ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

Implementation of Mining Algorithms for Improving Efficiencies in Product Session Data

Nikhil Gaikwad[1], Gaurav Chavan[2], Twinkle Samal[3], Avinash Sonule[4]

B.E Student, Department of Computer Engineering, A.C.Patil College of Engineering, Navi Mumbai, India[1, 2, 3, 4]

Abstract— For any business output means product should be perfectly formulated which is decided in meeting and to make meeting successful there should be some strategies analysis to review different opinion of different member in meeting, and based upon that product is successful in market or not. First it was given by Anvil tool which was bit difficult to analyze because it was working on gesture recognition and then it worked on tree-based structure mechanism which created and define the sessions so that each session can accommodate the individual's opinion, and after that session, if rectification is performed, but commands such as propose, acknowledgment, and negative response do not have a fixed structure or a notion that can differentiate them from one another. So in Proposed system we are using various Data Mining algorithms to label the nodes which differentiate from one another. The experimental results show that Threshold value which will decide future step of the product.

Index Terms—Threshold, bayesnet, Human Interaction, ARM, Classification.

I. INTRODUCTION

Meetings and human interactions provides a platform for communicating between members participating in a meeting. During a meeting, several kinds of human interactions may occur. In a particular meeting session following interactions occur: (i) proposing an idea, (ii) positively or negatively reacting to a proposal, (iii) acceptance of a proposal. We use mining methods to gather significant information regarding the success rate of the decision made in a meeting, using pattern detection.

Data mining is useful in discovering valuable information or knowledge from large datasets. Also, frequent pattern mining [1,2,3] is useful in finding frequently occurring patterns from meetings .The discovered interaction patterns helps to:

(i) Check the effectiveness of decisions made in meetings,

(ii) Conclude whether a meeting discussion is fruitful or not and (iii) index meetings for further ease of access in database.

Also existing meeting capture systems could use this technique as a smarter indexing tool to search and access particular semantics of the meetings [4, 5]. Second, the extracted patterns are useful for understanding human interaction in meetings. Now, the problem comes in labeling the patterns as well as analyzing the patterns which are discovered. Inspired by tree-based mining, we are going to use a Fuzzy Apriori-T approach[10] which is intended to address the crisp boundary problem encountered in traditional ARM(Association Rule Mining) along with FP-Tree algorithm separately and analyze the results generated by these algorithms.

The rest of the paper is organized as follows:

In Section 2: we will discuss previous study related to our work. Section 3 briefly introduces existing systems and proposed system. Section 4 contains analysis. Section 5 contains relative acknowledgments. Finally we conclude the paper in Section 6.

II. RELATED WORK

Human interaction in meetings [12] has proved to be useful for research in the fields of image/speech processing, and human-computer interaction [8].Several works have been done in discovering Human behaviour patterns by using random techniques. Bakeman and Gottman [10] applied sequential analysis to observe and analyze human interactions. To acquire the semantic knowledge, researchers extracted the meeting contents and represented them in a machine readable format. For instance, Waibel et al. [9] presented a meeting browser that describes the dynamics of human interactions. McCowan et al. [11] recognized group actions in meetings by modelling the joint behaviour of participants and expressed group actions as a two-layer process by a hidden Markov model framework. Otsuka et al. [7] used gaze, head gestures, and utterances to determine who

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

responds to whom in multiparty face-to-face conversations.

Yu et al. proposed a multimodal approach for interaction recognition; they also used a tree-based mining method to discover frequent patterns from human interactions in meetings. Such a method focuses mostly on capturing direct parent-child relations.

