ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

An Efficient Identification of Malnutrition with Unsupervised Classification Using Logical Decision Tree Algorithm

Aruna.S, Sudha.P

M.Phil. Research Scholar, Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, India

Assistant Professor, Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, India

ABSTRACT: Malnutrition is a quiet pandemic affecting millions of people all through the world. It has high community Health significance because of the complete number of people affected by malnutrition, the detail that susceptible populations disproportionately suffer the effects of malnutrition, and because these effects are severe, long lasting, and cumulative. In this the work presents Models based on the logical decision tree (LDT) algorithm showed the highest predictive capabilities with respect to recall and models based on the Decision Trees algorithm with low pruning had the highest precision. The proposed algorithm based on two functions namely bagging and logical decision rule. The Bagging based approach increasing the number of minority class instances by their replication and the Logical decision tree (LD) method which is used to find the covariation along with conjunction 0 to 1. Thus the result is obtained with good ensemble classification quality.

KEYWORDS: Data mining, Malnutrition, Decision Tree, Ensemble Approach.

I. INTRODUCTION

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large database. The patterns must be actionable so that they may be used in enterprise's decision making process [2]. Data mining or knowledge discover in databases (KDD) is a collection of exploration techniques based on advanced analytical methods and tools for handing for large amount of information. The techniques can find novel patterns that may assist an enterprise in understanding the business better and in forecasting. Many data mining techniques are closely related to some of the machine learning techniques that have been developed over the last 40 years. Others are related to techniques that have been developed in statistics, sometimes called exploratory data analysis. These techniques were developed some time ago and were designed to deal with a limit amount of data. The techniques have now been modified to deal with large amounts of data [1]. Decision tree which are used to predicate categorical variables are called classification tree because it plays instances in categories. Decision tree used to predicate continuous variable are called regression tree. Some of the decision trees are ID3, C5.0, Quest, CART and CHAID. The main advantages of decision tree are execution efficiency mainly due to its simple and economical representation and its ability to perform. Neural network is defined as a data processing system consisting of large number of simple highly interconnected processing elements in an architecture inspired by the structure of brain. There are three types of ANNs are Single layer feed forward network, Multi layer feed forward Network and recurrent network.

II. RELATED WORK

2.1 Insight into Data Mining Theory and Practice [2]: The data mining is an emerging technology that has made its way into science, engineering, commerce and industry as many existing inference methods are obsolete for dealing with massive datasets that get accumulated in data warehouses. This comprehensive and up-to-date text aims at providing the reader with sufficient information about data mining methods and algorithms so that they can make use of these methods for solving real-world problems. The authors have taken care to include most of the widely used

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

methods in data mining with simple examples so as to make the text ideal for classroom learning. To make the theory more comprehensible to the students, many illustrations have been used, and this in turn explains how certain parameters of interest change as the algorithm proceeds.

2.2 Knowledge Discovery in a Community Data Set: Malnutrition among the Elderly [3]: The study to design a prediction model that explains the characteristics of elderly adults at risk of malnutrition. A reliable decision support model was designed to provide accurate information regarding the characteristics of elderly individuals with malnutrition. The findings demonstrated the good feasibility of data mining when used for a large community data set and its value in assisting health professionals and local decision makers to come up with effective strategies for achieving public health goals.

2.3 Rule Based Classification to Detect Malnutrition in Children [4]: The data mining is an area which used in vast field of areas. Rule based classification is one of the sub areas in data mining. From this paper it will describe how rule based classification is used alone with Agent Technology to detect malnutrition in children. This system is implemented as an e-government system. Further it will try to research whether there is connection between numbers of rules which is used with the optimality of the final decision.

2.4 Analysis of Meal Patterns with the Use of Supervised Data Mining Techniques—Artificial Neural Networks And Decision Trees [5]: This article demonstrates how a coding system at the meal level might be analyzed by using data mining techniques. The objective was to evaluate the usability of supervised data mining methods to predict an aspect of dietary quality based on dietary intake with a food-based coding system and a novel meal based coding system. Artificial neural networks (ANNs) and decision trees were used to predict quintiles of the HEI based on combinations of foods consumed at breakfast and main meals.

2.5 Artificial Neural Networks in Medical Diagnosis [6]: The artificial neural networks are finding many uses in the medical diagnosis application. The goal of this paper is to evaluate artificial neural network in disease diagnosis. Two cases are studied. The first one is acute nephritis disease; data is the disease symptoms. The second is the heart disease; data is on cardiac Single Proton Emission Computed Tomography (SPECT) images. Each patient classified into two categories: infected and non-infected. Classification is an important tool in medical diagnosis decision support. Feed-forward back propagation neural network is used as a classifier to distinguish between infected or non-infected person in both cases.

