International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

Development Of Data Mining System To Compute The Performance Of Improved Random Tree And J48 Classification Tree Learning Algorithms

Dr.K.Suresh Kumar Reddy[1], Academic Consultant, Department of Computer Science and Engineering, SVU College of Engineering, Sri Venkateswara University, Tirupati, Andhra Pradesh, India, sureshreddy117@yahoo.com

Dr.M.Jayakameswaraiah[2], Assistant Professor, Department of Computer Applications, Madanapalle Institute of Technology & Science, Madanapalle, Chittoor Dist, Andhra Pradesh, India, drjayakameswar@gmail.com

Dr.S.Ramakrishna[3], Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India, drsramakrishna@yahoo.com

Dr.M.Padmavathamma[4],Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India,prof.padma@yahoo.com

Abstract — Data mining has evolved into a vital and active area of research because of the speculative challenges and practical applications related with the problem of discovering interesting and previously unknown knowledge from very big real-world databases. Several aspects of data mining have been investigated in several related fields. Although the database technologists have been seeking efficient way of storing, retrieving and manipulating data, the machine learning communities have focused on developing techniques for learning and acquiring knowledge from the data. To determine the fundamental model that governs the implementation of the physical world and encapsulate the same in theories that can be used for predicting the future. In this research we are developed an Improved Random Tree Learning Algorithm using Shannon entropy and compared the performance of the proposed algorithm with J48 Classification Algorithm on GPS Trajectory dataset to get improved performance and results.

Index Terms—Data Mining, Classification, Improved Random Tree, J48 Classifier.

I. INTRODUCTION

Data mining involves an integration of techniques from multiple disciplines such as data warehouse, machine learning, neural networks, information recovery, data visualization, pattern recognition, image and signal processing and spatial or temporal data study. That is, importance is placed on the capable and scalable data mining techniques. For an algorithm to designate scalable, its running time should be developed almost linear in proportion to the volume of the data, given the existing system resources such as the main memory and disk space. Data mining can be performed with motivating knowledge regularities, the sophisticated information can be extracted from databases and viewed or browsed from dissimilar angles[6].

II. LITERATURE REVIEW

A. Evolution of Data Mining:

Data mining can be viewed as a result of the expected development of information technology. The database system trade has witnessed an evolutionary course in the development of the following functionalities: data collection and database creation, data management (including data storage, retrieval and database transaction processing) and advanced data analysis (involving data warehousing and data mining). For illustration, the early growth of data collection and database design mechanisms served as a prerequisite for later development of effective mechanisms for information storage, recovery, query and transaction processing. With frequent database systems, offering query and transaction processing as frequent practice, complex data analysis has obviously become the next objective[4].

Since the 1960s, database and information technology have been evolving systematically from primitive file processing systems to complicated and dominant database systems. The research and improvement in database systems have progressed from early hierarchical and network database systems to the development of relational database systems, indexing and accessing methods.

Database technology since the mid-1980s has been characterized by the fashionable implementation of relational technology and an expansion of research and expansion activities on new and dominant database systems. These support

the development of sophisticated data models such as object-relational, extended-relational, deductive and object-oriented models. Application oriented database systems as well as spatial, temporal, multimedia, information bases, stream, and sensor, logical and engineering databases, and office information bases have flourished. Issues associated with the allocation, diversification and distribution of data have been considered extensively. Diverse database systems and Internet-based worldwide information systems such as the World Wide Web (WWW) have also emerged and take part in a very important task in the information industry[3].

B. Steps Involved in Knowledge Discovery:

Many people treat data mining as a synonym for one more widely used term, Knowledge Discovery from Data (KDD). On the other hand, data mining can be viewed simply as an essential step in the process of knowledge discovery[7]. Figure 1 shows the Knowledge Discovery consists of an iterative sequence of the following steps:

1. Data cleaning: Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistency in information. Data cleaning involves transformation to accurate the incorrect data. Data cleaning is performed as data preprocessing step before preparing the data for a data warehouse.
2. Data integration: Data Integration is a data preprocessing technique that merges the data from multiple heterogeneous data sources into a coherent data store. Data integration may absorb unpredictable data therefore needs data cleaning.
3. Data selection: Data Selection is the process where data relevant to the analysis task is retrieved from the database. Occasionally data transformation and consolidation are performed prior to data selection procedure.
4. Data transformation: In this step data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. Data mining: In this step intelligent methods are applied in order to extract data patterns.
6. Pattern evaluation: In this step, the data patterns are estimated.
7. Knowledge presentation: In this step, the knowledge is represented.