III. EXISTING AND PROPOSED SYSTEM

In previous works analysis of human Interactions were done using Apriori algorithm and FITM(Frequent Interaction Tree mining) algorithm[2] .The output was the frequent sets of items.Also Chopper algorithm was used for labelling precisely each node in a Tree. In proposed system we are going through FuzzyApriori-T algorithm. These techniques can be summarized as follows:

3.1 Apriori Algorithm

Apriori [2] is a classic algorithm for frequent item set mining and association rule learning over transactional databases. It identifies the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently in the database. These frequent item sets determined using Apriori algorithm determines the association rules which highlight general trends in the database: this has applications in domains such as market basket analysis. Each transaction has a set of items (an itemset). Given a threshold C, the Apriori algorithm identifies the item sets which are subsets of at least C transactions in the database. Apriori tends to use a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. It has some drawbacks viz. Candidate generation generates large numbers of subsets.

3.2 Isomorphic Trees

The main problem comes when we cannot decide if two unordered trees are the same. To define a binary relation isomorphic over non-empty trees, the following rules are considered:

1. A tree with a single node (the root) is isomorphic only to a tree with a single node.
2. Two trees with roots A and B , none of which is a single-node tree, are isomorphic if there is a 1-1 correspondence between the sub-trees of A and of B such that the corresponding sub-trees are isomorphic.

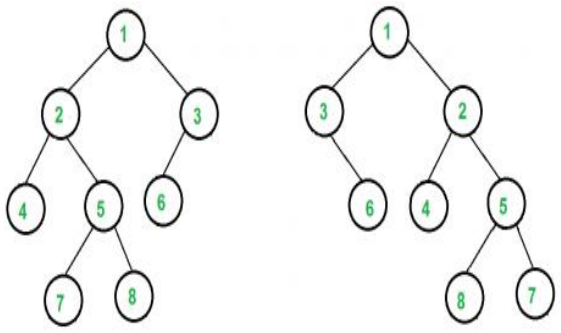


Figure 1: Isomorphic Trees

3.3 Chopper Algorithm

Chopper Algorithm gives the general idea to solve the problem of searching frequent sub-trees. Some properties of an ordered labeled tree will be brought out as follows. We choose the pre-order sequence as the basis of describing an ordered labeled tree, which cannot represent a tree distinctly. Therefore, we have to remember the level number of each element of the sequence in the tree. Thus, we can describe a tree uniquely with the combination of pre-order sequence and level sequence.

For instance, B1A2C3D3E2F3G3G2D3 is used to represent the following tree:

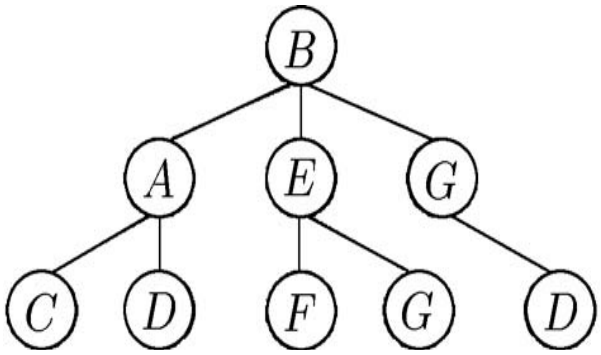


Figure 2: Chopper Algorithm

3.4 FITM

Human interaction flow [12] in a conversation is represented as a tree. Tree based mining algorithms are used to examine the structures of the trees and to discover the interaction flow patterns. These algorithms formulate the frequent tree pattern mining algorithm for every node in the tree .For every tree in TD (tree database), the algorithm first places interactions of

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

siblings to produce the full set of isomorphic trees (ITD). The key principle of generating isomorphic trees is to simplify string matching. Following algorithm generate the isomorphic trees then calculates support values of all trees at Steps 2-3. In Step 4, it selects the trees whose supports are larger than σ and detect isomorphic trees inside them. If m trees are Isomorphic, it selects individual of them and rejects the others. It finally outputs all frequent tree patterns with respect to σ.

3.5 FuzzyApriori-T algorithm[10]

Role of fuzzy sets helps transform quantitative values into linguistic terms, thus reducing possible item sets in the mining process. They are used in the AprioriTid data-mining algorithm to discover useful association rules from quantitative values.