2.6 A Review of Machine Learning Techniques and Statistical Models in Anaemia [7]: The blood diseases have in the recent past become a major cause of mortality and morbidity all over the world. Consequently, machine learning has emerged as one of the best and most fruitful methods of research in the present world, both in terms of proposing of new techniques with effective theoretical algorithms, and also in applying such methods in real life situations. These systems work through optimizing performance using certain algorithm in accordance with its maximization or minimization criteria, but also using experimental data instead of a given program.

2.7 Performance Comparison of Data Mining Techniques for Prediction and Diagnosis of Breast Cancer Disease Survivability [8]: The prediction of breast cancer survivability has been a challenging research problem for many researchers. In this paper we have discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis. Breast Cancer Diagnosis is distinguishing of benign from malignant breast lumps and Breast Cancer Prognosis predicts when Breast Cancer is to recur in patients that have had their cancers excised.

III. PROPOSED ALGORITHM

A. *Data Preprocessing Methods*

The data preprocessing techniques can be easily embedded in ensemble learning algorithms. Hereafter, we recall several data preprocessing techniques that have been used together with ensemble approach. A re-sampling techniques that study the effect of changing class distribution to deal with imbalanced data-sets, where it has been empirically proved that the application of a preprocessing step in order to balance the class distribution is usually a positive

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

solution. Re-sampling techniques can be categorized into three groups. Under sampling methods, which create a subset of the original data-set by eliminating instances (usually majority class instances); oversampling methods, which create a superset of the original data-set by replicating some instances or creating new instances from existing ones; and finally, hybrids methods that combine both sampling methods. Random under sampling: It is a non-heuristic method that aims to balance class distribution through the random elimination of majority class examples. Its major drawback is that it can discard potentially useful data, which could be important for the induction process. Random oversampling: In the same way as random under sampling, it tries to balance class distribution, but in this case, randomly replicating minority class instances. Several authors agree that this method can increase the likelihood of occurring over fitting, since it makes exact copies of existing instances.

B. *Splitting Decision Trees*

Decision tree learning is a common method used in data mining. Most of the commercial packages offer complex Tree classification algorithms, but they are very much expensive. Decision tree algorithms generate tree-structured classification rules, which are written in a form of conjunctions and disjunctions of feature values (or attribute values). These classification rules are constructed through 1) selecting the best splitting feature based on a certain criterion, 2) partitioning input data depending on the best splitting feature values, then 3) recursively repeating this process until certain stopping criteria are met. The selected best splitting feature affects not only the current partition of input data, but also the subsequent best splitting features as it changes the sample distribution of the resulting partition. Thus, the best splitting feature selection is arguably the most significant step in decision tree building, and different names are given for decision trees that use different splitting criteria, for example, C4.5 and ID3 for Shannon entropy-based splitting criteria such as and Information Gain ratio and CART for the Gini impurity measure.

Different splitting criteria use their own impurity measures, which are used to calculate “achievable” impurity reduction after a possible split. Consider a nominal feature X and target class Y.

The Node Impurity and Entropy impurity is defined as,

$$E(N) = -\sum_{j} P(\omega_j) log_2 P(\omega_j) \quad (1)$$

where, P(wj): fraction of patterns at node N in category wj

Gini impurity measure of expected error rate at node N if the category label is selected randomly from the class distribution present at N

$$E(N) = \sum_{i \neq j} P(\omega_i) P(\omega_j) = \frac{1}{2}\left[1 - \sum_{j} P^2(\omega_j)\right] \quad (2)$$

The Minimum probability that a training pattern will be misclassified at N

$$E(N) = 1 - max P(\omega_j) \quad (3)$$

C. *Bagging-Based Ensembles*

The bagging consists in training different classifiers with bootstrapped replicas of the original training data-set. That is, a new data-set is formed to train each classifier by randomly drawing (with replacement) instances from the original data-set (usually, maintaining the original data-set size). The bagging ensembles to deal with class imbalance problems due to its simplicity and good generalization ability. The hybridization of bagging and data preprocessing techniques is usually simpler than their integration in boosting. A bagging algorithm does not require re-computing any kind of weights; therefore, neither is necessary to adapt the weight update formula nor to change computations in the algorithm. In these methods, the key factor is the way to collect each bootstrap replica (Algorithm 4), that is, how the class imbalance problem is dealt to obtain a useful classifier in each iteration without forgetting the importance of the diversity. An easy way to overcome the class imbalance problem in each bag is to take into account the classes of the instances when they are randomly drawn from the original data-set. Hence, instead of performing a random sampling of the whole data-set, an oversampling process can be carried out before training each classifier. This procedure can be developed in at least two ways. Oversampling consists in increasing the number of minority class instances by their replication, all majority class instances can be included in the new bootstrap, but another option is to resample them trying to increase the diversity. Note that in all instances will probably take part in at least one bag, but each bootstrapped replica will contain many more instances than the original data-set.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Algorithm 1: Bagging.
Input: S: Training set; T: Number of iterations;
n: Bootstrap size; I: Weak learner
Output: Bagged classifier: H(x) = $\sum_{t=1}^{T} h_t(x)$sign where $h_t(x)$ [-1, 1] are the induced classifiers
for t = 1 to T do
St ← RandomSampleReplacement(*n, S*)
ht ←I(St)
end for

