ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 19, April 2016

ANALYSIS OF IMBALANCE DATASET USING CLASSIFICATION TECHNIQUES

[1]S. Sinduja,[2] Dharmendra Singh Rajput
School of Information Technology & Engineering,VIT University, Vellore (TN), India
[1]sinduharini19@gmail.com,[2] dharmendrasingh@vit.ac.in

Abstract—**An enormous amount of data is being collected and stored in databases everywhere across the world. These data bundles up and keep on increasing every year. Extracting information that are hidden in such databases and classifying that information extracted are most important tasks in data mining. If such datasets are imbalanced, then it becomes tough to handle it. Since Predicting future is one of the fundamental tasks in data mining. Working with imbalance datasets to predict the possible outcome is a very tedious task. The dataset is imbalanced when it is not classified correctly, when one class holds more instances than other. They are often represented as a positive class (minority) and negative (majority) class. The class that has less number of samples is called minority class, and one that has more is called majority class. Imbalance dataset causes many serious issues in data mining, mostly the standard classification algorithm considers the dataset as balanced which in turn is partial towards majority class. For applications like medical diagnosis, this causes a very serious effect. Hence balancing dataset is critical for many real-time applications. In this study, we will analyze various classifications techniques and algorithms that are used for balancing datasets like Sampling technique, K-nearest neighbour, Naive Bayesian, Decision Tree, SMOTE and an efficient method to handle imbalance datasets using classifier ensemble method is proposed. Ensemble based method are a combination of various classifiers which is an ensemble of new approach with other existing data and algorithmic approaches. Here we propose a new method that is the combination of new decision tree algorithm (which uses combined splitting criteria and distance measure approach to choose best splitting feature) and naive Bayesian approach. Since it combines advantages of many classifiers, this new approach outperforms previous methods.**

Keywords- *Imbalanced Datasets, Classification, Classifier, Sampling, KNN, SMOTE, Naive Bayesian, Decision Tree*

I. INTRODUCTION

Data mining is the process of extracting or discovering hidden and useful information from the database [1].Data mining methods used for analyzing and summarizing data in meaningful and understandable ways [2]. Data mining divided into two tasks. They are descriptive and predictive tasks. These two approaches used for data mining to uncover hidden information from a large amount of data or dataset. Predictive task predicts the future value of an attribute by using the values and results obtained from another data or attribute [3]. An example of a predictive task in the medical department is deciding whether a patient is suffering from a particular disease or not based on the diagnostic test results obtained from analyzing symptoms and other conditions. The descriptive model provides derivative patterns by analyzing past outcomes and gives abstract of underlying relationship between data [4]. An example of the descriptive model, In electrical companies to predict the optimal electric cost, the history of power usage can be examined and analyzed to make a plan. The descriptive task involves Clustering, Association Rule Discovery, Sequential pattern, summarizations. Predictive tasks are Classification, Regression and Deviation detection.

Clustering is the process of grouping objects that are similar into clusters. The training set not used in clustering hence it is called unsupervised learning [27].Marketing strategy is a good example of clustering. In this datasets of customers are split into sub-datasets based on similar characteristics, using this information a product can target to particular customers. Associative Rule discovery is a descriptive technique that analysis pattern based on the occurrence of objects. Sequential pattern discovery predicts pattern based on the occurrence of events and its sequence. Regression method predicts the value of one variable by considering the value of other variables [15]. A dataset is imbalanced when some samples in one class are more than the other. It often referred as minority and majority class. The class that has more samples are called majority class and the class that contains less number of samples called minority class.It is tough to learn from imbalance dataset. Imbalance dataset causes many issues. One of the main problems is the class that has less instances(minority class) represents the concept of interest in most cases[6] for example in the medical field, patients suffering from illness or disease may be minority class, and majority class represents healthy patient. In such case missing an illness patient who is the positive class (minority) but classified negative (false negative) leads to the serious problem. Thus, imbalance problem causes the serious effect in many real world applications.