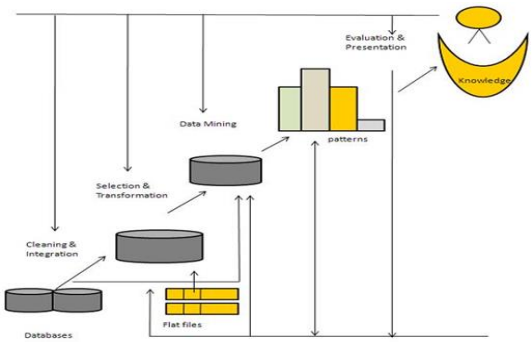

Figure 1: Process of Knowledge Discovery

Steps 1 to 4 are different forms of data preprocessing, wherever the data is ready for mining. The data mining step may interrelate with the user or the information base. The interesting patterns are offered to the user and may be stored as original data in the information base. According to this analysis, data mining is reminded simply as one step in the whole process, even though an important one because it covers known patterns for estimation. We agree that data mining is a step in the data discovery procedure. However, in industry, in media and in the database exploring environment, the term data mining is becoming more popular than the longer phrase of data discovery from data. So, here we choose to use the phrase data mining. We implement a broad view of data mining functionality. Data mining is the development of discovering interesting knowledge from huge amount of information stored in databases, data warehouses or additional information repositories. Based on this analysis, the architecture of a distinctive data mining structure may have subsequent major components[8].

C. Data Mining Functionalities:

Classification is the process of finding a model that elaborates the data classes or concepts. The principle is the ability to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of a set of training data[9,11]. The resulting model can be represented in the following forms

- Classification Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

1. *Classification:* It predicts the class of objects whose class label is unidentified. Its objective is to discover a derived model that describes and distinguishes data module or concepts. The resulting model is based on the analysis of a set of training data i.e. the data object whose class label is identified.
2. *Prediction:* It is used to predict missing or unavailable numerical data values rather than

class labels. Regression analysis is commonly used for the prediction. Prediction can also be used for the identification of distribution trends based on available data.

3. *Outlier analysis:* Outliers are data elements that cannot be grouped in a given class or cluster. The outliers can be considered as noise and discarded in some applications. The Outliers may be defined as the data objects that do not comply with general behavior or model of the data available.
4. *Evolution Analysis:* Evolution Analysis refer to description and model regularities or trend for objects whose behavior changes over time.

III. IMPLEMENTATION OF ALGORITHMS

A. J48 Algorithm:

J48 is a tree based learning algorithm and uses Divide-and-Conquer approach to split a root node into a subset of two partitions till leaf node occur in tree. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that every attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain that results from choosing an attribute for splitting the data[13]. To create the decision, the attribute with the maximum normalized information gain is used. Given a set T of total instances the following steps are used to construct the tree structure. Initially all the training records are at the root. The input to this algorithm consists of the training records E and the attribute set F. The algorithm works by recursively selecting the unsurpassed attribute to split the data and expanding the leaf nodes until stopping principle is met.

Algorithm:

```
TreeGrowth(E,F)
if stopping_cond(E,F) = true then
leaf = createNode().
leaf.label = Classify(E).
return leaf
else
root = createNode().
root.test_cond = find_best_split(E,F).
let V = {v|v is a possible outcome of
root.test_cond}.
for each v ε V do
= {e | root.test_cond(e) = v and e ε E}.
child = TreeGrowth(, F)
add child as descendent of root and label the edge
(root-child ) v.
end for
end if
return root
```

Step 1: A root node that has no incoming edges and zero or more outgoing edges. each of which has exactly one incoming edge and has no outgoing edges.

Step 2: Internal nodes, each of which has exactly one incoming edge and two or more outgoing edges. Internal nodes denote test on attributes.

*Step 3:*If all the instances in S belong to the same class or S is having fewer instances, than the tree is leaf labeled with the most frequent class in S.

Step 4: If step 1 does not occur then select a test based on a single attribute with at least two or greater possible outcomes.

Step 5: Then consider this test as a root node of the tree with one branch of each outcome of the test.

Step 6: Then partition S into corresponding S1, S2, S3..... according to the result for each respective cases.

Step 7: The same may be applied in recursive way to each sub node.

Step 8: Information gain and default gain ratio are ranked using two heuristic criteria.

B. Improved Random Tree Learning Algorithm:

A random tree is a collection of tree predictors that is called forest. It can deal with both classification and regression problems. The classification works as follows: the random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of votes . In case of a regression, the classifier response is the average of the responses over all the trees in the forest. All the trees are trained with the same parameters but on different training sets[12,15]. Improved Random Tree Leaning with Shannon entropy grows many classification trees. Each tree is grown as follows:

Step 1: If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data.

