International Journal of Computer Engineering & Technology (IJCET)
Volume 9, Issue 5, September-October 2018, pp. 109–115, Article ID: IJCET_09_05_014
Available online at
http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=9&IType=5
Journal Impact Factor (2016): 9.3590(Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6367 and ISSN Online: 0976–6375

ENHANCE DECISION TREE ALGORITHM FOR UNBALANCED DATA: RAREDTREE

Pratik A Barot
GEC Gandhinagar, India

H.B. Jethva
GEC Patan, India

ABSTRACT

Unbalanced data classification is important when misclassification rate of rare instances is huge. Medical diagnosis field is an example. Existing techniques of unbalanced data classification are based on sampling techniques which suffer from overlapping and increase learning time. To develop effective intelligent system for domain of unbalanced data, minority example should be classified with good accuracy. But traditional machine learning algorithms lack this features and they are biased towards majority class. We proposed new optimal algorithm based on decision tree algorithm. We modified the decision tree algorithm and developed new algorithm called RareDTree which classify minority instances with good accuracy without compromising the accuracy of majority class instances. RareDTree also eliminate the need of data sampling.

Key words: Unbalanced data classification, decision tree, RareDTree, Machine Learning.

Cite this Article: Pratik A Barot and H.B. Jethva, Enhance Decision Tree algorithm for Unbalanced Data: RareDTree. *International Journal of Computer Engineering and Technology*, 9(5), 2018, pp. 109-115.
http://www.iaeme.com/IJCET/issues.asp?JType=IJCET&VType=9&IType=5

1. INTRODUCTION

Data mining is knowledge extraction dataset [10]. Knowledge discovery is the process of analyzing data and summarizing it into knowledgeable information through knowledge representation techniques. Such knowledge used to increase organization benefit [11].

As per J Han et. al. [10], classification is used for label prediction. Classification is of two types; supervised and unsupervised. Supervised classification has two phases; learning and testing. In learning phase classifier model is prepared from pre-labeled training dataset. In testing phase constructed model used to test dataset and its accuracy is measure. If built classifier shows good accuracy then it used to classify further dataset. In unsupervised learning no pre-classified dataset is used, example is clustering.

1.1. Imbalanced Dataset

Imbalanced dataset divided into two categories: - Between-class imbalanced and Within-class imbalanced [16]. Between-class imbalanced dataset there is unequal distribution between classes. Imbalanced dataset which has only two classes is called binary-class imbalanced data. In multi-class imbalanced data there are more than two classes with uneven distribution Class with comparatively less number of instances called minority class or rare class [15].

As per Jerzy Stefanowski et al [16], between-class imbalance is evaluated by global imbalance ratio (IR). Some of datasets are highly imbalanced while some are little imbalanced. If dataset has little imbalanced then sampling techniques can be applied and found useful. But in highly imbalanced dataset, sampling become more complex and also increases dataset size or may remove some important information. This restricts the use of sampling method for absolute rare class [16].

Rare class classification is the data mining task for building a model that can correctly classify unbalanced data. [15].

High imbalance data is common in real-world domain. Therefore, researcher paid good attention on classification of imbalanced data and proposed different techniques for binary class imbalanced dataset [4]. However, classification becomes even more complicated and less accurate if dataset is multi-class imbalanced [7]. Traditional classification algorithms with accuracy close to 90% favor majority class and incorrectly classifying minority class instances. This high accuracy percentage is due to accurate classification of majority class instances.

In *within-class imbalance* dataset, target class contains several sub-concepts [7]. In addition to IR, other characteristics of data like small disjunction, overlapping of classes and borderline example need to be taken care for accurate classification. Overlapping and small disjunction makes cluster and outlier analysis techniques unreliable for unbalanced classification.

There are lots of researches have been performed in area of unbalanced classification. Most of researcher uses data sampling techniques to handle minority example. SMOTE algorithm and its different versions are widely used for data level sampling for unbalanced dataset [8]. However, SMOTE uses same sampling rate for all instances of minority class which is result into a poor classification [9]. Based upon data level sampling numerous algorithms have been proposed using decision tree, Bayesian network, support vector machine (SVM), Neural network etc [5, 12, 13, 14].

