Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 83 (2016) 1013 – 1018

Procedia Computer Science

The International Workshop on Data Mining for Decision Support (DMDMS 2016)

Pattern classification with imbalanced and multiclass data for the prediction of albendazole adverse event outcomes

Pınar Yıldırım[a]

[a]*Department of Computer Engineering, Faculty of Engineering&Architecture,Okan University,34959,Istanbul,Turkey*

Abstract

Class imbalance problem is one of the important problems for classification studies in data mining. In this study, a comparative analysis of some sampling methods was performed based on the evaluation of four classification algorithms for the prediction of albendazole adverse events outcomes. Albendazole is one of the main medications used for the treatment of a variety of parasitic worm infestations. The dataset was created from the public release of the FDA's FAERS database. Four sampling algorithms were used to analyze the dataset and their performance was evaluated by using four classifiers. Among the algorithms, ID3 with resample algorithm has higher accuracy results than the others after the application of sampling methods. This study supported that sampling methods are capable to improve the performance of learning algorithms.

Peer-review under responsibility of the Conference Program Chairs

Keywords: Imbalanced class,under sampling, over sampling, RBFNetwork, IBK, ID3, Randomtree

1. Introduction

Class imbalance problem is one of the important problems for classification studies in data mining. A data set is defined as imbalanced if one class has significantly more samples than the others. In recent years, the imbalanced problem has highlighted significant interest in many real-life applications in different domains such as fraud detection, medical diagnosis and text classifications[1].

The classification problem for imbalanced data is interesting and challenging to researchers because most standard data mining methods claim their assumption for balanced data but are not applicable for imbalanced one. Researchers have generally addressed two kinds of solutions for data classifications dealing with imbalanced problems: solving in data level by re-sampling, and solving in algorithm level by using design sophisticated

* Corresponding author. Pinar Yildirim Tel.: +90-216-677-1630; fax: +90-216-677-1647.
E-mail address: pinar.yildirim@okan.edu.tr

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2016.04.216

classification approaches, where the prior one is mostly preferred [2].

In this study, a comparative analysis of some sampling methods was performed based on the evaluation of four classification algorithms for the prediction of albendazole adverse events outcomes. Albendazole is one of the main medications used for the treatment of a variety of parasitic worm infestations and it has great interest in medical area. The aim of this study is to make contributions in the prediction of albendazole adverse events outcomes for medical research and present a detailed comparison of popular sampling methods.

2. Classification algorithms

2.1. RBFNetwork

Radial Basis Function (RBF networks) is the artificial neural network type for application of supervised learning problem[3]. By using RBF networks, the training of networks is relatively fast due to the simple structure of RBF networks. Other than that, RBF networks are also capable of universal approximation with non-restrictive assumptions[4]. The RBF networks can be implemented in any types of model whether linear on non-linear and in any kind of network whether single or multilayer [3].The design of a RBFN in its most basic form consists of three separate layers. The input layer is the set of source nodes (sensory units). The second layer is a hidden layer of high dimension. The output layer gives the response of the network to the activation patterns applied to the input layer. The transformation from the input space to the hidden-unit space is nonlinear. On the other hand, the transformation from the hidden space to the output space is linear [5,6].

2.2. IBK

K-nearest neighbour algorithm is called IBK in Weka software. In this algorithm,the training samples are described by n-dimensional numeric attributes. When given an unknown sample, a k-nearest neighbour classifier searches the pattern space for the k training samples that are closest to the unknown sample. The unknown sample is assigned the most common class among its k nearest neighbours[7,8].

2.3. ID3

This is a decision tree algorithm introduced in 1986 by Quinlan Ross. It is used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm. It learns decision trees by constructing them top down that is it is based on the divide and conquer strategy. The tree is constructed in two phases: tree building and pruning. ID3 uses information gain measure to choose the splitting attribute. It only accepts categorical attributes in building a tree model. ID3 does not support pruning. ID3 algorithm is used in knowledge acquisition for tolerance design[9,10].

2.4. Randomtree

Random Tree is a supervised classifier; it is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. In standard tree each node is split using the best split among all variables. In a random forest, each node is split using the best among the subset of predicators randomly chosen at that node. Random trees have been introduced by Leo Breiman and Adele Cutler[11]. The algorithm can deal with both classification and regression problems. Random tree is a collection (ensemble) of tree predictors that is called forest. Each tree produces a classification, and it can be called the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). In case of a regression, the classifier response is the average of the responses over all the trees in the forest [8,12,13].

