distributions the recall of the minority class is often 0, which means that no classification rules have been generated for the target class.

Commonly used graphical display of classifier accuracy include receiver operating characteristic curve (ROC), the precision-recall curve (PRC) and cost curves. For a binary classifier, ROC curve is a

graphical method to graphically represent the trade-off between TP rate and FP rate [11], [14]. A ROC plot provides a single performance measure called the Area under the ROC curve (AUC) score. AUC score is 0.5 for "chance" classifiers, which indicates the lack of any statistical dependence and is equivalent to random guessing and 1.0 for perfect classifiers. The Area under ROC Curve (AUC) can be used to compare the performance of multiple classifiers, but they are not very useful for class-imbalanced datasets. Precision-Recall curves (PRC) are often used instead of ROC plots to represent accuracy in the class-imbalanced datasets [23], [24]. The PRC plot shows precision values for corresponding recall or sensitivity values. While the baseline is fixed with ROC, the baseline of PRC is determined as P / (P + N). The area under the PR curve, denoted as AUC (PRC), is a better indicator for multiple classifier comparisons in the class-imbalanced datasets [14]. The cost curve (CC) is an alternative to the ROC plot, and they analyze classification performance by varying operating points, which are based on class probabilities and misclassification costs [14]. The probability cost function or PCF represents the operating points on the x-axis, and the normalized expected cost or NE[C] accounts for the classification performance on the y-axis.

2.4. Proposed Solution

The authors use the open source WEKA tool to create classifiers using both the algorithmic approach as well as the sampling approach. The authors picked WEKA, because of its popularity among researchers [25]. WEKA is a freely available, Java-based collection of many data mining implementations and visualization tools. Its easy to use GUI interface is better suited for non-technical users like the health care policy makers. Since the software is open-source, any researcher can modify the source and repeat experiments to compare results. A cost-benefit analysis using WEKA was done, and the classifier performance was compared to identify the best classifiers for the current class imbalanced health dataset. In the classifiers defined for prediction of chronic health conditions, we are specifically interested in reducing the false negatives because it has a higher cost. The classifier should be able to predict a significant number of the minority or target class instances. Once the core classifiers are studied, the authors attempt to improve the performance of the classifiers using an ensemble of classifiers and also data sampling techniques.

2. RESEARCH METHOD

The data for this experiment has been provided by the Rural Maternity and Child Welfare Homes (RMCWH) organization, which is the largest private integrated health care delivery network in Karnataka. RMCWHs are manned by the Department of Community Medicine, Kasturba Medical College, Manipal, India. The dataset has a total of 22,598 instances and 53 attributes. The predictor variables in the dataset record the patient's demographics, family details, socioeconomic status, and living conditions. The class attribute is a binary attribute which indicates if the patient has one or more of the following chronic diseases: diabetes, heart disease or high blood pressure. The class is imbalanced with 1311 patients with chronic illness and 20982 healthy patients. This dataset implies a rare case, wherein 5.8% of the total population is the rare positive case. The dataset is also unique because it contains 305 instances with missing class labels. The dataset contains an almost equal number of male and female records. The chronic disease was found in patients who are above the age of 40. The attributes which are highly correlated with the chronic illness occurrence are age, gender and marital status of the patient. The patients with chronic diseases were also found to be from the higher income group.

Based on literature review, the authors selected a subset of WEKA classifiers that are known to work well in the class-imbalanced datasets [25]. Classifiers which work with missing class values were chosen due to a large number of missing values in the health dataset. In the case of the chronic health dataset, the costs of FNs is much more that the cost of FPs. Although it is possible to compute the cost of the FP regarding the cost of diagnostic tests, the cost of late diagnosis and death cannot be easily quantified. The authors chose to represent the WEKA cost matrix in the ratio of 1:10, i.e., The cost of FN is ten times more than the cost of the FP. The widely used stratified 10-fold cross-validation was deployed for the testing of the classifiers, due to its relatively low bias and variance [7], [17]. The core classifiers were compared in terms of total cost and true positive rate. Cost benefit analysis was done with the results of basic classifiers, and the cost function was minimized so as to lower total costs in general as well as reduce the total number of FNs in the classifier.