The main drawback with SVM is its high computational cost [6]. Decision tree algorithm is somewhat greedy and uses top down divide and conquer method [2]. Decision tree uses information gain parameter which has natural bias towards majority class instances.

1.2. Decision Tree Algorithm

Decision tree algorithm builds a tree from training dataset. There are different versions of decision tree algorithm available like ID3, CART etc. [10]. ID3 algorithm uses information gain as attribute selection parameter. It selects the attribute which divide the given training dataset into the subsets. Selection of dividing attribute is based on maximization of information gain. Attribute selected in first iteration is used as root element of a tree. Based on that root attribute original training dataset is divided into subsets. In subsequent iteration this process is repeated for each subset and tree is built.

Because of the use of Information gain, Gain ratio or Gini index like parameter decision tree algorithm mostly concentrate on majority class instances and build a tree which classify majority class instances with good accuracy [2]. Due to small representation of minority class

instances they remain hidden in surrounding of majority class instances. Our new algorithm for decision tree based rare class classification called RareDTree alleviates this drawback by building a separate branch for minority class instances.

2. RAREDTREE ALGORITHM

RareDTree algorithm is modified decision tree algorithm. In this algorithm first decision tree is build using traditional decision tree algorithm like ID3. Then using causal relationship extraction process as proposed by Pratik Barot et al. [1], responsible causes of minority class instances are extracted. From this extracted causes CausalTree is built. In final step, CausalTree is merged with decision tree build by ID3 in first step. The CausalTree, which is now sub-tree of decision tree is responsible for accurate classification of minority class instances.

Algorithm: CausalTree
Input: Dataset (D), list of minority class (MC)
Output: CausalTree (CT);
Begin
    R=CausalRelationshipExtraction(D,MC)
    CT=Build_CausalTree(R);
    Return(CT);
End

Sample of causal relationship for two classes (Hypo and Hyper) of new-thyroid dataset is listed in TABLE-1.

Table 1 Sample Causal Relationship for New-Thyroid Dataset

Class	Causal Relationship
Hypo	thyroidstimulating='(5.73-33.38]'
Hypo	thyroxin='(-inf-2.98]'
Hyper	thyroxin='(22.82-inf)' triiodothyronine='(1.6-inf]'
Hyper	T3resin='(-inf-88.7]'
Hyper	thyroxin='(17.86- inf]'

Algorithm: RareDTree
Input: Dataset (D),
Output: T (Classification Model - Tree)
Begin
    CT = CausalTree(D);
    DT = ID3(D);
    T = TreeMerge(DT, CT);
    Return T ;
End

We used Weka 3.9 for implementation. Fig.1 shows decision tree mined for New-Thyroid dataset. Fig.2 shows one of the sub-tree which is generated from CausalTree algorithm for Hypo class of New-thyroid dataset.