3. Sampling methods

Inbalanced data set, a problem in real world applications can cause seriously negative effect on classification performance of machine learning algorithms. If a data set is imbalanced, it contains many more samples from one

class than from the other classes. Classifiers can have good accuracy on the majority class but very poor accuracy on the minority classes due to the influence that the larger majority class has on traditional criteria. Most classification algorithms try to minimize the error ratio; the percentage of the incorrect prediction of class labels[14].There are several algorithms and techniques that handle the imbalanced datasets. Sampling approach is one of the main methods to deal with imbalanced data. The sampling techniques are mainly divided into two subgroups: under sampling and over sampling [15].

3.1. Under sampling

Under sampling method removes examples from the majority class to make the data set balanced. This method tries to balance the distribution of class by randomly removing majority class samples. The drawback of under sampling method is that it can discard potentially useful information that could be important for classifiers[15].Under sampling methods are divided into random and informative. Random under sampling randomly eliminates examples from the majority class till the data set gets balanced. Informative under sampling method selects only the required majority class examples based on a pre-specifies selection criterion to make the data set balanced[15].

3.2. Over sampling

Over sampling is a sampling approach which balances the data set by replicating the examples of minority class. The advantage of this method is that there is no loss of data as in under sampling technique. The disadvantage of this technique is it may lead to over fitting and can introduce an additional computational cost if the data set is already fairly large but imbalanced[15].Like under sampling, oversampling is also divided into two types: random oversampling and informative oversampling. Random oversampling is the method which balances the class distribution by replicating the randomly chosen minority class examples. Informative oversampling method synthetically generates minority class examples based on a pre-specifies criterion[15].In summary, over sampling may cause longer training time of over-fitting. The alternative to over sampling is under sampling. This approach is better than over sampling in terms of time and memory complexity. In this study, following algorithms are used for sampling:

Resample: This algorithm produces a random subsample of a dataset, sampling with replacement [16,8].

SMOTE: This algorithm creates artificial data based on the feature space similarities between existing minority examples [17].

Spread Sub sample:This algorithm produces a random subsample with a given spread between class frequencies, sampling with replacement [8].

Stratified Removed Fold: Generates output a specified stratified cross-validation fold for the dataset [8].

4. Related work

There are several studies based on class imbalance problem and sampling methods in the literature. Thammasiri et al., compared different data balancing techniques to improve the predictive accuracy in minority class while maintaining satisfactory overall classification performance. Specifically, they tested three balancing techniques-over-sampling, under-sampling and synthetic minority over-sampling (SMOTE)-along with four popular classification methods-logistic regression, decision trees, neural networks and support vector machines. They used a large and feature rich institutional student data to assess the efficacy of both balancing techniques as well as prediction methods. Their results indicated that the support vector machine combined with SMOTE data-balancing technique achieved the best classification performance on their data [18].

Nguyen et al., presented a study to compare over-sampling and under-sampling techniques in the context of data streaming. They used the ten-fold cross-validation to evaluate sampling techniques on each UCI data set and proposed to use a multiple random under-sampling (MRUS) technique for applications with imbalanced and streaming data. Their experimental results showed that under sampling performs better than over-sampling at smaller training set sizes[19] .

Marcellin et al. proposed to evaluate the quality of decision trees grown on imbalanced datasets with splitting criterion based on an asymmetric entropy measure. They investigated the effect of asymmetric entropy on inbalanced data classification and they found that decision rules derived from a tree grown with an asymmetric entropy are more accurate for predicting the rare class [20].

5. Data description

The data in this study was created from the public release of the FDA's FAERS database by collecting data from DEMO, DRUG, REAC, OUTC and INDI datasets. The data in ASCII format were combined into a database using Microsoft SQL Server 2012. Then, adverse event reports for albendazole were collected from the database to create a dataset for classification. The dataset contains patient demographics such as age, gender, weight, occupation code, reporter country, route, adverse event outcome (class) and adverse event (Table 1). The attributes of the dataset were directly collected from database. The dataset consists of 12899 instances. As seen Table 1. the class distribution of the dataset is imbalanced. The dataset contains 7062(55%) hospitalization (HO), 2455(19%) other (OT), 1595(12%) death (DE), 459(0.4%) disability (DS), 1316(10%) life threatening(LF) ,12(0%)required Intervention to Prevent Permanent Impairment/Damage(RI) and 0(0%) CA- Congenital Anomaly. RI and CA outcomes are defined in the FDA's FAERS. However they have very few instances in our dataset, they were omitted.

Table 1. Characteristics of dataset.