Notation used in this paper is first stated as follows.
n: the total number of transaction data;
m: the total number of attributes;
$D^{(i)}$: the i-th transaction datum, $1 \le i \le n$;
A_j : the j-th attribute, $1 \le j \le m$;
$|A_j|$: the number of fuzzy regions for Aj ;
R jk : the k-th fuzzy region of Aj , $1 \le k \le Aj$;
$V^{(i)}_j$:the quantitative value of Aj for $D^{(i)}$
$f^{(i)}_j$: the fuzzy set converted from $V^{(i)}_j$.
$f^{(i)}_{jk}$: the membership value of $V^{(i)}_j$ in Region R jk;
$count_{jk}$: the summation of $f^{(i)}_{jk}$
$count_j^{max}$: the maximum count value among $count_{jk}$ values, k=1 to |Aj|;
R_j^{max}: the fuzzy region of Aj with count
α: the predefined minimum support level;
ɣ: the predefined minimum confidence value;
C_r: the set of candidate itemsets with r attributes (items);
C_r : the temporary set for recording the fuzzy values of r-items in each data;
L_r : the set of large itemsets with r attributes (items).

The proposed fuzzy mining algorithm first transforms each quantitative value into a fuzzy set with linguistic terms using membership functions. The algorithm then calculates the scalar cardinality of each linguistic term on all the transaction data using the temporary set Cr . Each attribute uses only the linguistic term with the maximum cardinality in later mining processes, thus keeping the number of items the same as that of the original attributes. The mining process based on fuzzy counts is then performed to find fuzzy association rules.
The detail of the proposed mining algorithm is described as follows:

The Algorithm:

INPUT: A body of n transaction data, each with m attribute values, a set of membership functions, a predefined minimum support value α, and a predefined confidence ɣvalue

OUTPUT: A set of fuzzy associate rules.
STEP 1: Transform the quantitative value $v^{(i)}_j$ of each transaction datum D(i) , i=1 to n, for each attribute A_j, j=1 to m, into a fuzzy set $f^{(i)}_j$ represented as :
$\left(\frac{f^{(i)}_{j1}}{R_{j1}} + \frac{f^{(i)}_{j2}}{R_{j2}} + \cdots + \frac{f^{(i)}_{jl}}{R_{jl1}}\right)$ using the given membership functions, where
R_{jk} is the k-th fuzzy region of attribute A_j, $f^{(i)}_{jk}$ is $v^{(i)}_j$'s fuzzy membership value in region R_{jk}, and l(=$|A_j|$) is the number of fuzzy regions for A_j
STEP 2: Build a temporary set $\overline{C_1}$ including all the pairs (R_{jk} , $f^{(i)}_{jk}$) of each data, where $1 \le i \le n$, $1 \le j \le m$, $1 \le k \le |A_j|$, and $f^{(i)}_{jk} \neq 0$.
STEP 3:For each R_{jk} stored in $\overline{C_1}$, calculate its scalar cardinality for all the transactions from $\overline{C_1}$:

$count_{jk} = \sum_{i=1}^{n} f^{(i)}_{jk}$.

STEP 4: Find $count_j^{max} = \underset{k=1}{\overset{|Aj|}{MAX}}(count_{jk})$, for j=1 to m, where $|A_j|$ is the member of fuzzy regions for A_j .Let R_j^{max} be the region with
count $_j^{max}$ for attribute A_j. R_j^{max} will be used to represent this attribute in later mining processing.

STEP 5 : Check whether the $count_j^{max}$ of each R_j^{max}, j=1 to m, is larger than or equal to predefined minimum support value, put it in the set of large 1-itemsets(L_1).That is,
$L_1 = \{R_j^{max} | \text{ count } _j^{max} \ge \alpha, 1 \le j \le m\}$.