After the best core classifiers had been identified, the authors conducted experiments to check if cost sensitive learning, filtered classifiers, and ensemble methods could be used to improve the results. WEKA supports ensemble-based classification: boosting, bagging and blending. Boosting was done with the AdaBoostM1 with different base classifiers to see if their results could be improved. Bagging with various base classifiers was performed to see if it results enhanced by the separation of data into samples. Blending was conducted using Stacking in WEKA which is based on the Stacked Aggregation method using a diverse

blend of algorithms. The choice of base classifiers for the ensemble was based on the assumption that base classifiers are independent of each other and that the base classifiers perform better than random guessing.

In the last experiment, the datasets were modified using under-sampling, over-sampling and SMOTE techniques to see their impact on the performance of classifiers. Each of these sampling techniques ensured that both class labels are balanced, using the WEKA filters "Resample," "SpreadSubSample" and "SMOTE." In the first strategy, the "SpreadSubSample" filter which produces a random subsample of a dataset was used. This filter performs under-sampling to ensure a uniform distribution of classes, which resulted in a dataset with 1074 positive instances and 1311 negative instances. In the second strategy, the "Resample" filter was used, with the "biasToUniformDistribution" option to get an over-sampled dataset with replacement. The over-sampled dataset resulted in a dataset with 11299 positive class instances and 11140 negative class instances. The SMOTE filter was also used to resample the dataset using five nearest neighbors to generate 14421 positive instances and 20982 negative instances. The three datasets were then used to produce classifiers using different classifier methods, and the results were ranked based on cost.

3. RESULTS AND DISCUSSION

In the first stage, core classifiers were built, and its performance for the class-imbalanced dataset was analyzed. The data also contains missing values, and only those classifiers which support missing class values were evaluated for their performance. The Support Vector Machine based WEKA implementation LIbSVM, produced an effective classifier with the sigmoid kernel while other kernels like linear and radial resulted in "chance" classifiers which were equivalent to a random guess. This also implies that the solution space is not linearly separable. Chance classifiers were eliminated, and the remaining classifiers were shortlisted and ranked based on the total cost of the classifier (see Table 1). The 'Total Cost' was calculated based on the cost matrix; the costs were further reduced by performing a cost-benefit analysis to ensure a minimum number of FNs. The 'Total Cost (Optimized)' represents the costs after a cost-benefit analysis. Fewer FNs (which results in less cost) and high recall is desirable. The total number of FNs before and after cost-benefit analysis were tabulated. The AUC values in column 8 prove that all these classifiers are not equivalent to random guess but can classify the data in spite of class imbalance.

Table 1. Costs and performance of core classifiers (Top 10)

Rank	WEKA Classifier	Total Cost	Total Cost (Optimized)	Recall	FNs (out of 1311)	FNs (Optimized)	Accuracy Rate (%)	AUC	AUC (PRC)
1	Bayesian Net	5504	4953	0.73	357	174	89.72	0.925	0.379
2	Naïve Bayesian	5602	4988	0.72	367	143	89.69	0.924	0.379
3	Logistic	10480	4928	0.22	1021	172	94.21	0.924	0.387
4	Random Tree	10707	9662	0.24	993	806	92.06	0.657	0.142
5	J48	11137	7467	0.17	1091	528	94.08	0.836	0.286
6	VotedPerceptron	11488	11327	0.14	1127	1105	93.97	0.572	0.117
7	JRIP	11447	11468	0.14	1122	1122	93.95	0.576	0.128
8	LibSVM – Sigmoid kernel	11945	11971	0.12	1153	1311	92.97	0.550	0.083
9	SimpleCart	12054	9158	0.09	1194	743	94.13	0.768	0.219
10	IBK, K=5	12548	10029	0.05	1249	577	94.14	0.721	0.178