Attribute	**Type**		
Age	*Numeric (*Mean: 29.897 years)		
Gender	*Nominal* (Male, Female,Null)		
Weight	*Numeric* (Mean=67.882 kg)		
Occupation code	*Nominal*		
	MD- Physician		
	OT- Other health professional		
	CN- Consumer		
	PH- Pharmacist		
	Null		
Reporter Country	*Nominal*(7 distinct values)		
Route	*Nominal*		
	Oral		
	Transplacental		
	Ophthalmic		
	Intravenous		
	Topical		
	Parenteral		
	Disc, Nos		
Adverse event outcome	*Nominal*	**N(Number of records)**	**Class ratio(Percentage)**
	HO-Hospitalization	7062	55%
	OT-Other	2455	9%
	DE-Death	1595	12%
	DS-Disability	459	0.4%
	LT-Life threatening	1316	10%
	RI- Required Intervention to	12	0%
	Prevent Permanent	0	0%
	Impairment/Damage		
	CA- Congenital Anomaly		
Adverse event	*Nominal*(220 distinct values)		

6. Experimental results

Albendazole dataset was used to compare different sampling methods for the prediction of adverse event outcomes. Four classification algorithms (RBFNetwork, IBK, ID3, Randomtree) introduced above were selected to evaluate classification accuracy. At first, sampling algorithms were used to dataset and then, classification algorithms were applied to evaluate the algorithms. Respectively, Resample, SMOTE, Spread Sub Sample and Stratified Removed Fold algorithms were used for sampling. Same experiment was repeated for four classifiers. WEKA 3.7.3 software was used. WEKA is a collection of machine learning algorithms for data mining tasks and is an open source software. The software contains tools for data pre-processing, feature selection, classification, clustering, association rules and visualization[21].

Table 2. Evaluation of sampling algorithms.

Classifier	Sampling Method	Precision	Recall	F-Measure	RMSE
RBFNetwork	No sampling	0.6910	0.6620	0.6710	0.2476
	Resample	**0.7060**	**0.6740**	**0.6830**	**0.2450**
	SMOTE	0.6780	0.6630	0.6680	0.2480
	SpreadSubsample	0.6790	0.6670	0.6700	0.2468
	Stratified Removed folds	0.6430	0.6470	0.6440	0.2592
IBK	No sampling	0.6450	0.6500	0.6470	0.2308
	Resample	**0.7070**	**0.6830**	**0.6910**	**0.2301**
	SMOTE	0.6430	0.6490	0.6460	0.2308
	SpreadSubsample	0.6400	0.6480	0.6400	0.2501
	Stratified Removed folds	0.6400	0.6480	0.6400	0.2501
ID3	No sampling	0.6450	0.6500	0.6470	0.2307
	Resample	**0.7080**	**0.6830**	**0.6910**	**0.2300**
	SMOTE	0.6430	0.6490	0.6460	0.2308
	SpreadSubsample	0.6430	0.6510	0.6420	0.2490
	Stratified Removed folds	0.6430	0.6510	0.6420	0.2490
Randomtree	No sampling	0.6450	0.6500	0.6470	0.2307
	Resample	**0.7070**	**0.6830**	**0.6910**	**0.2302**
	SMOTE	0.6430	0.6490	0.6460	0.2307
	SpreadSubsample	0.6400	0.6470	0.6380	0.2508
	Stratified Removed folds	0.6400	0.6470	0.6380	0.2508

Table 3.Performance comparison of classifiers with sampling algorithms.

Classifier	Sampling Method	Execution time(Seconds)
RBFNetwork	No sampling	10.38
	Resample	8.77
	SMOTE	8.99
	SpreadSubsample	19.1
	Stratified Removed folds	0.73
IBK	No sampling	**0**
	Resample	**0**
	SMOTE	**0**
	SpreadSubsample	**0**
	Stratified Removed folds	**0**
ID3	No sampling	0.05
	Resample	0.06
	SMOTE	0.03
	SpreadSubsample	0.05
	Stratified Removed folds	0
Randomtree	No sampling	0.03
	Resample	0.02
	SMOTE	0.02
	SpreadSubsample	0.02
	Stratified Removed folds	**0**

There are many performance measures for the evaluation of the classification results, where TP/TN is the number of True Positives/Negatives instances, FP/FN is the number of False Positives/Negatives instances but some of them are used in this study. Precision is a proportion of predicted positives which are actual positive ($TP/(TP+FP)$). Recall is a proportion of actual positives which are predicted positive ($TP/(TP+FN)$). Precision and recall measures are utilized to find the best method, but it is not easy to make decision. Thus, F-measure was used to get a single measure to evaluate results. The F-measure is the harmonic mean of precision and recall ($2TP/(2TP+FN+FP)$). The comparison analysis by root mean squared error was also performed and described in Table 2. Where n is the number of data patterns, $y_{p,m}$ indicates the predicted, $t_{m,m}$ is the measured value of one data point m and $\bar{t}_{m,m}$ is the mean value of all measure data points [22]. Root Mean Squared Error (RMSE) can be written as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{m=1}^{n} (y_{p,m} - t_{m,m})^2}{n}} \quad (1)$$

Table 2 shows the performance metrics of the classification algorithms with 10-fold cross-validation and sampling algorithms. According to table 2, the highest precision values were obtained for the dataset with ID3 algorithm with resample sampling method. For example, the precision of ID3 with resample is 0.7080 which is the highest value in the Table 2. and has the lowest RMSE with 0.2300. These results highlighted that ID3 with resample is superior to the others. Similarly, IBK, ID3 and Random tree algorithms with resample have the same and highest recall and f-measure values.

