American Journal of Bioinformatics Research 2017, 7(1): 1-8
DOI: 10.5923/j.bioinformatics.20170701.01

Ensemble Learning Methods to Deal with Imbalanced Disease and Left-Skewed Cost Data

Songul Cinaroglu

Department of Healthcare Management, Hacettepe University, Ankara, Turkey

Abstract Rare events and class imbalance is very often in classification problems. Rare diseases which are good example for rare events are life-threating and vast majority of them are genetically determined. Moreover, traditionally appropriate analysis of cost data generated by clinical trial is problematic. The distribution of cost data is generally highly skewed because a few patients faced with large costs. Several ensemble learning methods (ELM) were applied to health care datasets such as predicting individual expenditures and disease risks for patients. These methods are consists of a set of individual training classifiers such as Bagging and Boosting. This study aims to compare ELM classification performances applied on thyroid disease dataset. Data came from UCI Machine Learning Repository. Diagnosed as a hyperthyroid determined as a dependent variable for classification. ID.3, C4.5, CART, NB, KNN, RF, SVM, NN were used as ELMs. Bagging and Boosting were implemented to improve prediction performances. “k” 10 fold cross validation and AUC was examined to evaluate classication performances of ELMs. Study results reveal that single ELM have superior prediction performance compared with Bagging and Boosting applications. In addition to that kNN, RF and NN have superior classification performance compared with other ELMs. Future research is needed to better understand the role of ELM to improve prediction performance of rare disease data.

Keywords Imbalanced Data, Rare Diseases, Ensemble Learning, Left-Skewed Cost Data

1. Introduction

Rare events that occur with low frequency, if one class contains significantly more samples than the others. These events frequently causes imbalanced data problem in statistics. Presenting imbalanced data to the classifier will produce undesirable results [1]. Rare diseases are good examples for rare events. Most rare diseases are life-threating and the vast majority of them are genetically determined. Low prevalence of rare diseases requires special attention to improve diagnosis, care and prevention [2]. Fraudulent credit card transactions [3], word mispronunciation [4], oil spills [5], train derailments [6], tornadoes [7] are popular examples of rare events. The low prevalence of rare diseases improves their social and economic impact. Economic impacts are associated with specialized health and educational services, loss of income for caregivers and loss of productivity for society are some of these impacts [8].

Disease prediction is becoming a prominent research area due to the increasing popularity of big datasets. Large public datasets are valuable sources and they are still a valuable resources to obtain useful information about rare diseases. First, they may provide population level clinical information. Second, they are available to develop methodologies for clinical decision support systems that can be employed for electronic medical records [1]. On the other hand, the appropriate analysis of cost data generated by clinical trial is problematic. While the usual outcome of interest is the population mean cost for a particular treatment, the distribution of cost data is generally highly skew because a few patients incur very large costs [9]. The distribution of costs and expenditures for health care shares a number of characteristics that make their use in economic analysis difficult. The costs are typically highly skewed to the right. Under these circumstances analysts have often found that use of standard least squares estimators often leads to analytical problems from highly influential outliers [10]. A minority of patients are responsible for a high proportion of health care costs is one of the biggest reason of that rare events are more observed than severe cases [11].

Learning classifiers from imbalanced datasets is important and observed very often in practice. Traditional classifiers tend to classify all the data into the majority class, which is usually the less important class. Scholars suggest that it is hard to achieve good prediction performance results while using traditional methods, several machine learning techniques were applied to healthcare datasets to improve future prediction of diseases. These modern statistical

* Corresponding author:
songulcinaroglu@gmail.com (Songul Cinaroglu)

Published online at http://journal.sapub.org/bioinformatics

learning algorithms are effective methods for predicting individual expenditures and disease risks for patients. Modern methods combine the predictions of multiple base learners to form ensembles, which typically achieve better predictive performance than individual base learners [12]. Several machine learning techniques were applied to health care datasets these include such as predicting individual expenses and disease risks for patients.