Classification techniques classified as data level, algorithm level, cost sensitive level, feature selection level and ensemble level [8]. In data level approach datasets are balanced in preprocessing stage [8]. Sampling methods are data level approach which will be discussed briefly in literature review section. The other classification algorithms are Decision Tree, SVM, K-nearest neighbour, Bayesian classifier, artificial neural network, fuzzy rule-based classification and regression trees [9]. Decision tree method is used to develop a tree like structure that has leaf and decision nodes. The decision nodes represents attributes in data set which are further expanded using splitting criterias [24]. More about decision tree and splitting criteria are discussed in later sections. K-nearest neighbour algorithm known as the lazy learning algorithm. In this algorithm, the object is classified based on the neighbours associated with it. The object belongs to that category its majority belongs to which is decided by distance similarity [11]. For example, if it has five neighbours three neighbours belongs to a class A and two belong to other class B then this new object belong to class A.If it has the single neighbour when k=1 then it belongs to that category. KNN calculates distances using Euclidian distance for numerical value and Hamming distance for categorical attributes [9]. Naïve Bayesian classifier method is a statistical method for classification. It based on Bayes theorem which is named after Tomas Bayes. It uses Bayesian formula and calculates the probability to assume predictor value. Naive Bayesian though it's simple it outperforms other classification methods [12].

II. LITERATURE REVIEW

Class imbalance is a major problem which has faced by many real world application fields and domains. Several approaches and algorithms have been developed and proposed to overcome this issue. Techniques that used for solving imbalance datasets classified as Algorithm level, Data level, Cost-Sensitive level, Future selection level and Ensemble level [8]. At Data Level Approach, to Balance the dataset, some instances are either removed from or added to the majority and minority classes respectively [13]. According to Chawla, In data level approach, datasets are put together before they trained by classifiers. Hence, it is easier to balance as compared to algorithm level approach (Chawla 2010). Sampling techniques are used to balance the imbalance datasets at the data level, which represented as two types, Over Sampling and under Sampling. Sampling techniques which involve reducing majority class samples by removing some instances called Under-Sampling. When reduction of Samples done with some knowledge it is known as informed under sampling [14], when performed randomly then it is called random under sampling [16]. The oversampling method adds some instances by generating minority class samples, so in original dataset essential information gets included which in turn improves the performance of classifier [17]. Hence, Oversampling methods are useful than Under sampling..

The Sampling methods are further Categorize as many methods based on the generation of synthetic samples. Some of the methods are Synthetic Minority Oversampling techniques (SMOTE) [18]. Adaptive Synthetic Sampling Technique (ADASYN)[19] and Borderline SMOTE[20]. Data level approaches are sometimes better when compared with algorithm level, but both techniques have many disadvantages. The main drawback of the under-sampling technique is the loss of useful data, since some data are removed randomly from majority class. Whereas in Oversampling, adding new instances to minority class induces overfitting problem [21]. The accuracy of Classification has been affected because of this issue. In algorithm level, existing algorithm is modified to recognize instances in minority class[22]the main drawback is it depends on classifiers and is sometimes difficult to handle. Decision tree method is another classification technique which is used to build classification model as a tree structure. It uses divide and conquer rule and divides datasets into subsets as child nodes. The best predictor attribute made as root node [23].

Decision tree algorithm is mainly of three types as ID3, CART and C4.5. These three algorithms have its issues and advantages. Disadvantages of id3 are Over classification, over fitting problem, missing values are not properly handled, not suitable for large examples. CART may sometimes result in unstable decision tree and can only splits one variable at a time.C4.5 may develop branches that contains nodes with zero values and can result in complex tree which is bigger It is based on the application to decide which one best suited. The performance mainly depends on splitting criteria used. Brieman analyzed various splitting criteria Shannon entropy, Gini impurities and Twoing and briefly discussed its effect on classification [24]. Further analysis on splitting criteria done by Deitterich and according to his analysis splitting criteria has the huge effect on the performance of the tree [25]. Ensemble classifier is another classifier technique which also known as multiple classifiers. It improves the performance of single classifier by combining other classifiers, and the new classifier obtained which outperforms all other classifiers [26]. One such ensemble method is proposed in this study.

III. METHODOLOGY ADAPTED

In this study, we proposed an Ensemble method which is the combination of many classifiers techniques. An ensemble approach integrates advantages of many classifiers and thus produces better performance than other classifiers. In this study, we propose a decision tree algorithm by analyzing different splitting criteria that used for ID3, CART and C4.5 and combining these splitting criteria to find best splitting attribute. The splitting criteria used are Shannon entropy used in ID3 and C4.5 decision tree and Gini impurity measure which is used in CART.

The splitting criteria used for ID3 and C4.5 is Shannon entropy: Shannon entropy for target class X=x and feature Y is

$$H(Y/x) = -\sum_{y} p(y/x) \log p(y/x) \text{..................(i)}$$

Where Y represents the distinct values count and X the count of a total number of values of the target attribute. Consider Play tennis dataset, in which play game is target attribute the values are yes and no. If there are two distinct values for target attribute play game, Then number of yes is 5 and number of No are 3 then Shannon entropy is calculated for positive class as p(5/8) and P(3/8) .