In general, most of the top classifiers exhibited low total cost, and the number of FNs was drastically reduced using the cost-benefit analysis. The Bayesian classifiers had the best performance regarding the cost of the core classifiers in the class imbalanced health dataset (Table 1). The WEKA based paired t-test proved that there is no difference in the performance of the Naïve Bayesian and Bayesian net classifiers. The Bayesian net classifier is preferred because initial exploration shows a strong correlation among the patient features. Bayesian classifiers are known to work well in situations like medical diagnosis, wherein the relationship between the attribute set and class variable is non-deterministic. Bayesian classifiers are also robust to noise, irrelevant attributes and confounding factors that are not included in the classification. All the other algorithms like Logistic, Random Tree and Voted Perceptron have a high number of false negatives which drastically increases the cost of the classifier. The results also contradict the results of Weiss [12] who advocated the use of instance based learners for the class imbalance problem. The rule-

based classifier, JRIP which is an implementation of a propositional rule learner, is well suited for handling class imbalances and appears in the top 10 classifiers. This result is in line with previous results which suggested that kernel based SVMS work better in class imbalance problems [12], [13], [15]. Even though the classifier accuracy for all classifiers is high, between 89% and 94 %, only the Bayesian classifiers, have a high recall (0.73) and low number of FNs.

In the second experiment, the effect of ensembling methods like Random Forest, Boosting, Bagging, Stacking and Voting on these baseline classifiers was studied. The results are tabulated in Table 2. The cost sensitive learning and the meta cost implementations in WEKA were also evaluated. The classifier performance was again ranked based on total cost which was further optimized using cost-benefit analysis and tabulated in Table 2.

Table 2. Costs and performance of Ensemble & Cost based Classifiers (Top 10)

Rank	Classifier	Total Cost	Total cost (optimized)	Recall	FNs (out of 1311)	FNs (optimized)	Accuracy Rate (%)	AUC	AUC (PRC)
1	Cost Sensitive (BayesNet)	4968	5007	0.90	137	169	83.25	0.924	0.378
2	Filtered class, Class Balancer, (BayesNet)	5150	5000	0.91	117	169	81.62	0.924	0.378
3	Cost sensitive (JRIP)	5281	5448	0.85	197	214	84.26	0.857	0.262
4	MetaCost (BayesNet)	5426	5185	0.93	95	192	79.49	0.919	0.355
5	Cost sensitive (Logistic)	5507	5607	0.85	196	223	83.21	0.904	0.312
6	Bagging (BayesNet)	5575	4969	0.72	368	171	89.85	0.924	0.380
7	Filtered class, Class Balancer, (Logistic)	5618	5547	0.88	153	211	80.98	0.907	0.320
8	Filtered class Class Balancer, (J48)	6240	6287	0.70	399	403	88.12	0.797	0.231
9	Vote (BayesNet with Logistic)	6443	4936	0.60	520	191	92.09	0.928	0.394
10	Vote (BayesNet with random forest)	6543	4836	0.59	534	154	92.20	0.927	0.400

The cost sensitive learning implementation with different core classifiers exhibited the best performance. In the case of the JRIP and Logistic classifiers, the cost-sensitive learning approach almost halves the total cost of the core implementation. The filtered classifier with the class balancer filter produces good results with Logistic and J48 methods which were previously affected by the class imbalance. However, the cost-benefit analysis sometimes led to an increase in overall cost even though the number of FNs were low. The increase in total cost was due to a massive increase in FPs as a result of threshold moving. As indicated in the literature, the ensemble methods like Voting and ADABoostM1 significantly increase the performance of classifiers in imbalanced class datasets in comparison to core classifiers [19].

In the third experiment, the effect of sampling to balance the classes was done using techniques like under-sampling, oversampling and SMOTE (Table 3). The results were ranked based on total cost, and a cost-benefit analysis was performed to see if costs could be reduced. In general, as indicated in the literature, under-sampling seems to work better than over-sampling and SMOTE [22]. The authors recommend the usage of random under-sampling as a solution for class imbalanced datasets because it is also computationally less expensive to implement than SMOTE or over-sampling. It also reduces the size of the dataset, which will improve time complexity without sacrificing classification performance. In the case of the J48 and IBK, it was observed that all three sampling strategies improved the cost dramatically. The sampling results indicate that J48 and IBK work better in class balanced datasets. This result contradicts some previous empirical results; the sampling methods outperformed the cost-sensitive methods [11].