Ensemble learning methods (ELMs) are effective methods to deal with imbalanced data. They consists of a set of individual classifiers, such as decision trees. Decision trees are combined novel instances and they are more accurate than any of the single classifiers in the ensemble. ID.3 (Interactive Dichotomizer 3), C4.5 and CART (Classification and Regression Trees) are well known examples of decision tree based ELMs. ID.3 algorithm uses the concept of information gain. It uses information entropy minimization criteria in tree growing process. C4.5 is an another popular decision tree algorithm, it is an extended version of ID.3 algorithm. CART generates binary decision tree constructed by splitting the data in a node into small nodes repeatedly, starting with the root node that includes the whole learning sample [13]. RF (Random Forest) is an ensemble learner and a method that generates many classifiers and gathers their results. RF will produce multiple CART trees. Each tree in the RF will take a vote for some input x, then the output of the classifier is determined by majority voting of the trees. RF can handle with high dimensional data and use a large number of trees in the ensemble [1]. RF is an effective method to overcome imbalanced data. It also estimates the importance of variables used in the classification. kNN (k-Nearest Neighbors) is an another well-known ensemble learning method to handle imbalanced data. It is a non-parametric method used for classification and regression. The input consists of the “k” closest training examples in the feature space. In k-NN classification, the output is a categoric variable. An object is classified by a majority vote to its neighbors in this algorithm. Thus, an object is classified by a majority vote to its neighbors, with the object being assigned to the class most common among its k nearest neighbor [14]. SVM (Support Vector Machine) is a binary classifier and it is assumed to be linearly separable in the input sphere. For binary case it is suggested that a hyper plane exist such that all points belongs to one class are on the one side, and all points belonging to the other class are on the other side of the hyperplane [13]. NB (Naive Bayes) is an another ensemble learning method assign class labels to problem instances. All NB classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable [15]. In the last decade the use of artificial intelligence has become widely accepted in medical applications. NN (Neural Networks) are popular methods of artificial intelligence. Ease optimization and accuracy of prediction, flexible non-linear modelling of large datasets and potential to support clinical decision making are advantages of using NN [16].

Bagging and Boosting are comparably new methods for generating ensembles [17-19]. Bagging predictor is a method for generating numerous versions of predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a majority vote when predicting a class [17]. Bagging can make weak learners to learn parallel since random dataset is used for training [20]. Boosting [18] includes a family of methods. The focus of these methods is to produce a series of classifiers. The training set used for each number of the series is chosen based on the performance of the earlier classifier(s) in the series. In Boosting, instances that are incorrectly predicted by former classifiers in the series are selected more often than examples that were correctly predicted. Thus, Boosting produce new classifiers that are better able to predict examples for which the current ensemble’s performance is bad. AUC (Area Under the ROC Curve) is a performance measure of a plot that represents the performance of a binary classifier system. The curve is created by plotting the true positive rate against the false positive rate. AUC values lies between 0.5 and 1 where 0.5 is a bad classifier and 1 denotes an excellent classifier [13].

Here are some examples from the literature emphasizing prediction performance differences between ELMs. Davis et al. (2008) was used ELMs to predict individual disease risk based on medical history. The prediction was performed multiple times for each patient, each time employing different sets of variables. In the end, the clustering were combined to form an ensemble [21]. Moturu et al. (2007) predicted future high cost patients, data taken from Arizona Medicaid program and 20 non-random data samples created, each sample with 1.000 data points to overcome the problem of imbalanced data. Variety of classification methods such as: SVM, Logistic regression, Logistic Model Trees, AdaBoost and LogitBoost were used in the analysis [22]. Mantzaris et al. (2008) predicted Osteoporosis using NN [23]. Hebert et al. (1999) identified persons with diabetes using Medicare claims data [24]. They have constructed a problem where the diabetes claims occur less frequently to be sensitive indicators for persons with diabetes. Yu et al. (2010) examined a method using SVM for detecting persons with diabetes and prediabetes [25]. Zhang et al. (2009) compared prediction performances of AdaBoost, LogitBoost and RF to logistic regression and SVM in the classification of breast cancer metastasis [26]. They concluded that ensemble learners have higher accuracy compared to the non-ensemble learners. There are large number of studies in the literature emphasizing the prediction performance differences of ELMs. However, lack number of studies that has been carried out specifically to compare prediction performance of ELMs while using rare diseases and left-skewed cost data. To fill this void in the literature this study aims to compare prediction performances of ELMs while implementing bagging and boosting algorithms on rare disease and left-skewed cost data. The next sections of the paper describe study materials and methods, analysis, results and conclusions.

2. Metarials & Methods

2.1. Aims

The aim of this study is to compare classification performance of ELMs using rare disease and left skewed cost data. For this aim algorithms which are ID.3, C4.5, CART, NB, KNN, RF, SVM, NN were implemented on thyroid disease data while performing Bagging and Boosting respectively.