Gini impurity measure:

$$Gini(Y/x) = \sum_{y} p(y/x)(1 - P(y/x)) \text{......(ii)}$$

Gain ratio (S, A)=

$$\frac{Information\ gain(S,A)}{Splitting\ information(S,A)} \text{..(iii)}$$

The new gain ratio formula applied after calculating splitting criteria, where split information is best splitting criterion which is calculated using splitting criteria formula

Classification rules to select best splitting criteria are

1. Select best splitting attribute based on splitting criterion
2. Split input data based on the splitting attribute values
3. Repeat the process recursively until the subsets outcome is all positive or all negative i.e. pure subset.

Drawback in using Information gain:

The best splitting attribute is chosen in id3 based on information gain values. The attribute that has highest information gain is selected as the root node and based on the values of chosen attribute again another attribute that is closer to root value was chosen; thus, the decision tree is developed by recursively deciding best attribute to split. Information gain used by Id3 may classify the data perfectly, but it is biased towards attribute that has more values than those that has lesser. An attribute may have many possible values, so the information gain value will be high, but that particular attribute may turn out to be a poor predictor[28]. To overcome the drawback faced in using information gain. It improved by calculating Gain ratio,

$$\text{Gain ratio(S, A)} = \frac{Information\ gain(S,A)}{Split\ information(S,A)} \text{......(iv)}$$

$$\text{Split information(S, A)} = \sum_{i=1}^{c} \frac{Si}{S}\ log_2\ \frac{Si}{S} \text{....(v)}$$

Where S represents the subset, Si values range from 1 to c .C is some distinct values for attribute A. Split information is entropy value on the values of attribute A.Id3 algorithm information gain entropy value is calculated on target attribute.

Drawback in using Gain ratio: The main drawback of using gain ratio is the denominator value may be zero or very small[28].

Using Distance based measure with splitting criterion:

In order to overcome the drawback faced by using gain ratio and split information Distance based measure introduced Lopez de Mantaras can be used to improve performance and to produce small decision tree as it considers Distance values of attributes that are closer[28]. Distance Measure Attribute

$$\text{(DMA)} = \frac{Gini\ index}{Information\ gain\ value} \text{....(vi)}$$

Thus, a best splitting criteria is derived using all splitting criterion formulas, since it eliminates the drawbacks of each splitting criterion. Best splitting attribute is chosen to develop decision tree which is helpful to derive future outcomes. Ensemble classification technique combines new algorithm with existing classification method. Here, On top of the new decision tree that is developed using combined formula and the new gain ratio is obtained and the decision tree is generated then the performance of classifier and time consumption is analyzed then ensemble approach developed by combining the decision tree with the naïve Bayesian approach.

```
ALGORITHM 1:
Input: Training data (attributes A1, A2….An and Target class)
Output: Decision
Tree START
Step1: Node-> Select best attribute An which gives closer gain ratio based on splitting criterias and DMA
Step 2: If (subset is pure (achieving all yes or all no))
            Then Stop growing
        Else
            Select best attribute An which gives
        closer gain ratio based on splitting criterias and
        DMA
Step 3: For each child node repeat step 1 and
2. END
```

Algorithm 2: Procedure of ensemble classification technique

START
Step 1: Load dataset into the database.

Step 2: Split data from the database into two subsets i.e. subset1 and subset2, then calculate number of positive class and negative class that present in the database.

Step 3: Calculate Information Gain using equation 1. Step 4: Calculate Gain Ratio using equation 4 and 5. Step 5: Calculate Gini Index using equation 2.

Step 6: Make decision to choose which attribute is best attribute to split based on DMA equation 6.
Step 7: Develop decision tree using Algorithm 1.

Step 8: On top of the decision tree we develop using algorithm 1 naïve Bayesian method using Bayes theorem i.e.
Bayesian formula P(c|x) = P(x|c) P(c)/ P(x)

where P(c|x) is posterior probability(x|c) is likelihood of predictor(c) is class prior probability(x) is predictor prior probability .