Step 2: This sample will be the training set for growing the tree.

Step 3: If there are M input variables, a number mM is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node.

Step 4: The value of m is held constant during the forest growing.

Step 5: Each tree is grown to the largest extent possible. There is no pruning.

To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

a) Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

b) Entropy using the frequency table of two attributes:

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

C. Benefits of Random Tree Learning Algorithm: Accuracy

- Runs efficiently on large data bases
- Handles thousands of input variables without variable deletion
- Gives estimates of what variables are important in the classification
- Generates an internal unbiased estimate of the generalization error as the forest building progresses
- Provides effective methods for estimating missing data
- Maintains accuracy when a large proportion of the data are missing
- Provides methods for balancing error in class population unbalanced data sets
- Offers an experimental method for detecting variable interactions

IV. WEKA

WEKA is a data mining software developed by the University of waikato in New Zealand that apparatus data mining algorithms using the JAVA language. WEKA is a milestone in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption. The algorithms are directly to a database. WEKA implements algorithms for data pre-processing, classification, regression, clustering and association rules; It also includes visualization tools.

In this research experiment we use WEKA 3.8 and Window 7 to evaluate the J48 Classification Algorithm and our proposed Improved Random Tree Algorithm for generating effective classification approach using respective parameters using GPS Trajectory Dataset from the UCI Machine Learning Repository. Data Set is taken for this algorithm; the input data set is an integral part of data mining application. The data used in our experiment is either real world data obtained from UCI machine learning repository or widely accepted data set available in WEKA Toolkit[1,2,13]. GPS Trajectory data set consists of 163 instances and 10 attributes while some of them contain missing values.

A. Attribute Information of GPS Trajectory Dataset:

go_track_tracks.csv: A list of trajectories
id_android : it represents the device used to capture the instance;
speed : it represents the average speed (Km/H)
distance : it represent the total distance (Km)
rating : it is an evaluation parameter. Evaluation the traffic is a way to verify the volunteers perception about the traffic during the travel, in other words, if volunteers move to some place and face traffic jam, maybe they will evaluate 'bad'. (3- good, 2- normal, 1-bad).
rating_bus : it is other evaluation parameter. (1 - The amount of people inside the bus is little, 2 - The bus is not crowded, 3- The bus is crowded.
rating_weather : it is another evaluation parameter. (2-sunny,1-raining).
car_or_bus-(1-car,2-bus)
linha : information about the bus that does the pathway

V. RESULTS

A. Performance Comparison between J48 and Improved Random Tree Classifiers:

Algorithm	Test Option	Correctly Classified Instances	Incorrectly Classified Instances
J48	Full Training set	85 %	15 %
Improved Random Tree	Full Training set	100 %	0 %

