Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Evaluating the Performance of Rule, Density and Tree Techniques Using Classification and Clustering Algorithms

[1]Supreet Kaur, [2]Amanjot Kaur
[1]Research Scholar, [2]Assistant Professor
[1, 2] Department of Computer Science and Engineering, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India

Abstract: Data mining is the course of the action support of searching huge record for pattern used mainly to discover association between variables .In this technique, clustering is universally used for fastening the data according to the probing bond between the data. Later than agreement of information, classification of information is equipped, currently essentials are appreciate under categorized cluster. Judgment based is also speculating based winning the classification. The association linking clustering and classification is depending ahead the pole apart algorithm beginning to end analyzing data which grows into supplementary well-ordered record. The inescapable potential of this paper is to get links of the unusual algorithms line of attack of clustering and classification to poke approximately in the sort of strengthen algorithm to the obligation. Evaluator the Clustering algorithms like Cobweb, Hierarchical, DBSCAN, OPTICS and classification algorithms: RIDOR, PART, BF Tree, AD Tree.

Keywords: Data mining, Clustering algorithms: COBWEB, Hierarchical, DBSCAN, OPTICS, classification algorithms: RIDOR, PART, BF Tree, AD Tree.

I. METHODOLOGY

My line of attack is extremely easy. It is attractive the past development information beginning the repositories and relate it classification and clustering. I am applying different- different clustering algorithms and classification algorithms expect a useful product so as extremely supportive for the creative users and innovative researchers. Technique research effort is as.

1. Dataset – used datasets like sick, lung cancer, liver disorder, hepatitis
2. Classification Rule based Algorithms
   PART and RIDOR
   Classification tree based algorithms
   BF Tree and AD Tree
3. Clustering density based algorithms
   DBSCAN, OPTICS.
   Clustering tree based algorithms
   COBWED, Hierarchical
4. Act of factors
   No. of instances, No. of attributes, Classified Instances, Clustering instances and Time taken.

II. DATASET

For performing arts the relationship analysis requires the earlier period development datasets. In this do research attractive in sequence of two information repositories. ISBSG and PROMISE data repositories make available the earlier period development information. This must in use the different natural history. These repositories are exceptionally kind intended for the researchers. We know how to right away be relevant in sequence within the data mining classification and clustering expects the result. In this do research article apply for datasets like sick, lung cancer, liver disorder, hepatitis. Clustering and classification algorithms are applied on these datasets.

III. INTRODUCTION

Data mining is the course of action of involuntary classification of exceptional special effects based on ground rules patterns obtained in vision of the dataset. Data mining furthermore distinguished as KDD. Knowledge discovery procedure of act consists of an iterative chain of steps such as facts crackdown, data incorporation, data selection, data transformation, data mining, prototype estimation also knowledge appearance.

Data mining involves six well-known modules of tasks:

- Preprocessing: The discovery of extraordinary data proceedings, that might be attention-grabbing or data bugs that oblige auxiliary exploration
- Association: Invention for relative's line by variables.

- Clustering: Is the undertaking of discovering groups and structures in the data that are in a quantity of manner or a further "analogous", exclusive of by means of agreed structures in the facts.
- Classification: Is the undertaking of generalizing well-known configuration to be relevant to innovative data.
- Regression: Attempts to stumble on a task which models the data with the smallest amount inaccuracy.
- Summarization: On provision so as to a supplementary full to capacity in display of the data set, mutually with apparition and description development.

IV. CLUSTERING

Clustering is a data mining ability of assembly situate of information material into many groups or clusters so that bits and pieces surrounded by the cluster have remote over the opinion equal, other than are extremely dissimilar to substance in the supplementary clusters. Clustering algorithms are used to arrange out the information, sort out information, for data density and representation construct, for uncovering of outliers etc. Clustering is a mainly central dependability of explorative information insertion, and a wide-ranging capacity for algebraic information stop working used in many fields, collectively with method culture, prototype acknowledgment, representation investigation, in sequence reclamation, and bioinformatics.

Clustering Algorithms: In this phase of research paper, I have to choose the four clustering algorithms: Cobweb, Hierarchical, DBSCAN, and OPTICS.