2.2. Dataset

The thyroid disease dataset came from UCI Machine Learning Repository-Center for Machine Learning and Intelligent Systems. Table 1 shows description of thyroid disease dataset. The thyroid is an endocrine gland in the neck, consisting of two lobes connected by an isthmus. The thyroid gland secretes thyroid hormones, which primarily influence the metabolic rate and protein synthesis. The thyroid disease (ann-thyroid) dataset is a classification dataset. The problem is to determine whether a patient referred to the clinic is hypothyroid [27, 28]. Basal thyroid-stimulating hormone (TSH) test costs for individual patients considered as cost variable. This cost is in Canadian dollars and the cost information is from the Ontario Health Insurance program's fee schedule (Table 1).

Table 1. Hyperthroid Dataset Description

Variable	Explanation /Categories	Type
Age (A)	Age of the patient	Continuous
Sex (S)	Male, Female	Categoric
On Thyroxine (OT)*	True, False	Categoric
Query on Thyroxine (QoT)	True, False	Categoric
On Antithyroid Medication (OAM)*	True, False	Categoric
Thyroid Surgery (TS)	True, False	Categoric
Query Hypothyroid (QHPT)*	True, False	Categoric
Query Hyperthyroid (QHRT)*	True, False	Categoric
Pregnant (P)	True, False	Categoric
Sick (S)	True, False	Categoric
Tumor (T)	True, False	Categoric
Lithium (L)*	True, False	Categoric
Goitre (G)*	True, False	Categoric
TSH* Test Costs (TSHC)	Continuous	Continuous

Explanations:*Thyroxine: *is a hormone the thyroid gland secretes into the bloodstream.* ***Antithyroid Medication:*** *sometimes written as anti-thyroid medications- are a common treatment for hyperthyroidism.* ***Hypothyroid:*** *Hypothyroid relating to or affected with hypothyroidism.* ***Hyperthyroid:*** *Hyperthyroid is the condition that occurs due to excessive production of thyroid hormone by the thyroid gland.* ***Goiter:*** *Goiter is the most common thyroid abnormality in lithium-treated patients, occurring in approximately 40 to 50 percent* ***TSH:*** *Thyroid Stimulating Hormone measured by radioimmuno assay. The cost information is from the Ontario Health Insurance Program's fee schedule. The cost is in Canadian dollars. The costs in this file are for individual tests, considered in isolation.*

Source: Thyroid disease records supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Sydney, Australia. UCI Machine Learning Repository-Center for Machine Learning and Intelligent Systems.

2.3. Analysis

Diagnosed as a hyperthyroid determined as a dependent variable all other variables are determined as covariates. ID.3, C4.5, CART, NB, KNN, RF, SVM, NN were used, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 trees were generated for RF. Bagging and Boosting were implemented to improve the accuracy of classification performances. "k" 10 fold cross validation was performed and an area plot was used to visualize prediction performances of ELMs. AUC was used to evaluate classication performance of ensemble methods. All variables are normalized to have zero means and unit variances.

This will allow to handle with parameters of different units and scales. Logarithmic transformation implemented to TSH cost data. AUC was used for comparison of classification performances. 5-fold cross validation was performed in the analysis. Prediction performances of different ELMs are visualized on an area graph and Kruskall Wallis variance analysis was performed for comparison of ELMs prediction performance differences.

3. Results

3.1. Descriptive Statistics

As previously mentioned, in this study being diagnosed by thyroid disease determined as a dependent variable for classification based model. 4.8% (151) of patients were diagnosed by thyroid and called "hyperthyroid" group. However, 95.2% (3012) of them were non-hyperthyroid. Descriptive statistics for predictive variables are as follows: 69% of patients were female, 85.4% of patients didn't on thyroxin, 98.3% of them didn't query on thyroxin, 98.7% of them didn't take antithyroid medication, 96.7% of them didn't have thyroid surgery, 92.4% of them didn't query hypothyroid, 98% of them were not pregnant, 96.9% of them were not sick, 98.7% of them didn't have a tumor, 99.9% of them didn't have a lithium, 96.9% of them didn't have a goiter and 92.3% of them didn't query hyperthyroid.