TABLE 1: Performance Comparison of J48 and Improved Random Tree

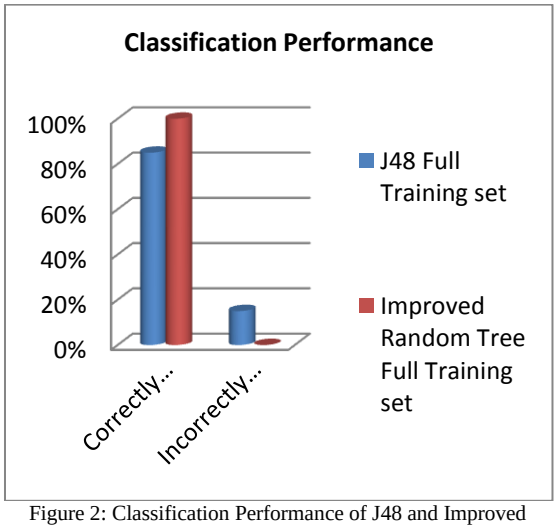

Figure 2: Classification Performance of J48 and Improved Random Tree

Figure 3: Run Information of J48

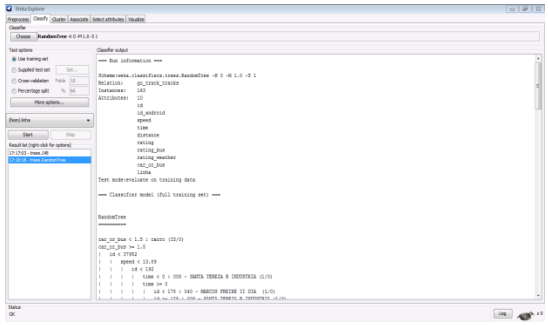

Figure 4: Run Information of Random Tree

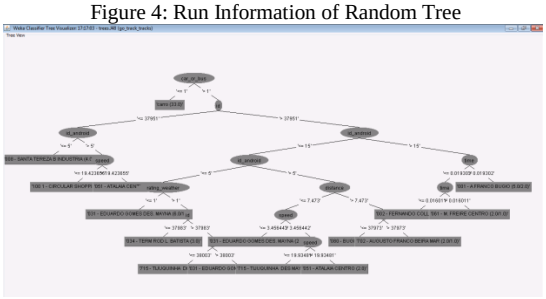

Figure 5: Constructed Tree with J48 Classifier

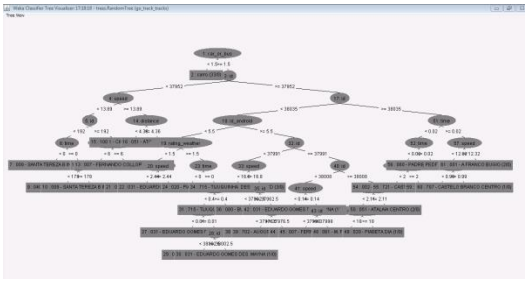

Figure 6: Constructed Tree with Improved Random Tree Classifier

VI. CONCLUSION

Data mining is a wide area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence and database systems, for the analysis of large volumes of data. There have been a large number of data mining algorithms embedded in these fields to execute different data analysis tasks. In this research we compared the classification results of the J48 and the our Improved Random Tree Classifiers using shannon entropy. The results shows that the Improved Random Tree classification algorithm takes good classification performance to classify GPS Trajectory data set and also it gives better accuracy. The specific approaches for classification are characterized, we developed the WEKA method is based on choosing the file and selecting attributes to convert .csv file to flat file and discussed features of WEKA performance. Our work extends to utilize the implementation of different dataset. In future work we can use these decision tree algorithm in medical, banking, stock market and various data mining areas.

REFERENCES

[1]. J. Han., M. Kamber. "Data Mining: Concepts and Techniques (2nd edition)." Morgan Kaufmann Publishers (2006).

[2]. Baowei Song., Chunxue Wei. "Algorithm of Constructing Decision Tree Based on Rough Set Theory." *International Conference on Computer and Communication Technologies Agriculture Engineering*, IEEE (2010).

[3]. Baoshi Ding, Yongqing Zheng and Shaoyu Zang, "A New Decision Tree Algorithm Based on Rough Set Theory." *Asia-Pacific Conference on Information Processing*, IEEE, (2009):326-329.

[4]. B. Othman, Md. Fauzi, and T. M. S. Yau. "Comparison of different classification techniques using WEKA for breast cancer." 3*rd Kuala Lumpur International Conference on Biomedical Engineering 2006*. Springer Berlin Heidelberg (2007).

[5]. Ganti, V., Gehrke, J., and Ramakrishnan, R. "Mining Very Large Databases." *IEEE Computer*, Special issue on Data Mining (1999).

[6]. H. Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD Explorations Newsletter 11.1 (2009):10-18.

[7]. Krakovsky, R., R. Forgac. "Neural network approach to multidimensional data classification via clustering." *Intelligent Systems and Informatics (SISY), 2011 IEEE 9th International Symposium* (2011): 169–174.

[8]. M.Jayakameswaraiah, Prof.S.Ramakrishna, "A Study on Prediction Performance of Some Data Mining Algorithms", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 10, October 2014, ISSN: 2321-7782.

[9]. Mehmed Kantardzic., Jozef Zurada. "Next Generation of Data-Mining Applications." New York : Wiley-IEEE Press 3 (2005).

[10]. Q. Wang., Y. Wu., J. Xiao., and G. Pan. "The Applied Research Based on Decision Tree of Data Mining In Third-Party Logistics*." Automation and Logistics," presented at 2007 IEEE International Conference* (2007).

[11]. Q. Wang., Y. Wu., J. Xiao., and G. Pan. "The Applied Research Based on Decision Tree of Data Mining In Third-Party Logistics*." Automation and Logistics," presented at 2007 IEEE International Conference* (2007).

[12]. Raj Kumar., Dr. Rajesh Verma. "Classification Algorithms for Data Mining: A Survey." *IJIET* 1.2(2012).

[13]. WEKA Software. "The University of Waikato". [http://www.cs.waikato.ac.nz/ml/weka].

[14]. Wei Peng., Juhua Chen., and Haiping Zhou. "An Implementation of ID3 - Decision Tree Learning Algorithm." (2012).

[15]. Weiguo Yi., Jing Duan, Mingyu Lu. "Optimization of Decision Tree Based on Variable Precision Rough Set." *International Conference on Artificial Intelligence and Computational Intelligence IEEE* (2010).