STEP 6: Set r=1, where r is used to represent the number of items kept int the current large itemsets.
STEP 7: Generate the candidate set C_{r+1} from L_r. Restated, the algorithm joins L_r and L_r under the condition that r-1 items in the two itemsets are the same and the other one is different. Store in C_{r+1} the itemsets which have all their sub-r-itemsets in L_r.
STEP 8: Build an empty temporary set C_{r+1} .
STEP 9: Do the following substeps for each newly formed (r+1)-itemsets with items
(s_1 , s_2 , ..., s_{r+r1}) in C_{r+1}:
For each transaction datum $D^{(i)}$, calculate its fuzzy value on s as
$f_s^{(i)} = f_{s1}^{(i)} \Lambda f_{s2}^{(i)} \Lambda \ldots \Lambda f_{s_{r+1}}^{(i)}$ using $\overline{C_r}$, where $f_{sj}^{(i)}$ is the fuzzy

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

membership value $D^{(i)}$ in region S_j.If the minimum operator is used for the intersection, then $f_s^{(i)} = \min_{j=1}^{r+1} f_{sj}^{(i)}$.

(b)Store the pair (s, $f_s^{(i)}$) of Case I in $\overline{C_{r+1}}$, where $1 \le i \le n$, $f_s^{(i)} \neq 0$.

(c) Set $count_s = \sum_{i=1}^{n} f_s(i)$ using $\overline{C_{r+1}}$

(d)If $count_s$ is larger than or equal to the predefined minimum support value α, put s in L_{r+1}.

STEP 10: IF L_{r+1} is null, then do the next step; otherwise, set r=r+1 and repeat STEPs 7 to 10.

STEP 11: Construct the association rules for all large q-item set s with items($s_1, s_2, \ldots, s_q$), $q \ge 2$, using following substeps :

(a) Form all possible association rules as follows:

$S_1 \Lambda \ldots \Lambda S_{k-1} \Lambda s_{k+1} \Lambda \ldots \Lambda s_q \rightarrow s_k$, k=1 to q.

(b) Calculate the confidence values of all association rules using :

$$\frac{\sum_{i=1}^{n} f_s^{(i)}}{\sum_{i=1}^{n} \left(f_{s1}^{(i)} \Lambda \ldots \Lambda f_{Sk-1}^{(i)}, f_{Sk+1}^{(i)} \Lambda \ldots \Lambda f_{Sq}^{(i)} \right)}$$

STEP 12: Output the rules with confidence values larger than or equal to the predefined confidence threshold λ.

After STEP 12, the rules constructed are output and can act as the meta-knowledge for the given transactions.

```
LOAD DATA FILE:
inputFileName = D:\Fuzzy\test1.num
Number of records = 23
Reading input file:
D:\Fuzzy\test1.num
Number of columns = 50
---------------------------------
READ OUTPUT SCHEMA FILE:
inputFileName = D:\Fuzzy\test1.schema
Number of lines (attributes) in output schema file = 50
---------------------------------
INPUT SUPPORT THRESHOLD:
Support threshold (%) = 0.5
---------------------------------
INPUT CONFIDENCE THRESHOLDS:
Confidence threshold (%) = 0.6
---------------------------------
SORT INPUT DATA:
---------------------------------
SORT AND PRUNE INPUT DATA:
Operation failed!
---------------------------------
FUZZY APRIORI-T (WITH X CHECK):
Apriori-T with X-Cchecking
Minimum support threshold = 0.5 (11.5 records)
Levels in T-tree = 2
Generation time = 0.0 seconds (0.0 mins)
Number of frequent sets = 3.
---------------------------------
GENERATE ARs (SUPPORT AND CONFIDENCE FRAMEWORK):
Fuzzy Apriori-T
Confidence Thold = 0.6
Number of rules = 2
---------------------------------
FUZZY APRIORI-T (WITH X CHECK):
```

Figure 3: Output of Fuzzy T Apriori Algorithm

IV. ANALYSIS OF VARIOUS ALGORITHMS USED

Analysing Different Algorithm

Naive Bayes

A Naive Bayes[13] classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions A more descriptive term for the underlying probability model would be "independent feature model" An overview of statistical classifiers is given in the article on Pattern recognition In Naive Bayes Algorithm the TP Rate is 0.97 and FP Rate is 0.017 .