4.1 Cobweb: COBWEB is an incremental classification for hierarchical clustering. Cobweb incrementally organizes accepting into a classification chain of command. Each one knob in a classification hierarchy symbolizes a category with a probabilistic impression that summarizes the aspect significance distributions of substance not to be disclosed lower the handle. Classification hierarchy able to used to estimate not there attributes or the grouping of a latest object. Cobweb uses a heuristic estimation determines called sort effectiveness to conduct manufacture of the hierarchy. It percentage increase integrate substance keen on a classification hierarchy in categorize to obtain the uppermost category effectiveness. Cobweb employs four indispensable operations in construction the classification hierarchy. The operations are:

1. Amalgamation two nodes.
2. Splitting a knob.
3. Inserting attach of joint.

Pseudocode: The COBWEB algorithm

Procedure: children := {copy(core)}
new category(database) \\ adds child with record's feature values.
insert(database, core) \\ update root's statistics
else
insert(database, core)
for child in root's children do
calculate cu for insert(database, child),
set best1, best2 children w. best cu.
end for
if newcategory(database) yields best cu
then
newcategory(database)
else if merge(best1, best2) yields best cu
then
merge(best1, best2)
COBWEB(core, database)
else if split(best1) yields best cu
then
split(best1)
COBWEB(core, database)
else
COBWEB(best1, database)
end if
end

Table:1 COBWEB

Dataset Name	No. of instances	No. of attributes	Clustered Instances	Time taken
SICK	3772	30	3335	155.21 sec
LUNG CANCER	32	57	32	0.03 sec
LIVER DISORDER	345	7	302	0.39 sec
HEPATITIS	155	20	146	0.14 sec

4.2 Hierarchical: It is set of thread complementary a progression of position of scales by creating a cluster hierarchy or dendrogram. In hierarchical clustering sequence are not division complex in a challenging cluster in a division swiftness. as prospect, a chain of partition takes place, which able to run construct a particular cluster containing all individual matter to n clusters every containing a particular aim.The sequence is not a particular circumstance of clusters, apart from quite a multilevel chain of command; Allows constructing a finale stage or level of clustering so as to practically each and every suitable for your purpose. Agglomerative method and divisive method are for the most part used in hierarchical clustering.

Hierarchical Agglomerative Clustering: Hierarchical agglomerative clustering begin from bottom, through each datum during individual singleton cluster and combine cluster.

Pseudocode:The Hierarchical Agglomerative Method algorithm

```
Procedure: ws=new set (point);
KDTree kdtree=newKDTree (points);
While (right){
Foreach (factor p in ws) {
If (p.hascluster ()) go on;
Point r=kdtree.findnearest (p);
}
If(q==qull)break;//discontinue but p is end element
Point r=kdtree.final Nearest(q);
If(p==r){//make original cluster e to be contain by a and b
Element e= cluster (p,q);
newwork.add(e);
}
Else{//cant't cluster yet,try again later
newWork.add(p);//add back to worklist
}
}
If(newWork.size()==1)//we have a single cluster
Break;
Ws.addAll(newWork);//add new nodes to worklist
Kdtree.addAll(newWork);
NewWork.clear();
```

Table: 2 HIERARCHICAL

Dataset Name	No. of instances	No. of attributes	Clustered Instances	Time taken
SICK	3772	30	0	688.12 sec
LUNG CANCER	32	57	2	0.02 sec
LIVER DISORDER	345	7	2	0.34 sec
HEPATITIS	155	20	1	0.06 sec

4.3 Density Based Methods: DBSCAN Clusters is course of creature shapes & as a rule reproduce clusters while intense regions of substance in the information hole so as to be separated through regions of low density. OPTICS extends DBSCAN toward construct a *cluster ordering* obtained beginning a large collection of parameter settings.

4.3.1 DBSCAN: Density-based spatial clustering of applications with noise is exacting a position of points in a capacity of space; it groups of simultaneously points so as to intimately pack as a group. It uses the conception of density get to capacity and density fix capacity. Begin through a subjective point p starting with record and recover every point density-reachable beginning p with consider to Eps and MinPts. If p is a center point, the course of action yields a cluster with regard to Eps and MinPts and the points is classified. If p is a edge points, no points are density –reachable as of p and DBSCAN visit the subsequently unclassified point during the record.

Pseudocode: The DBSCAN algorithm

```
Procedure:clustered:=nextID(NOISE);
Foreach  p ∈ SetOfPoint do
If p.classifiedAs==UNCLASSIFIED
then
If Expandcluster (SetOfPoints ,p ,clustered ,Eps , MinPts)
then
clustered++;
endif
endif
endforeach
```


Table: 3 DBSCAN

Dataset Name	No. of instances	No. of attributes	Clustered Instances	Time taken
SICK	3772	30	75	112.8 sec
LUNG CANCER	32	57	32	0 sec
LIVER DISORDER	345	7	2	0.13 sec
HEPATITIS	155	20	2	0.05 sec

4.3.2 OPTICS: OPTICS is stands for ordering points to identify the clustering structure. Algorithm innovated from density-based clusters in spatial information. Its fundamental design is related to DBSCAN, except it addresses individual of DBSCAN's most important weaknesses: the difficulty of detecting significant clusters in information of changeable density. Represented by dendrogram.