Step 9: Improved Decision Tree. END

Here in this study, a new decision tree is applied first and after analyzing the performance of ensemble classifier To further improve this study naive Bayesian classification technique is applied which uses Bayesian formula and calculates probability and likelihood, naïve Bayesian outperforms all other classifiers [12] and results are analyzed which proves that since it combines many classification techniques it eventually improves performance than other classifiers.

IV. DISCUSSION AND RESULTS

Combining Decision tree with another classification improves efficiency, because decision tree approach has many advantages, Such as it is simple to understand, it can handle both numerical and categorical data and is Robust. Hence using other existing classification methods over the decision tree further improves performance. Moreover the issues faced by all other existing decision trees Id3,CART and c4.5 are analyzed and to overcome all those issues the best splitting criteria should be chosen. Various splitting criterion which includes Shannon entropy, Gain ratio and Gini are combined and modified using distance measure attribute. Hence ,the modified splitting criterion results in selecting best splitting attribute which eventually results in improved decision tree which is small and effectively predicts future outcome.

Combining this modified decision tree with naive Bayesian method and other classification increases classifiers performance and the time consumption of the proposed algorithm evaluated for different datasets. The data sets used here are extracted from UCL machine learning repository, KEEL and are modified.

Table 1 denotes comparison of time consumption of existing id3 algorithm and proposed algorithm. The analysis and comparison study made by Ally Leung [29] gives time in milliseconds of existing id3 algorithm which is compared with our proposed work done in this study.

Dataset	Records	ID3 existing algorithm (ms)	Proposed Algorithm
1	460	30.3	23.0
2	1152	41	34.0
3	1728	56	35.1

Table 1: Comparison of Execution time between ID3 and Proposed algorithm

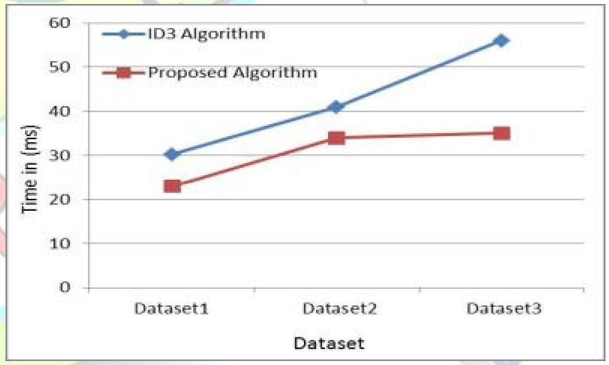

Figure 1: Comparison of Execution time between ID3 and Proposed algorithm in different datasets

Figure 1 represents the time consumption of different datasets. X axis represents time in milliseconds and y axis different datasets. Blue line denotes time consumption of existing id3 algorithm and red denotes the time consumption of the proposed approach which clearly shows that time consumption is less than the existing one as it reduces decision tree by selecting only necessary attributes that are needed to predict future values.

V. CONCLUSION

In this study, we have analyzed various classification techniques for imbalance dataset and developed ensemble approach which involves developing a proposed decision tree algorithm and Naive Bayesian method applied on top of it which results in efficient and reduced decision tree. The decision tree is developed by combining different splitting criterias like Information gain, Gain ratio, Gini impurity measure and best splitting attribute is selected by using Distance based measure along with different splitting criteria. Since selecting splitting attribute plays important role in performance of decision tree. The improved decision tree is efficient in predicting the class outcome from the dataset and takes less time in producing the outcome. Thus combining these classification techniques improves the classifier performances and overcomes certain issues faced by other existing methods discussed earlier in this paper.

REFERENCES

[1] Osmar R Zaiane,"Principles of Knowledge Discovery in Databases", University of Alberta,1999.

[2] Ranshul Chaudhary, Pradheep Singh, Rajiv Mahajan, "A Survey on Data mining techniques", International Journal of Advanced Research in computer and Communication engineering, Vol 3 No 1, pp. 5002-5003, 2014.

[3] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, "Knowledge discovery and data mining", International conference on knowledge discovery and data miming,1996.

[4] Fadzilah Siraj,Mansour Ali Abdoulha,"Mining Enrolment Data Using predictive and descriptive Approaches",InTech,Knowledge-Oriented Applications in Data Mining,pp.53-72,2011.

[5] Christo Ananth, S.Esakki Rajavel, S.Allwin Devaraj, P.Kannan. "Electronic Devices." (2014): 300.