Bayesnet

A Bayesian network [13], Bayes network, belief network, Bayes(ian) model or probabilistic directed acyclic graphical model is a probabilistic graphical model In Bayes Net Algorithm the TP Rate is 1and FP Rate is 0.

Naive Bayes Multinomial

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

The multinomial Naive Bayes classifier [14] is suitable for classification with discrete features. Algorithm the TP Rate is 0.68 and FP Rate is 0.55

J48

J48 is an open source Java implementation of the C4.5 algorithm in the weka data mining tool. C4.5 is an algorithm used to generate a decision tree. C4.5 is often referred to as a statistical classifier.

SMO

Sequential minimal optimization (SMO) is an algorithm for solving the optimization problem which arises during the training of support vector machines. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically.

Random Forest

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

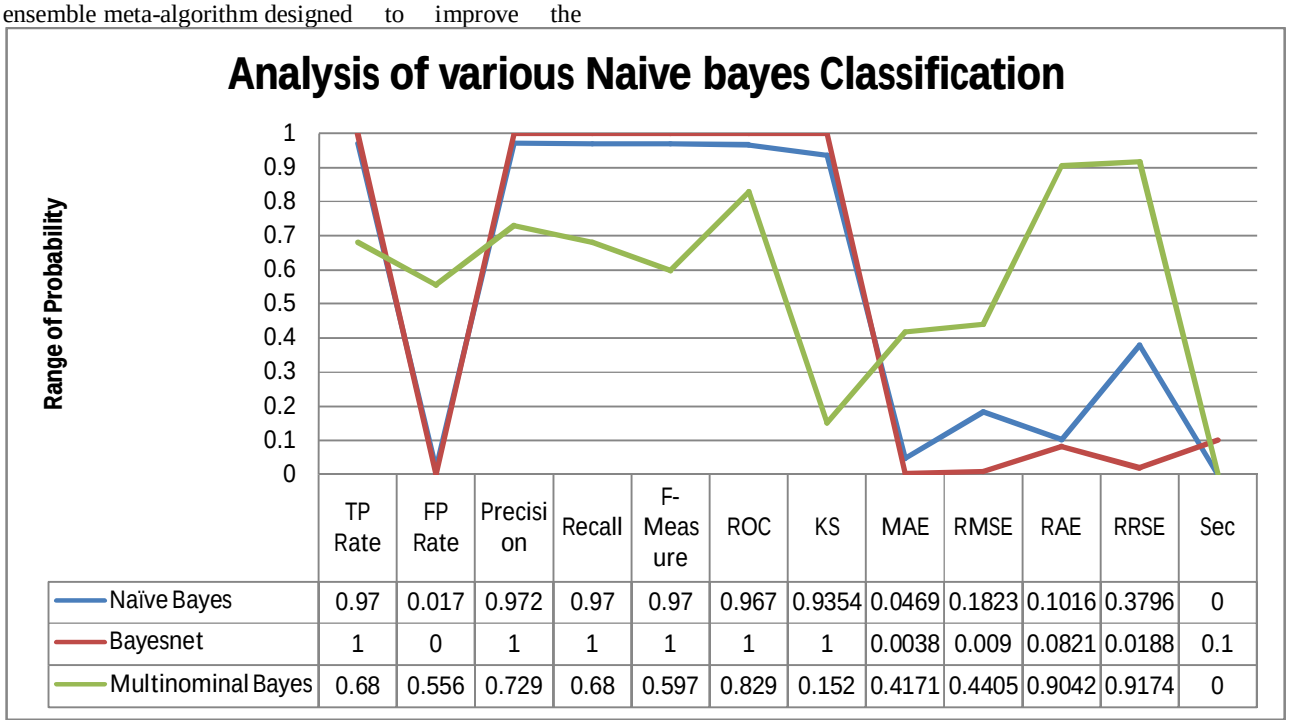
Bagging

Bootstrap aggregating (bagging) is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid over fitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach.