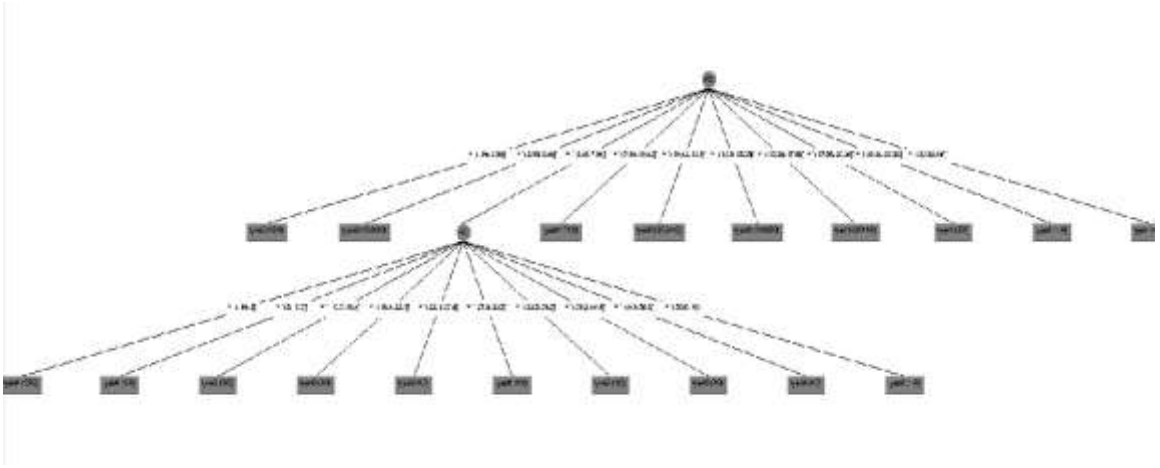

Figure 1 Decision Tree for New-Thyroid

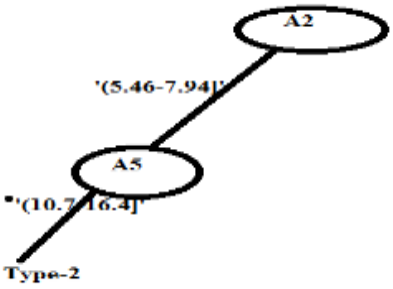

Figure 2 Causal Tree for minority-class Type-2 (Hypo)

RareDTree generates many such branches for different strong causal relationship between target class and features. In next phase, tree merge algorithm merges decision tree of Fig.1 and causal tree of Fig.2 and generate new enhance decision tree.

3. RESULT ANALYSIS

We have used four datasets from KEEL repository [3] for performance evaluation. TABLE-2 shows dataset description.

Table 2 Dataset Description

Sr No.	Name	#instance	IR ratio	Description
1	Hayes-roth	132	1.70	Class distribution:- Type0 and Type1 has 51 instances each and Type2 has 30 instances.
2	New-Thyroid	215	4.84	An imbalanced version of the Thyroid Disease (New Thyroid) data set, with 3 classes: hyper and hypo are rare class. **hypo is smallest class with 30 instances and normal is largest class with 150 instances.**
3	Ecoli	336	71.50	An imbalanced version of the Ecoli data set. Class distribution:- cp=143 | im=77 | pp=52 imU=35 | om=20 |omL=5|imL=2 |imS=2
4	Pageblocks	548	164.0	type1=492 | type2=33 type3=3 | type4=8 type5=13

Fig. 3 shows classification result of traditional decision tree algorithm for new-thyroid dataset. As shows in fig.3, the highlighted portion shows misclassification of type-2(Hypo) instance as instances of type0 (normal). In new-thyroid dataset, type0 class is majority class and type2 is minority class.

Figure 3 Classification Result of Decision Tree.

Such misclassification is due to the biasing of decision tree towards the majority class. RareDTree algorithm removes such biasing by developing separate sub-tree for minority class instances which exhibits unique causes. Fig.4 shows result of RareDTree. As shown in figure, the misclassification is not there in result and it correctly classifies the instance of type2 as type2 due to additional branch of minority class instances.

Figure 4 Classification Result of RareDTree.

TABLE-3 lists the accuracy of J48, SMOTE sampling, and RareDTree. As shown in table, RareDTree outperforms traditional decision tree algorithm and sampling technique. Our proposed algorithm shows best result without data level sampling.

Table 3 Performance(Accuracy) Comparison

Dataset	J48	SMOTE Accuracy	RareDTree Accuracy
Hayes-roth	71.21	76.54	80.06
New-Thyroid	91.62	95.91	96.53
Ecoli	72.91	87.84	90.13
Pageblocks	90.32	81.92	87.81

Fig.5 shows comparison in form of line chart. Except Pageblocks dataset, for all other datasets RareDTree algorithm outperform other algorithms.