Table 3. Costs and performance of Classifiers using sampling techniques (Top 10)

Rank	Classifier	Total Cost	Total cost (optimized)	Recall	Accuracy Rate (%)	FNs	FNs (optimized)	AUC	AUC (PRC)
1	Oversampling Random Tree	766	530	1.00	96.58	0/11299	5	0.979	0.959
2	UnderSampling JRIP	1198	1349	0.92	84.10	91/1074	68	0.841	0.711
3	UnderSampling Bayesnet	1215	756	0.91	86.04	98/1074	11	0.886	0.777
4	OverSampling J48	1265	1292	0.99	95.20	21/1074	14	0.963	0.934
5	UnderSampling J48	1455	1170	0.89	83.89	119/1074	65	0.838	0.693
	UnderSampling VotedPerceptron	1533	1039	0.88	85.53	132/1074	56	0.847	0.709
6	UnderSampling Logistic	1682	804	0.86	84.19	145/1074	11	0.875	0.748
7	UnderSampling IBK k=5	2205	1096	0.82	78.49	188/1074	9	0.831	0.709
8	UnderSampling Random Tree	3271	1548	0.72	75.30	298/1074	0	0.750	0.611
9	UnderSampling LIbSVM (sigmoid kernel)	3528	1431	0.71	70.57	314/1074	105	0.687	0.541
10	Oversampling IBK k5	5033	3991	0.99	83.95	159/11299	956	0.962	0.938

4. CONCLUSION

This work is relevant to public health policy makers, who can use the classifiers to predict the occurrence of chronic disease in the population and also identify the factors that are correlated with chronic diseases. The classifiers will help health care providers in improving their prognosis, diagnosis and treatment plans. Experiments were conducted based on various approaches suggested in the literature, to tackle the class imbalance problem. The WEKA based classifiers were used to record and analyze the classifier performance in terms of cost. The Bayesian classifiers were identified as the best classifiers for the class imbalanced dataset. The authors recommend the Bayesian Net classifier because of underlying correlation among patient features. The cost sensitive implementations and cost-benefit analysis can further reduce the total cost while maintaining the accuracy. However, the ensemble methods is a complex solution wherein there are a huge number of solutions that still needs to be explored. Under-sampling is an efficient data pre-processing approach with low computation costs, and it is recommended for building cost effective classifiers. The under-sampling can dramatically improve the performance of methods like J48, IBK which are affected by the class imbalance. The current work assumes that the cost of FNs is ten times more than the cost of TPs. The work can be improved by actually quantifying the actual costs in generating classifier errors. In future work, the effect of feature selection on the classifier cost will be studied. Though irrelevant features are not known to improve classification performance significantly, they can slow down the classifier process.

ACKNOWLEDGEMENT

We thank Dr. Harishchandra Hebbar, Professor, School of Information Sciences, Manipal and the Department of Community Medicine, KMC, Manipal for sharing with us valuable data. We thank Dr. Veena Kamath, Professor, Department of Community Medicine, KMC, Manipal, for extending us her subject expertise.

REFERENCES

[1] Tomar D, Agarwal S. "A survey on Data Mining approaches for Healthcare", *International Journal of Bio-Science and Bio-Technology*. 2013; 5(5): 241-266.