The mean values of continuous predictive variables are as follows; age 51 (±17.88) and TSH costs 19.40 (±8.08). Figure 1 shows distribution of TSH cost data, it is seen that the distribution of TSH cost data is negatively skewed. As a part of preliminary analysis procedure Z transformation was implemented on TSH cost data. After that, natural logarithm (Ln) of TSH cost was taken.

It is seen that the distribution of TSH cost data become positively skewed after transformation (Figure 1).

Table 2. Descriptive Statistics for Categorical Variables

Variables	Categories	n	%	Variables	Categories	n	%
Sex	**Male**	981	31	**Pregnant**	**True**	63	2
	Female	2182	69		**False**	3100	98
On Thyroxine	**True**	461	14.6	**Sick**	**True**	99	3.1
	False	2702	85.4		**False**	3064	96.9
Query on Thyroxine	**True**	55	1.7	**Tumor**	**True**	40	1.3
	False	3108	98.3		**False**	3123	98.7
On Antithyroid Medication	**True**	42	1.3	**Lithium**	**True**	2	0.1
	False	3121	98.7		**False**	3161	99.9
Thyroid Surgery	**True**	104	3.3	**Goitre**	**True**	99	3.1
	False	3059	96.7		**False**	3064	96.9
Query Hypothyroid	**True**	241	7.6	**Query Hyperthyroid**	**True**	243	7.7
	False	2922	92.4		**False**	2920	92.3
Total		3163	100	**Total**		3163	100

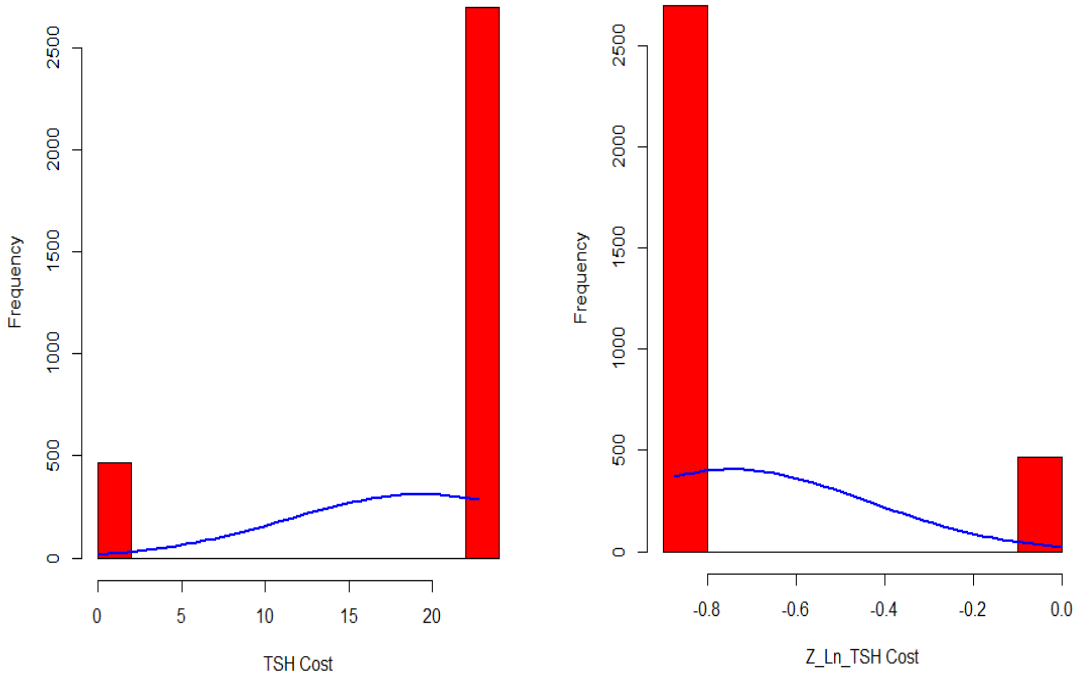

Figure 1. Distribution of TSH Cost Data

Table 3. Descriptive Statistics for Continuous Variables

Continuous Variables		
Age	**Mean**	**SD.***
	51.13	17.88
TSH Costs	**Mean**	**SD.**
	19.40	8.08

***SD:** *Standard Deviation*

3.2. Correlations between Independent Variables and ELM Performance Comparison

As a part of preliminary analysis procedure, correlations between independent variables examined by using Pearson correlation coefficient. Standardized Z scores of study variables was used in examination of correlation coefficients of study variables. Table 4 shows matrix of Spearman correlation coefficients of study variables. Literature suggests that the magnitude of 0.70 and higher indicate variables which have high correlations [29]. All correlation coefficients are lower in this table in other words, there is no multi collinearity problem was detected between independent variables. Table 5 shows ELMs performance comparisons. AUC was used as a performance measure and k=5 fold cross validation was implemented as a part of cross validation procedure. Figure 2 visualize performance results differences between prediction methods. Kruskall Wallis variance analysis verifies statistical significance of prediction performance differences between different ELMs.