Ada Boost M1

It is ameta-algorithm, and can be used in conjunction with many other learning algorithms to improve their performance

AdaBoost is adaptive in the sense that subsequent classifiers built are tweaked in favor of those instances misclassified by previous classifiers..



	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC	KS	MAE	RMSE	RAE	RRSE	Sec
Naïve Bayes	0.97	0.017	0.972	0.97	0.97	0.967	0.9354	0.0469	0.1823	0.1016	0.3796	0
Bayesnet	1	0	1	1	1	1	1	0.0038	0.009	0.0821	0.0188	0.1
Multinominal Bayes	0.68	0.556	0.729	0.68	0.597	0.829	0.152	0.4171	0.4405	0.9042	0.9174	0

Figure 4: Analysis of various Naive Bayes Classification

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

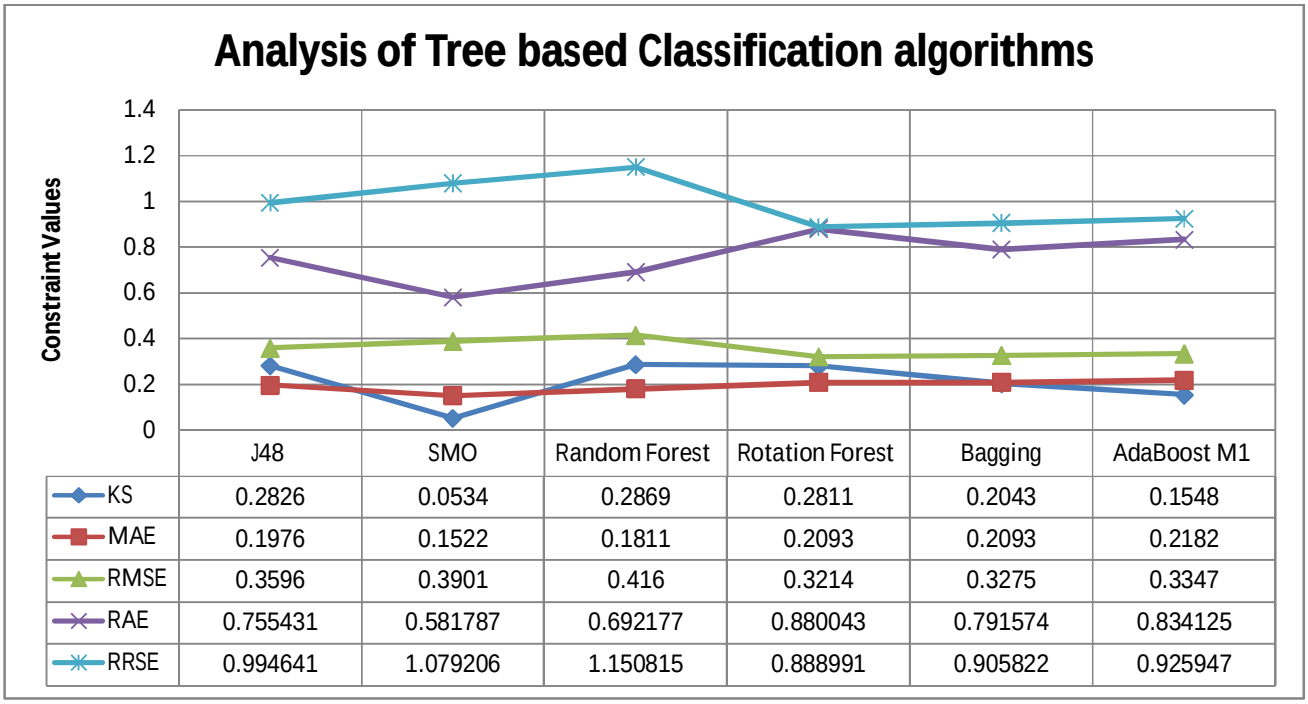
International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India