[6] N.V.chawla,N.Japkowicz and A.Kolcz,"Special issue learning Imbalanced datasets",SIGKDD Explor.Newsl,Vol 6,No 1,2004

[7] Victora Lopez, Alberto Fernandez,Salvador Garcia,Vasile Palade,Francisco Herrera," An insight into classification with imbalanced data:Empirical results and current trends on using data intrinsic characteristics",information sciences,pp.113-141, 2013.

[8] Dr.D.Ramyachitra,P.Manikandan,"Imbalanced Dataset Classification and solutions: A Review",IJCBR, vol 5,No 4,2014.

[9] Megha Gupta,Naveen Aggarwal,"Classification Techniques analysis",National Conference on Computational Instruction, pp.19-20, 2010 .

[10] Lili Diao, Chengazhong Yang and Hao Wang, "Training SVM email classifiers using very large imbalanced dataset" , Journal of Experiment &Theoretical Artificial intelligence, Vol. 24, No 2, pp. 193-210,2012.

[11] Oliver Sutton, "Introduction to K Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction" ,2012

[12] Dr.Saed Sayad ,"http://www.saedsayad.com/naivebayesian.htm ", 30th dec 2015.

[13] Piyasak Jeatrakul,Kok wai Wong," Enhancing classification of Multi Class Imbalanced data Using the OAA-DB Algorithm",IJCNN 2012.

[14] X.Y.Liu,J.Wu,and Z.H.Zhou,"Exploratory Under Sampling for Class Imbalance Learning",proc.Int'l conf.Data mining,pp.965-969,2006.

[15] Tan,Steinbach,Kumar,"Introduction to Data Mining",18th April 2004

[16] J.Zhang andI.Mani,"KNN Approach to Unbalanced Data Distributions:A Case Study Involving Information Extraction",Proc.Int'l Conf.Machine Learning,Workshop Learning from Imbalanced Datasets,2003.

[17] Date Shital Maruti,"Minority oversampling Technique for Imbalanced data",International Journal of Scientific and reasearch publications,Vol.5,Issue 4,April 2015.

[18] N.V.Chawla, K.W.Bowyer, L.O.Hall and W.P.Kegelemeyer," SMOTE: Synthetic Minority Oversampling Technique",J.Artificial Intelligence Research,vol 16,pp.321-357,2002.

[19] H.He,Y.Bai,E.A.Garcia and S.Li,"ADASYN:Adaptive Synthetic Sampling Approach for Imbalanced Learning,"Proc.Int'l Joint Conf Neural Networks,pp.1322-1328,2008.

[20] H.Han,W.Y.Wang,and B.H.Mao,"Borderline-SMOTE:A New Oversampling Method in Imbalanced Data Sets Learning,"Proc.Int'l Conf.Intelligent Computing,pp.878-887,2005.

[21] Chris Seiffert,Taghi M Khoshgoftaar.Jason Van Hulse, Amri Napolitano,"RUSBoost:A Hybrid approach to Alleviating Class Imbalance",IEEE Transactions on Systems, man and Cybernetics-Part A:Systems And Humans,Vol 40,No 1,2010.

[22] M.Mahdizadeh,M.eftekhari,"Designing fuzzy Imbalanced Classifier based on the Subtractive clustering and genetic Programming", In

Proceeding of:13th Iranian Conference on Fuzzy Systems(IFSC),pp.2-6,2013.

[23] Dr.Saed Sayad, ” http://www.saedsayad.com/decision_tree.htm “,30 dec,2015.

[24] Sonia Singh,Priyanka Gupta,”Comparative Study ID3,CART and C4.5 Decision tree Algorithm:A Survey”, International Journal of advanced Information Science and Technology,Vol 27,No.27,2014.

[25] L.Breiman,”Technical note: Some properties of splitting Criteria in Machine learning”,Vol 24, pp.41-47,1996.

[26] Lior Rokach,”Ensemble- based Classifiers”,Artif Intell Rev 2010,
33:1-39 DOI 10.1007/s10462-009-9124-7.

[27] Jiawei Han,Micheline Kamber,"Data Mining Concepts and Techniques,2nd edition,The Morgan Kaufmann Series in Data Management Systems”, Jim Gray,Morgan Kaufmann Publishers, March 2006.ISBN 1-55860-901-6.

[28] Thomas M. Mitchell,”Machine learning: chapter 3 Decision Tree learning”,MCGraw-Hill,pp.73-74,1997.

[29] Ally Leung ,"Multimedia, Communication and Computing
Application”, Proceedings of the 2014 International Conference on
Multimedia, Communication and Computing Application (MCCA 2014),CRC press,pp 260,2014.