Table 4. Correlations between Independent Variables

	A	S	OT	QoT	AM	TS	QHPT	QHRT	P	S	T	L	G	TSHC
A	1													
S	-0.02	1												
QT	-0.01	0.13^{**}	1											
QoT	-0.02	-0.04^{**}	-0.01	1										
AM	-0.05^{**}	0.02	-0.02	-0.01	1									
TS	-0.01	0.03^{*}	0.02	0.01	-0.06	1								
QHPT	0.04^{**}	0.07^{**}	0.07^{**}	-0.02	-0.03	0.01	1							
QHRT	-0.08^{**}	0.06^{**}	-0.08^{**}	-0.03^{*}	0.12^{**}	0.02	-0.02	1						
P	-0.15^{**}	0.09^{**}	-0.01	-0.01	0.03	-0.01	-0.01	0.06	1					
S	0.06^{**}	-0.01	-0.06^{**}	-0.02	-0.02	-0.03	-0.05^{**}	-0.05	-0.02	1				
T	-0.06^{**}	0.09	-0.02	0.05^{**}	-0.03	-0.02	-0.03	-0.03	0.04	-0.02	1			
L	-0.02	0.01	-0.01	-0.03	-0.03	-0.05	0.04^{*}	-0.07	-0.04	-0.05	-0.03	1		
G	-0.03^{*}	0.09^{**}	-0.02	0.04	-0.02	-0.02	-0.02	0.04	0.07	-0.03	-0.02	-0.05	1	
TSHC	-0.20^{**}	0.03	-0.07^{**}	0.10^{**}	0.02	-0.03^{*}	-0.05^{**}	0.13	-0.08	-0.04	0.03	-0.01	0.07	1

Pearson correlation coefficient **p <0.01, *p <0.05

Table 5. Ensemble Learning Methods Performance Comparison Using AUC Values

ELM Single	CVM	AUC	ELM Bagging	CVM	AUC	ELM Boosting	CVM	AUC
ID.3	k=5	0.6000	ID.3 Bag.	k=5	0.5491	ID.3 Boost.	k=5	0.5190
C4.5	k=5	0.5883	C4.5 Bag.	k=5	0.5539	C4.5 Boost.	k=5	0.5052
CART	k=5	0.5976	CART Bag.	k=5	0.5523	CART Boost.	k=5	0.5187
NB	k=5	0.4808	NB Bag.	k=5	0.4965	NB Boost.	k=5	0.5002
kNN	k=5	0.6031	kNN Bag.	k=5	0.5831	kNN Boost.	k=5	0.6302
RF 10	k=5	0.6203	RF 10 Bag.	k=5	0.5097	RF 10 Boost.	k=5	0.5000
RF 20	k=5	0.6257	RF 20 Bag.	k=5	0.5029	RF 20 Boost.	k=5	0.5000
RF 30	k=5	0.6493	RF 30 Bag.	k=5	0.5032	RF 30 Boost.	k=5	0.5000
RF 40	k=5	0.6502	RF 40 Bag.	k=5	0.5032	RF 40 Boost.	k=5	0.5000
RF 50	k=5	0.6532	RF 50 Bag.	k=5	0.5032	RF 50 Boost.	k=5	0.5000
RF 60	k=5	0.6551	RF 60 Bag.	k=5	0.5032	RF 60 Boost.	k=5	0.5000
RF 70	k=5	0.6537	RF 70 Bag.	k=5	0.5000	RF 70 Boost.	k=5	0.5000
RF 80	k=5	0.6539	RF 80 Bag.	k=5	0.5000	RF 80 Boost.	k=5	0.5000
RF 90	k=5	0.6542	RF 90 Bag.	k=5	0.5000	RF 90 Boost.	k=5	0.5000
RF 100	k=5	0.6509	RF 100 Bag.	k=5	0.5000	RF 100 Boost.	k=5	0.5000
SVM	k=5	0.5700	SVM Bag.	k=5	0.5046	SVM Boost.	k=5	0.5000
NN	k=5	0.6393	NN Bag.	k=5	0.5084	NN Boost.	k=5	0.5000