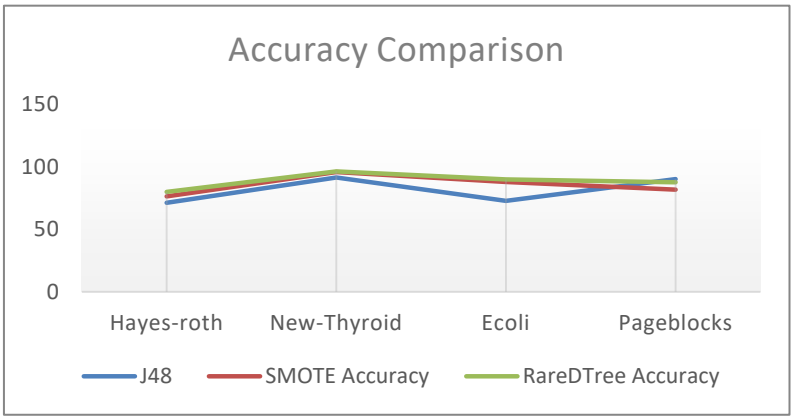

Figure 4 Accuracy Comparison

4. CONCLUSIONS

Unbalanced data classification is important when minority class is more important than majority class. Traditional algorithm biased towards majority class due to their inherent nature. We present new improve algorithm which is based on decision tree algorithm. RareDTree algorithm develops special branch for minority class instances and then merge it with the main decision tree. Using RareDTree we remove the need of data sampling and thus removes drawback associated with is as well. Our future work is to test performance of RareDTree algorithm for other applications and on large datasets.

REFERENCES

[1] Pratik A Barot, H B Jethva, Statistical Study to Prove Importance of Causal Relationship Extraction in Rare Class Classification, In: proceeding of International Conference on Information and Communication Technology for Intelligent Systems (ICTIS '17), Springer Series: Smart Innovation, Systems and Technologies-2017

[2] C A Ratanamahatana, D Gunopulos,Scaling up the Naïve Bayesian Classifier: Using Decision Trees for Feature Selection, Appl. Artif. Intell. 2003.

[3] KEEL dataset repository, Feb-2017, http://sci2s.ugr.es/keel/datasets.php

[4] Nitesh V. Chawla, Data Mining and Knowledge Discovery Handbook, Chapter-40, Data Mining for Imbalanced Datasets: An Overview.

[5] Jose A. Saez, Julian Luengo , Jerzy Stefanowski , Francisco Herrera, SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, Information Science, Elsevier, Aug-2014.

[6] Mehmait Sait Vural, Mustafa Gok, Criminal Prediction using Naïve Bayesian Theory, Neural Comput. & Applic. Springer, Feb-2016.

[7] Astha Agrawal, Herna L Viktor, Eric Paquet SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling, In: Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR, pages 226-234, SCITEPRESS – IEEE explore, 2016.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[9] Kun Jiang, Jing Lu, Kuiliang Xia, A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE, Springer, May-2016.

[10] Jiawei Han, M Kamber, J Pei, Data Mining Concepts and Techniques; Third Edition,Elsevier 2012.

[11] Varsha Mashoria, Anju Singh, "Literature Survey on Various Frequent Pattern Mining Algorithm", IOSRJEN, Vol-3, Jan-2013.

[12] Zhongbin Sun, Qinbao Song, Xiaoyan Zhu, Heli Sun, Baowen Xu, Yuming Zhou, A Novel Ensemble Method for Classifying Imbalanced Data, Pattern Recognition, http://dx.doi.org/10.1016/j.patcog.2014.11.014 - Elsevier-2014.

[13] Randa Oqab Mujalli, Griselda Lopez, Laura Garach, Bayes Classifiers for Imbalanced Traffic Accidents Datasets, Accident Analysis and Prevention, Elsevier, Dec-2015.

[14] S. Garcıa, F. Herrera, Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy, Evolutionary Computation 17 (2009) 275–306

[15] Kittipong Chomboon, Kittisak Kerdprasop, Nittaya Kerdprasop, Rare Class Discovery Techniques for Highly Imbalanced Data, in: proceeding of the International MultiConference of Engineers and Computer Scientists, Hong Kong, 2013.

[16] Jerzy Stefanowski, Dealing with Data Difficulty FactorsWhile Learning from Imbalanced Data, Springer International Publishing Switzerland, 2016.