Pseudocode: The OPTICS algorithm.
Procedure:OPTICS(SetOfObjectsMinPts, OrderedFile)
OrderedFile.open();
FOR i FROM 1 TO SetOfObjects.size DO
Object := SetOfObjects.get(i);
IF NOT Object.Processed
THEN
ExpandClusterOrder(SetOfObjects, Object,,MinPts, OrderedFile)
OrderedFile.close();
END; // OPTICS

Table: 4 OPTICS

Dataset Name	No. of instances	No. of attributes	Clustered Instances	Time taken
SICK	3772	30	3772	45.06 sec
LUNG CANCER	32	57	32	0.14sec
LIVER DISORDER	345	7	345	0.25sec
HEPATITIS	115	20	115	0.17sec

V. CLASSIFICATION

Classification is way of action of arrangement simultaneously credentials or proceedings that have analogous properties or are unified. Classification is the quandary of identifying to which of a lay down of categories a innovative surveillance belongs, on the root of a working out situate of information containing interpretation. Classification is glowing planned an illustration of supervised tutoring, i.e. awareness where a effective position of in the permitted approach accepted understanding is accessible. The analogous unsubstantiated course of action is known as clustering.In this term paper, our core mean is to measure up to singular clustering and classification algorithms by means of dissimilar parameters and act upon the investigational outcome.

Classification Algorithms: In this phase of research paper, I have to choose the four classification algorithms: AD Tree, BF Tree, PART, RIDOR.

5.1 AD TREE: Alternating decision tree function on top down approach[22]. This approach select furthermost grouping so as to separates the classes of the dataset be valid the project partition rule. Alternating decision tree is build by recursive utility so as to create each and every nodes in the hierarchy.

Algorithm: The AD Tree
Procedure:If(precondition)
If (condition)
return score_one
else
return score_two
end if
else
return 0
end if

Table: 5 AD TREE

Dataset Name	No. of instances	No. of attributes	Classified Instances	Time taken
SICK	3772	30	3699	0.41sec
LUNG CANCER	32	57	32	0.02sec
LIVER DISORDER	345	7	206	0.03sec
HEPATITIS	155	20	118	0.02sec

5.2 BF Tree Classification: BF (Best First) tree algorithms [8] are binary trees surrounded by which the "best" join is delayed each one position. The "best" joint is the joint whose divide leads near upper limit fall of impurity involving each and every nodes accessible used for splitting. The hierarchy rising technique attempts to exploit within-node homogeneity. The scope of which a node does not symbolize a homogenous dividing up cases is a suggestion of impurity.

5.2.1 Pseudocode : BEST FIRST TREE

Procuder: Best <— Root
Bestscore 00; Estimate Assess(Best)
1 if State(Best) = Closed
then return failure endif (1)
if State(Best) = Final
then return Description(Best) endif (2)
Descriptors Generate-Next(Best) (3)
if Descriptors = nil
then State(Best) Closed
else Evaluate(Best) (4)
New Create-S uccessor(Best, Descriptors)
if Distractors(New) = nil then (5)
State(New) Final (6)
if Bestscore > Score(New) then (7)
Bestscore Score(New)
for every node n do
if (State(n) = Open) and (Bestscore
< (Score(n) + Minassess(n))) (8)
then State(n) Cut-off endif (9)
next endif endif endif
for every node a do
if Assess(n) < Estimate
then Estimate (Score(n) + Assess(n))
Best n endif (10)
next
goto Step I

Table: 6 BFTree

Dataset Name	No. of instances	No. of attributes	Classified Instances	Time taken
SICK	3772	30	3767	3.29 sec
LUNG CANCER	32	57	29	0.03 sec
LIVER DISORDER	345	7	259	0.06 sec
HEPATITIS	155	20	149	0.2 sec

5.3 PART Classifier: PART (Partial Decision Tree) [4] is not straight presentation for assemble collection of method. It put together construct use of of partial conclusion hierarchy to generate the being system as well as hierarchy is induced by C4.5 classifier. in a while hierarchy production, method are derived in a straight line beginning the incomplete hierarchy initial by means of the sincere leaf node in mixture with every node next to the path towards the core. in that case, the incomplete ending hierarchy is uncomplicated.