Abbreviations: Bag. *Bagging,* ***Boost.*** *Boosting,* ***ELM:*** *Ensemble Learning Methods,* ***AUC:*** *Area Under the ROC Curve,* ***CVM:*** *Cross Validation Method,* ***k:*** *"k" fold cross validation,* ***ID.3:*** *Iterative Dichotomiser 3* ***C4.5:*** *Extended version of ID.3,* ***CART****: Classification and Regression Trees,* ***NB:*** *Naive Bayes,* ***kNN:*** *k-Nearest Neighbor,* ***RF:*** *Random Forest,* ***SVM:*** *Support Vector Machine,* ***NN:*** *Neural Network (Table 5).*

Figure 2 shows comparison of AUC prediction performances of ELMs. It is seen that the application of single ELMs have superior performance results compared with Bagging and Boosting applications. In addition to that kNN, RF and NN has superior prediction performance compared with other ELMs.

Kruskall Wallis Variance analysis test results shows that the difference between prediction performance results of ELM are statistically significant ($X^2 = 26.33$; p<0.001).

Table 6. Kruskall Wallis Variance Analysis Test Results Difference

ELM	N	Mean Rank	Chi-Square	p
ELM-Single	17	40.53	26.33	<0.001
ELM-Bagging	17	21.47		
ELM-Boosting	17	16		

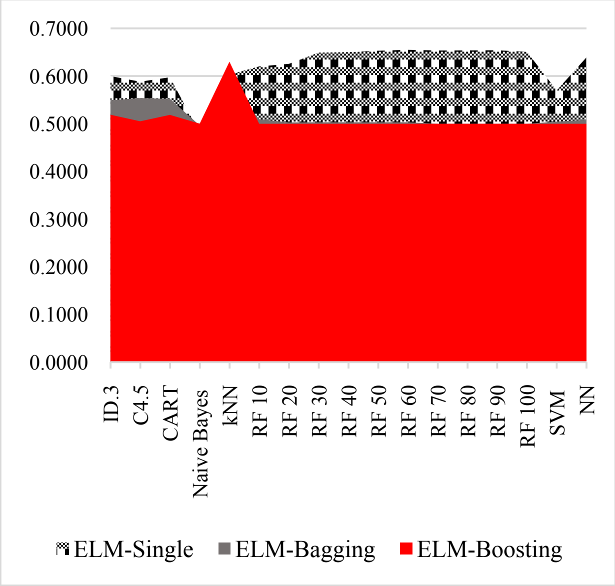

Figure 2. Comparison of AUC Prediction Performances of ELM

4. Conclusions

Rare events and class imbalance are critical to prediction in the field of data mining and particularly data classification [30]. Thyroid disease dataset used in this study which is highly imbalanced. In this data 4.8% of the patients have hyperthyroid, 95.2% of them are not. ELM examined and compared to improve prediction performances and will achieve high classification accuracy. ID.3, C4.5, CART, NB, kNN, RF, SVM and NN were implemented on thyroid disease data. k=5 fold cross validation was implemented to achieve better performance results. Study results show that single ELM have superior performance compared with Bagging and Boosting applications. Moreover kNN, RF and NN have better prediction results compared with other ELM.

Despite scholars suggest that “all models are wrong” and “no data are normally distributed” [31, 32]. This study examined a detailed preliminary analysis procedure to overcome measurement unit differences of predictive variables, multicollinearity problem and class imbalance of predictive variable and left-skewed distribution of cost data. Z transformation, Pearson correlation coefficient and natural logarithm of cost data implemented into the dataset. Highly skewed nature of cost data is not a new issue. It is highly believed that parametric models not work to handle highly positive skewed nature of cost data. However, for this study TSH cost data has a left-skewed distribution. QALYs distribution which takes less attention in the literature has heavy left tails the same with TSH cost data in this study. As a part of the preliminary analysis procedure TSH cost data normalized with Z transformation and after that natural logarithm was implemented. After normalization and Ln transformation the shape of TSH cost data became positively skewed. Finally, study results supports that kNN, RF and NN are prior algorithms are to handle imbalanced dataset to predict thyroid disease.