[2] Dissanayaka C, Abdullah H, Ahmed B, Penzel T, Cvetkovic D. *"Classification of Healthy Subjects and Insomniac Patients Based on Automated Sleep Onset Detection"*. In International Conference for Innovation in Biomedical Engineering and Life Sciences: ICIBEL2015; 2015; Putrajaya, Malaysia.
[3] Chien C, Pottie GJ, "A Universal Hybrid Decision Tree Classifier Design for Human", In 34th Annual International Conference of the IEEE EMBS; 2012; San Diego, USA.
[4] Wang Y, Li Pf, Tian Y, Ren Jj, Li Js, "A Shared Decision Making System for Diabetes Medication Choice Utilizing Electronic Health Record Data", *EEE Journal of Biomedical and Health Informatics*. 2016; pp(99):1-1.
[5] Konda S, Balmuri KR, Basireddy RR, Mogili R, "Hybrid Approach for Prediction of Cardiovascular Disease Using Class Association Rules and MLP", *International Journal of Electrical and Computer Engineering.* 2016; 6(4):1800.
[6] Boris Milovic, Milan Milovic, "Prediction and Decision Making in Health Care using Data Mining", *International Journal of Public Health Science*, December 2012; 1(2): 69-78.
[7] Pollettini JT, Panico SRG, Daneluzzi JC, Tinós R, Baranauskas JA, Macedo AA, "Using Machine Learning Classifiers to Assist Healthcare-Related Decisions: Classification of Electronic Patient Records", *Journal of Medical Systems*, 2012; 36: 3861-3874.
[8] Herland M, Khoshgoftaar TM, Wald R, "A review of data mining using big data in health informatics", *Journal of Big Data*, 2014; 1:2.
[9] Takeda F, Tamiya N, Noguchi H, Monma T, "Relation between Mental Health Status and Psychosocial Stressors among Pregnant and Puerperium Women in Japan: From the Perspective of Working Status", *International Journal of Public Health Science*, 2012; 1(2): 37-48.
[10] Milovic B, Milovic M, "Prediction and Decision Making in Health Care using Data Mining", I*nternational Journal of Public Health Science*, 2012; 1(2): 69-78.
[11] Chawla NV. Data Mining for Imbalanced Datasets: an Overview. In Data Mining and Knowledge Discovery Handbook, Springer US; 2005, 853-867.
[12] Weiss GM., "Mining with Rarity: A Unifying Framework, ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets", June 2004; 6(1):7-19.
[13] Japkowicz N, *"The Class Imbalance Problem: Significance and Strategies"*, In the 2000 International Conference on Artificial Intelligence [ICAI]; 2000; Las Vegas, USA.
[14] Saito T, Rehmsmeier M, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets", PLoS ONE 10(3): e0118432. doi:10.1371/journal.pone.0118432
[15] Lusa L, Blagus R, *"The class-imbalance problem for high-dimensional class prediction"*, in 2012 11th International Conference on Machine Learning and Applications; 2012.
[16] Hempstalk K, Frank E, Witten IH, "One-class Classification by Combining Density and Class Probability Estimation", Machine Learning and Knowledge Discovery in Databases, 2008; 5211: 505-519.
[17] Tan Pn, Steinbach M, Kumar V. Introduction to Data Mining: Pearson Publication; 2014.
[18] Joshi MV, Agarwal RC, Kumar V, *"Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Induction"*, in the 2001 ACM SIGMOD international conference on Management of data; 2001; New York, USA.
[19] Joshi MV, Agarwal RC, Kumar V, *"Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong?"*, in the eighth ACM SIGKDD international conference on Knowledge discovery and data mining; 2002; New York , USA.
[20] Joshi MV, Kumar V, *"CREDOS: Classification using Ripple down Structure [A Case for Rare Classes]"*, In the 2004 SIAM International Conference on Data Mining; 2004; Florida, USA.
[21] Dittman DJ, Khoshgoftaar TM, RandallWald , Napolitano A, "Comparison of Data Sampling Approaches for Imbalanced Bioinformatics Data", In the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference; 2014; Florida.
[22] Blagus R, Lusa L, "SMOTE for high-dimensional class-imbalanced data", BMC Bioinformatics, 2013; 14:106.
[23] Davis J, Goadrich M, "The Relationship between Precision-Recall and ROC Curves", In 27 rd International Conference on Machine Learning; 2006; Pittsburgh, USA.
[24] Janez Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets", *Journal of Machine Learning Research*, 2006; 7: 1-30.
[25] Frank E, Hall MA, Witten IH, "The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann; 2016 [cited 2016 May 01, Available from: http://www.cs.waikato.ac.nz/ml/weka/.