	J48	SMO	Random Forest	Rotation Forest	Bagging	AdaBoost M1
KS	0.2826	0.0534	0.2869	0.2811	0.2043	0.1548
MAE	0.1976	0.1522	0.1811	0.2093	0.2093	0.2182
RMSE	0.3596	0.3901	0.416	0.3214	0.3275	0.3347
RAE	0.755431	0.581787	0.692177	0.880043	0.791574	0.834125
RRSE	0.994641	1.079206	1.150815	0.888991	0.905822	0.925947

Figure 5: Comparison analysis of Human Interaction dataset by using Tree based classification Algorithms

ACKNOWLEDGMENT

We dedicate this paper to Our HOD Prof. Nitin .P.Chawande, Our Guide Prof.Avinash Sonule and a Special Thank to Mr. Hemant Palivela for giving encouragement to write this paper.

REFERENCES

[1] Ahmed, C.F., Tanbeer, S.K., Jeong, B.-S., Lee, Y.-K.: An efficient candidate pruning technique for high utility pattern mining. In: PAKDD 2009. LNAI 5476, pp. 749–756. Springer (2009).

[2] Leung, C.K.-S., Mateo, M.A.F., Brajczuk, D.A.: A tree-based approach for frequent pattern mining from uncertain data. In: PAKDD 2008. LNAI 5012, pp. 653–661. Springer (2008)

[3] Tanbeer, S.K., Ahmed, C.F., Jeong, B.-S., Lee, Y.-K.: Discovering periodic-frequent patterns in transactional databases. In: PAKDD 2009. LNAI 5476,pp. 242–253. Springer (2009)

[4] W. Geyer, H. Richter, and G.D. Abowd, "Towards a Smarter Meeting Record—Capture and Access of Meetings Revisited," Multimedia Tools and Applications, vol. 27, no. 3, pp. 393-410, 2005.

[5] Z. Yu, M. Ozeki, Y. Fujii, and Y. Nakamura, "Towards Smart Meeting: Enabling Technologies and a Real-World Application," Proc. Int'l Conf. Multimodal Interfaces (ICMI '07), pp. 86-93, 2007.

[6] McCowan, L., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, D.: Automatic analysis of multimodal group actions in meetings. IEEE TPAMI 27(3), 305–317 (2005).

[7] Otsuka, K., Sawada, H., Yamato, J.: Automatic inference of cross-modal nonverbal interactions in multiparty conversations: "who responds to whom, when, and how?" from gaze, head gestures, and utterances. In: ICMI 2007, pp. 255–262. ACM (2007)

[8] Waibel, A., Bett, M., Finke, M. Stiefelhagen, R: Meeting browser: tracking and summarizing meetings. In: DARPA Broadcast News Transcription and Understanding Workshop (1998)

[9] R. Bakeman and J.M. Gottman, Observing Interaction: An Introduction to Sequential Analysis. Cambridge Univ. Press, (1997.)

[10] A fuzzy AprioriTid mining algorithm with reduced computational time TP Hong, CS Kuo, SL Wang - Applied Soft Computing, 2004 - Elsevier

[11] Palivela, Hemant, et al. "Novel Approach for Finding Patterns in Product-Based Enhancement Using Labeling Technique." Intelligent Computing, Networking, and Informatics. Springer India, 2014. 1249-1256.

[12] Yu, Zhiwen, Zhiyong Yu, Xingshe Zhou, Christian Becker, and Yuichi Nakamura. "Tree-based mining for discovering patterns of human interaction in meetings." Knowledge and Data Engineering, IEEE Transactions on 24, no. 4 (2012): 759-768.

[13] Hemant, Palivela, and Thotadara Pushpavathi. "A novel approach to predict diabetes by Cascading Clustering and Classification." Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on. IEEE, 2012.

[14] Kibriya, Ashraf M., et al. "Multinomial naive bayes for text categorization revisited." AI 2004: Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2005. 488-499.