A number of studies in the literature supports our study results and emphasize the superior performance of kNN, RF and NN. They have concluded that kNN, RF and NN are computationally efficient and better handle with highly imbalanced dataset. Moreover, they are more vulnerable to noise detection compared with other methods [33]. In addition to that, previous work has demonstrated that Bagging and Boosting are very effective methods for decision trees. However, there has been little empirical testing with NN. Previous authors are concentrated on decision trees due to their fast training speed and well-established default parameter settings [34]. As a support for previous study results, after performing extensive normalization, transformation and cross validation exercises, study results show that three of the ELMs performed well enough to be used in many application. One question is necessary to answer is why kNN was more effective than other classifiers. Horton and Nakai (1997) answers that question and suggests that other classifiers suffer from some shortcomings like data fragmentation and repeatedly partitioning [35]. From the other point of view, as kNN performs well with small number of input variables, but struggles when the number of inputs is very large. Also kNN works well if all the data has same scale. Thus, data normalization is a good idea to improve performance of kNN [36]. Chernozhukov et al. (2016) suggest that modern supervised machine learning methods are designed to solve prediction problems very well [37]. In order to avoid overfitting problem and improve prediction performance results, it is advisable to use a very broad set of ELMs to improve prediction performances. A number of studies in the literature compare prediction performances of different ELM. However, there is little evidence focus on rare diseases as an example of rare events. It is hoped that in the light of this study, future studies will applied aggregated methods on rare disease data to solve rare events problems.

REFERENCES

[1] M. Khalilia, S. Chakraborty and M. Popescu “Predicting Disease Risks from Highly Imbalanced Data Using Random Forest”, *BMC Medical Informatics and Decision Modelling*, vol. 11, no. 51, pp.1-13, 2011.

[2] S. Ayme and J. Schmidtke “Networking For Rare Diseases: An Necessity for Europe”, Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz, pp. 1477-1483, 2007.

[3] P.K. Chan and S. J. Stolfo “Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection”, *in Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pp.164–168, AAAI Press, 1998.

[4] B. Busser, W. Daelemans and A. Bosch “Machine Learning Of Word Pronunciation: The Case Against Abstraction”, *Sixth European Conference on Speech Communication and*

Technology-EUROSPEECH, Budapest, Hungary, 1999.

[5] M. Kubat, R. C. Holte and S. Matwin (1998) "Machine Learning for the Detection of Oil Spills in Satellite Radar Images", Machine Learning, vol. 30, no. 2, pp. 195-215.

[6] J. Quigley, T. Bedford and L. Walls "Estimating Rate of Occurrence of Rare Events with Empirical Bayes: A Railway Application", *Reliability Engineering and System Safety*, vol. 92, no. 5, pp. 619-627, 2007.

[7] T.B. Trafalis, H. Ince and M.B. Richman "Tornado Detection with Support Vector Machines", *International Conference on Computational Science*, pp. 289-298, 2003.

[8] Y. Zurynski, K. Frith, H. Leonard and E. Elliot "Rare Childhood Diseases: How Should We Respond", *Arch Dis Child*, vol.93, pp. 1071-1074, 2008.

[9] A. Briggs, R. Nixon, S. Dixon and S. Thompson (2005) "Parametric Modelling of Cost Data: Some Simulation Evidence", *Health Economics*, vol. 14, no. 4, pp. 421-428.

[10] W. Manning "Dealing with Skewed Data on Costs and Expenditures" Chapter 41, pp. 439-454. Jones A.M. (2006) *The Elgar Companion to Health Economics*, Second Edition, 2006.

[11] A. Manca and S. Palmer (2005) "Handling Missing Data in Patient-Level Cost-Effectiveness Analysis Alongside Randomized Clinical Trials", *Appl Health Econ Health Policy*, vol. 4, no. 2, pp. 66-75.

[12] L. Rokach "Ensemble Based Classifiers" *Artificial Intelligence Review*, vol. 33, no.1, pp. 1-39, 2010.

[13] R. Chattamvelli "Data Mining Methods", Alpha Science International, Oxford, UK, 2009.

[14] N. S. Altman "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression", *The American Statistician*. vol. 46, no. 3, pp. 175–185, 1992.

[15] S. Russell and P. Norvig [1995]. Artificial Intelligence: A Modern Approach (2nd Ed.). Prentice Hall. ISBN 978-0137903955, 2003.