D. *First Order Decision Tree Representation*

The decision tree induction presented in Algorithm 1. Consider a data set with features $\{X'i\}_{ji}=1$ and an imbalanced binary target class Y . To keep the notation simple, to transform the features $\{X'i\}_{ji}=1$into binary features$\{Xi\}ni=1$, i.e., $\{Xi\}_{ni}=1$Binarize($\{X'i\}_{ji}=1$). The inputs to the Decision Tree (DT) algorithm are the binarized features $\{Xi\}_{ni}=1$, the corresponding class labels Y , and a pre-specified α.

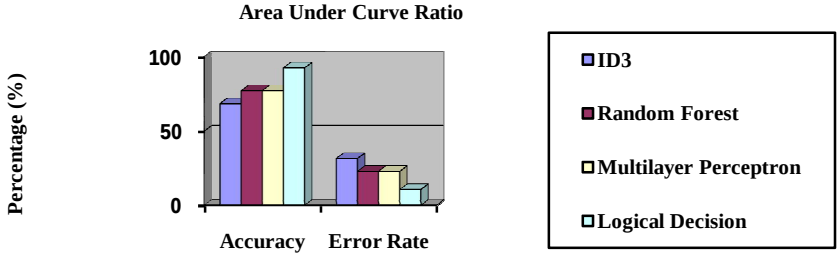
IV. SIMULATION RESULTS

The experimental results background information of the women malnutrition in the sample. The women were between 18 and 50 years of age. They had a mean BMI of 27 kg/m2. According to the WHO BMI classification groups, 3 % were underweight, 28 % had normal weight, 46 % were overweight, and 22 % were obese. Average fluid intake was 1978 ml/day. Table 1 describes the proposed logical decision Accuracy and Error rate of imbalance classification of Area under curve ration of exisiting method ID3, Random Forest and Multilayer perceptron give below,

$$AUROC\ Accuracy = \frac{Logical\ diversified\ AUROC}{original\ AUROC} \quad (4)$$

Table 1: Comparison Methods of imbalance classification of AUROC values

Methods	Accuracy	Error Rate
ID3	68.50	31.50
Random Forest	77.17	22.83
Multilayer Perceptron	77.17	22.83
Logical Decision	**92.60**	**10.58**



ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

V. CONCLUSION

In this work it proposed the logical decision tree (LDT) algorithm showed the highest predictive capabilities with respect to recall and models on the Decision Trees algorithm with low pruning had the highest precision. The proposed algorithm based on two functions namely bagging and logical decision rule method which is used to find the co-variation along with conjunction 0 to 1. Thus the result is obtained with good ensemble classification quality. The ensembles of logical decision tree algorithm break down a complex decision tree into a collection of simpler decision tree. Ensuring that people's Women's learns the importance of eating a balanced diet, means ensuring he or she is free of these diseases and grows up to be a healthy adult. In this work WHO (World Health Organization) database is analyzed for logical decision tree construction using decision algorithm.

REFERENCES

1. Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques: Concepts and Techniques", third edition.
2. K.P.Somen, Shyam Diwakar and V,Ajay, "Insight into Data Mining Theory and practice", Prentice hall of India Private Limited,2008.
3. Myonghwa Park, PhD, Hyeyoung Kim, PhD, Sun Kyung Kim, MSN, " Knowledge Discovery in a Community Data Set: Malnutrition among the Elderly ", Healthcare Informatics and Research,2014.
4. Xu Dezhi and Gamage Upeksha Ganegoda , "Rule Based Classification to Detect Malnut,rition in Children", International Journal on Computer Science and Engineering (IJCSE),2011.
5. Aine P Hearty and Michael J Gibney, "Analysis of meal patterns with the use of supervised data mining techniques—artificial neural networks and decision trees", American Society for Nutrition", 2008.
6. Qeethara Kadhim Al-Shayea ," Artificial Neural Networks in Medical Diagnosis", IJCSI International Journal of Computer Science,2011.
7. Jameela Ali Akrimi, Abdul RahimAhmad, Loay E. George(IJSR), "Review of Machine Learning Techniques in Anemia Recognition", International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064, 2013.
8. K.R. Lakshmi , M.Veera Krishna and S.Prem Kumar," Performance Comparisonof Data Mining Techniques for Predicting of Heart Disease Survivability", International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013 ISSN 2250-3153.