5.3.1 Pseudocode: PARTRule(D, Rt)

Procedure: R = blank set; // original position of rules learned is
blank
Do
Tree node T = PART Tree (D, Rt)
make a rule r from T
R = R + r;
X=X – r // take away every optimistic aim tuples
fulfilling r from X;
While (X)
Return R;
End

Table: 7 PART

Dataset Name	No. of instances	No. of attributes	Classified Instances	Time taken
SICK	3772	30	3758	0.21 sec
LUNG CANCER	32	57	30	0 sec

LIVER DISORDER	345	7	297	0.02 sec
HEPATITIS	155	20	148	0.01 sec

5.4 RIDOR Classifier: Ripple Down Rule (RIDOR) [7], [6], [1], [2], [3**]** is well during straight line categorization procedure. original and principal it constructs the evasion rule and after that produces the exceptions used for the default law by means of lowest mistake rate. For each exclusion, the the majority brilliant exceptions are produced thereby produces the Tree-like development of exceptions. The exceptions stand for a position of rules that forecast module additional than the default. Incremental Reduced Error Pruning IREP [5] is used to create the exceptions.

5.4.1 Pseudocode Ridor (D, Rt)

Procedure: Rule set R = empty set;
if |Rt| < MIN_SUP **then** return
Rule r =empty rule
Set Rt active
Repeat
Find a rule r in dynamic relation or dealings
Joinable among dynamic relation
Learn except branch and if not branch
Set relation of r to active
R= R + r
X= X- r // obtain left each optimistic aim tuples
Fulfilling r from X;
Until (X=NULL)
situation each active relations into inactive
Return R;
End

Table: 8 RIDOR

Dataset Name	No. of instances	No. of attributes	Classified Instances	Time taken
SICK	3772	30	3709	0.36 sec
LUNG CANCER	32	57	28	0 sec
LIVER DISORDER	345	7	259	0.02 sec
HEPATITIS	155	20	131	0 sec

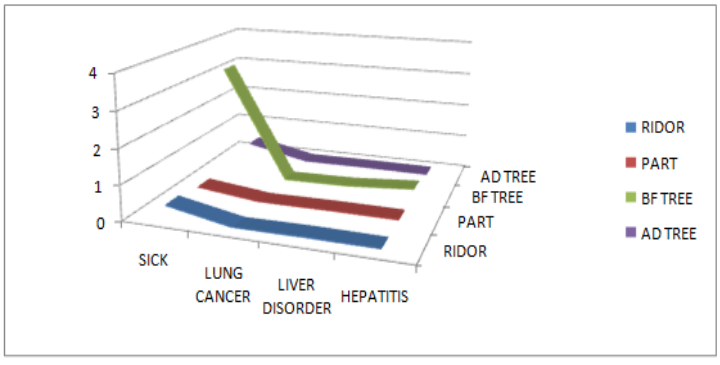
VI. EXPERIMENTAL AND GRAPHS RESULTS

Sick, Lung Cancer, Liver Disorder, Hepatitis data set is with clustering algorithm Cobweb, Hierarchical, DBSCAN, OPTICS and classification algorithms: RIDOR, PART, BF Tree, AD Tree. Algorithm with evaluation. Total time taken to build model .This paper compares various classification algorithms Cobweb, Hierarchical, DBSCAN, OPTICS and classification algorithms: RIDOR, PART, BF Tree, and AD Tree data set. Four dataset applied on algorithms and results related to the time taken built model. table describes the time taken built models the algorithms and conclusion draw as the graphs .table and graph one describe the classification algorithms time taken and table and graph two describe the clustering algorithms time taken. Table and graph three describe both classification and clustering algorithms time taken.