[16] P. J. Lisboa and A. F. G. Taktak "The Use of Artificial Neural Networks in Decision Support in Cancer: A systematic review", *Neural Networks*, vol. 19, no. 4, pp. 408-415, 2006.

[17] L. Breiman "Bagging Predictors", *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[18] Y. Freund and R. E. Schapire "Experiments with a New Boosting Algorithm", http://www.public.asu.edu/~jye02/CLASSES/Fall-2005/PAPERS/boosting-icml.pdf, 1996, Accessed on: 15.3.2017.

[19] D. Opitz and R. Maclin "Popular Ensemble Methods: An Empirical Study", *Journal of Artificial Intelligence Research*, vol. 11, pp. 169-198, 1999.

[20] E. Bauer and R. Kohavi "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants", *Machine Learning*, vol. 36, no.1, pp. 105-139, 1999.

[21] D. A. Davis, N. V. Chawla N., Blumm, N. Christakis, and A. L. Barabasi "Predicting Individual Disease Risk Based on Medical History", *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 769-778, 2008.

[22] S. T., Moturu, W. G. Johnson and L. Huan "Predicting Future High-Cost Patients: A Real World Risk Modeling Application", *Bioinformatics and Biomedicine, BIBM, IEEE International Conference*, 2007.

[23] D. Mantzaris, G. C. Anastassopoulos and D. K. Lymberopoulos (2008) "Medical Disease Prediction Using Artificial Neural Networks", http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4696782, 2008, Accessed on: 15.3.2017.

[24] P. L. Hebert, L. S. Geiss and E. F. Tierney, M. M. Engelgau, B. P. Yawn and A. M. McBean "Identifying Persons with Diabetes Using Medicare Claims Data", *Am J Med Qual*, vol. 14, no. 6, pp. 270-277, 1999.

[25] W. Yu, T. Liu, R. Valdez, M. Gwinn and M. J. Khoury "Application of Support Vector Machine Modeling for Prediction of Common Diseases: The Case of Diabetes and Pre-Diabetes", *BMC Medical Informatics and Decision Making*, 10(1), pp.1-7, 2010.

[26] W. Zhang, F. Zeng, X. Wu, X. Zhang and R. Jiang "A Comparative Study of Ensemble Learning Approaches in the Classification of Breast Cancer Metastasis", *Bioinformatics, System Biology and Intelligent Computing International Conference*, pp. 242-245, 2009.

[27] Hall J. Guyton and Hall Textbook of Medical Physiology (12th ed.). Philadelphia, Pa.: Saunders/Elsevier. ISBN 978-1-4160-4574-8, 2011.

[28] UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems, Thyroid Disease Data Set, https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease, Accessed on: 9.3.2017.

[29] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning Data Mining, Inference and Prediction, Springer, Second Edition, 2009.

[30] M. Maalouf and T. B. Trafalis "Robust Weighted Kernel Logistic Regression in Imbalanced and Rare Events Data", *Computational Statistics & Data Analysis*, vol. 55, no. 1, pp. 168-183, 2011.

[31] G. E. P. Box "Science and Statistics", *J Am Statist Assoc*, vol. 71, pp. 791-799, 1976.

[32] M. R. Nester "An Applied Statistician's Creed", *Appl Statist*, vol. 45, no. 4, pp. 4001-410, 1996.

[33] E. L. Cohen, C. A. Caburnay, D. A. Luke, S. Rodgers, G. T. Cameron and M. W. Kreuter (2004) "Cancer Coverage in General-Audience and Black Newspapers", *Health Communication*, vol. 23, no. 5, pp. 427-435, 2004.

[34] J. R. Quinlan "Bagging, Boosting and C4.5", *Proceedings of the Thirteenth National Conference on Artificial Intelligen*ce, pp. 725–730, 1996.

[35] P. Horton and K. Nakai "Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier", *ISMB-97 Proceedings*, pp. 147-151, 1997.

[36] C. M. Ma, W. S. Yang and B. W. Cheng "How the Parameters Of K-Nearest Neighbor Algorithm Impact On The Best Classification Accuracy: In Case of Parkinson Dataset", *Journal of Applied Sciences*, vol. 14, no. 2, pp. 171-176, 2014.

[37] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen and W. Newey “Double Machine Learning for Treatment and Causal Parameters”, Cornell University Library, https://arxiv.org/abs/1608.00060, 2016, Accessed on: 15.3.2017.