The performance evaluation of classifiers with sampling algorithms was performed and the results were obtained in Table 3. The Table 3 revealed that IBK algorithm with sampling algorithms took a short time to classify instances and therefore the performance of this algorithm is better than others.

7. Conclusion

Imbalanced dataset is an important issue in data mining studies and many machine learning algorithms can hardly cope with imbalanced class distribution. Thus, sampling algorithms became a necessity for many studies. In this study, a comparative analysis was performed on the basis of sampling algorithms to predict albendazole adverse events outcomes. Four sampling algorithms were used to analyze the dataset and their performance was evaluated by using four classifiers. The evaluation of results was performed based on accuracy measures and execution time. Among the algorithms, ID3 with resample has higher accuracy results on the dataset than the others after the application of sampling methods. This study supported that sampling methods are capable to improve the performance of learning algorithms and resample algorithm performed better results than the others. The results of this study can make contributions in the prediction of adverse event outcomes in medical research and provide a comparison of sampling methods for machine learning studies.

References

1. Lopez V, Fernandez A, Garcia S, Palade V, Herrera F. An insight into classification with imbalanced data:Empirical results and current trends on using data intrinsic characteristics. *Information Sciences* 2013. 03;50:3-4.
2. Wang K-J, Makond B, Chen K-H, Wang K-M. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. *Applied Soft Computing* 2014; 20: 15-24.
3. Orr MJL. Radial Basis Function Networks. Edinburgh, Scotland.1996
4. Park J and Sandberg IW. Approximation and Radial-Basis-Function Networks. *Neural Computation* 1993*;* 5: 305-316.
5. Haykin S. Neural Networks a Comprehensive Foundation, New Jersey, PrenticeHall 1994.
6. Aruna S, Rajagopalan SP and Nandakishor LV. An empirical comparison of supervised learning algorithms in disease detection. *International Journal of Information Technology Convergence and Services (IJITCS)*2011*;*1; 4.
7. Han J Kamber M. Data Mining Concepts and Techniques.Morgan Kaufmann 200.
8. Witten IH, Frank E. Data Mining: Practical Machine Learning Tool and Technique with Java Implementation. Morgan Kaufmann; 2000.
9. Quinlan JR.Induction of decision trees.*Machine Learning* 1986;81-06.
10. Adeyemo OO & Adeyeye TO. Comparative Study of ID3/C4.5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever. *African Journal of Computing & ICT* 2015; 8;1.
11. Breiman L. Random Forests. Machine Learning 2001:45;5-32.
12. Kalmegh SR. Comparative Analysis of WEKA Data Mining Algorithm RandomForest, RandomTree and LADTree for Classification of Indigenous News Data. *Int. Journal of Emerging Technology and Advanced Engineering* 2015; 5;1.
13. Pfahringer B. Random model trees: an elective and scalable regression method. Working Paper Series, ISSN 1177-777X, 2010.
14. Ganganwar V. An overview of classification algorithms for imbalanced datasets. *International Journal of emerging Technology and Advanced Engineering 2012*; 2(4): 42-47.
15. KrishnaVeni CV, Sobha Rani T. On the classification of imbalanced datasets. *International Journal of Computer Science&Technology* 2011;2:145-148.
16. Ratnoo PS. A comparative study of instance reduction techniques. *Int. Journal of Advances in Engineering Sciences* 2013; 3(3).
17. Haibo H and Garcia EA. Learning from imbalanced data. *IEEE Transaction on knowledge and data engineering* 2009;21;9: 1293-1284**.**
18. Thammasiri D, Delen D, Meesad P, Kasap N.A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* 2014; 41: 321-330.
19. Nguyen HM, Cooper EW, Kamei K.A comparative study on sampling techniques for handling class imbalance in streaming data. SCIS-ISIS 2012, Kobe, Japan, November 20-24.
20. Marcellin S, Zighed DA and Ritschard G. evaluating decision trees grown with asymmetric entropies.*ISMIS* 008: 58-67
21. WEKA: Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka.
22. Kuçuksille EU, Selbas R, Şencan A. Prediction of thermodynamic properties of refrigerants using data mining. *Energy conversion and management* 2011; 52: 836-848.