COMPARISON OF CLASSIFICATION & CLUSTERING ACCORDING TO TIME TAKEN

Table 9: Performance Measures according to time take clustering algorithms

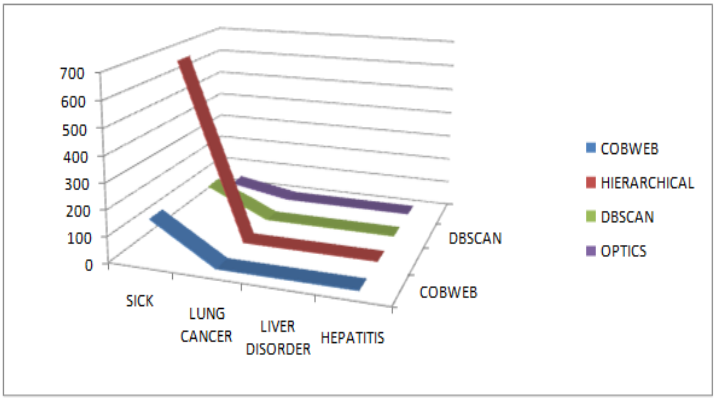
DATASET NAME	RIDOR	PART	BF TREE	AD TREE
SICK	0.36	0.21	3.29	0.41
LUNG CANCER	0	0	0.03	0.02
LIVER DISORDER	0.02	0.02	0.06	0.03
HEPATITIS	0	0.01	0.2	0.02



Graph: 1 Graphical representaion of time taken to form classification

Table 10: Performance Measures according to time take clustering algorithms

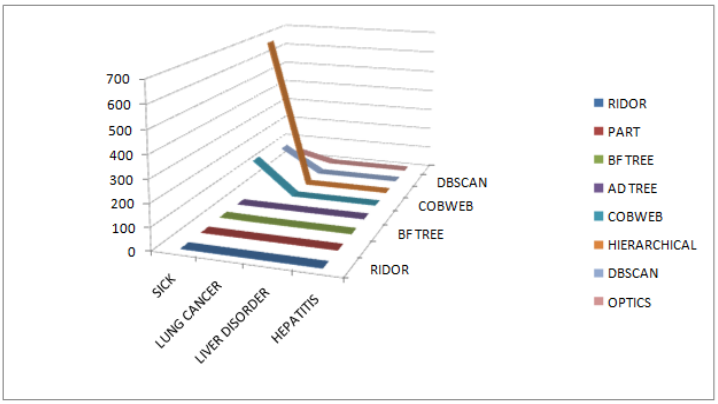
DATASET NAME	COBWEB	HIERARCHICAL	DBSCAN	OPTICS
SICK	155.21	688.21	112.8	45.06
LUNG CANCER	0.03	0.2	0	0.14
LIVER DISORDER	0.39	0.34	0.13	0.25
HEPATITIS	0.14	0.4	0.05	0.17



Graph 2: Graphical representaion of time taken to form clustering

Table 11: EVALUATION OF CLASSIFICATION & CLUSTERING ACCORDING TO TIME TAKEN

DATASET NAME	RIDOR	PART	BF TREE	AD TREE	COBWEB	HIERARCHICAL	DBSCAN	OPTICS
SICK	0.36	0.21	3.29	0.41	155.21	688.21	112.8	45.06
LUNG CANCER	0	0	0.03	0.02	0.03	0.2	0	0.14
LIVER DISORDER	0.02	0.02	0.06	0.03	0.39	0.34	0.13	0.25
HEPATITIS	0	0.01	0.2	0.02	0.14	0.4	0.05	0.17



Graph 3: Graphical representation of time taken formed using different dataset Evaluation of classification & clustering according to time taken

VII. CONCLUSION

This research paper covers data mining technique, Begin the working through data mining models, require the familiarity with the algorithms. The core of this paper to give complete introduction of clustering algorithms and classification algorithms. Experimental results and graphs gives us the performance and working of a choice of algorithms. all algorithm have own significance and use them on the actions of records, except the basis study start from the clustering algorithm pseudocode and classification algorithm pseudocode are compared . As a result necessary the familiarity with algorithms for working. Using four clustering algorithms and four classification algorithms well-known rule, tree and density based ,techniques namely BFTREE, AD Tree, PART, RIDOR, COBWEB, HIERARCHICAL, OPTICS, DBSCAN. As of experimental results formulate virtual study of these algorithms and their applicability on these databases. Compare Algorithms theoretical and experimental parameters such as time taken, clustered instances,etc. beginning of these data sets analyzed algorithms on four different datasets and from that conclude classification algorithms is better than clustering algorithms. Classification algorithms are less time taken to build model according to clustering algorithms. Calculate these algorithms on multi datasets and there results shown through tables. represent of this estimation is good clustering algorithm as classification technique is most popular technique for using different fields like Insurance,medical,banking *etc*.

VIII. FUTURE WORK

In Future we will implement and evaluate four clustering algorithms and four classification algorithms well-known rule, tree and density based techniques namely BFTREE, AD Tree, PART, RIDOR, COBWEB, HIERARCHICAL, OPTICS, DBS-CAN algorithm and will be able to present the results of these algorithms with practical examples. After this we will compare all these eight algorithms with practical examples and find that which algorithm is best among these eight algorithms. Take different parameters. As a future work we can carry out the same result using other types of database types as in this paper only sequential database is used.

REFERENCES

[1] Catlett, 1992. Ripple-Down-Rules as a mediating representation in interactive induction. In Proceedings of the Second Japanese Knowledge Acquisition for Knowledge-Based Systems Workshop, Kobe, Japan.

[2] Compton, P., Edwards, G., et al., Ripple down rules: turning knowledge acquisition into knowledge maintenance, Artificial Intelligence in Medicine **4**: 47-59.

[3] Compton, P., Edwards, G., Kang, B., Lazarus, L., Malor, R. 1991. Ripple down rules: Possibilities and limitations.

[4] Frank, E., Witten, I., J.1998. Generating Accurate Rule Sets without Global Optimization, Machine Learning: Proceedings of the Fifteenth International Conference, pp.144-151, Madison, Wisconsin, Morgan Kaufmann, San Francisco.

[5] Furnkranz, J., Widmer, G. 1994. Incremental Reduced Error Pruning. In Machine Learning: Proceedings of the 11th Annual Conference, New Brunswick, New Jersey, Morgan Kaufmann.

[6] Gaines, B.R., Compton, P.J. 1992. Induction of Ripple Down Rules, AI Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, World Scientific, Singapore.

[7] Gaines, B.R., Paul Compton, J. 1995. Induction of Ripple-Down Rules Applied to Modeling Large Databases, Intell. Inf. Syst. 5(3):211-228.

[8] H. Shi, "Best-first decision tree learning," Citeseer, 2007.

[9] N. Landwehr, M. Hall, and E. Frank, " Logistic model trees". for Machine Learning.,Vol. 59(1-2),pp.161-205, 2005.

[10] ―S.HanumanthSastry‖ and ―Prof.M.S.PrasadaBabu‖. ―Cluster Analysis of Material Stock Data of Enterprises.‖ ―International Journal of Computer Information Systems (IJCIS)‖ 6.6 (2013): pp. 8-19

[11] LiorRokach, OdedMaimon, ―DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK‖ 2010, Springer USA; pp 322-350

[12] M. Thangaraj, Ph.D.and C.R.Vijayalakshmi ," Performance Study on Rule-based Classification Techniques", *Volume 5– No.4, March 2013*

[13] D. L. Gupta, A. K. Malviya and Satyendra Singh," Performance Analysis of Classification Tree Learning Algorithms", *Volume 55– No.6, October 2012*

[14] Helmut Horacek, "Best-First Search Algorithm for Generating Referring Expressions" Helmut HoracekUniyersitat des Saarlandes, FR 6.2 Informatik Postfach 151150, D-66041 Saarbdicken, Germany

[15] Priyanka Sharma," Comparative Analysis of Various Decision Tree Classification Algorithms," Volume: 3 Issue: 2.

[16] S.HanumanthSastry and Prof.M.S.PrasadaBabu," ANALYSIS & PREDICTION OF SALES DATA IN SAP-ERP SYSTEM USING CLUSTERING ALGORITHMS **"**Vol.1, No.4, November 2013

[17] Kaushik H. Raviya and Kunjan Dhinoja," An Empirical Comparison of K-Means and DBSCAN Clustering Algorithm" Volume : 2 | Issue : 4 | April 2013 ISSN - 2250-1991

[18] Narendra Sharma 1, Aman Bajpai 2, Mr. Ratnesh Litoriya," Comparison the various clustering algorithms of wekaTools' 3ISSN 2250-2459, Volume 2, Issue 5, May 2012

[19] Mr.Mitesh Thakkar and Prof.J.S.Shah ," Logistic Model Tree: A Survey"Vol. 2, Issue 2,Mar-Apr 2012, pp.588-594

[20] Md. Mostofa Ali Patwary1,†, Diana Palsetia1, Ankit Agrawal1,Wei-keng Liao1, Fredrik Manne2, Alok Choudhary1**,**"Scalable Parallel OPTICS Data ClusteringUsing Graph Algorithmic Techniques.Northwestern University, Evanston, IL 60208, USA 2University of Bergen, Norway

[21] Ryan P. Adams,"Hierarchical Agglomerative Clustering".

[22] Anilu Franco-Arcega1, Guillermo S´anchez-D´ıaz1, Jos´e Ruiz-Shulcloper2,"ADT: A decision tree algorithm based on concepts" AUGUST 25-28, 2006.

