View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by CiteSeerX

International Journal of Database Management Systems (IJDMS) Vol.5, No.6, December 2013

COLUMN-STORE: DECISION TREE CLASSIFICATION OF UNSEEN ATTRIBUTE SET

Tejaswini Apte[1], Dr. Maya Ingle[2] and Dr. A.K. Goyal[2]

[1]Symbiosis Institute of Computer Studies and Research
[2]Devi Ahilya Vishwavidyalaya, Indore

ABSTRACT

A decision tree can be used for clustering of frequently used attributes to improve tuple reconstruction time in column-stores databases. Due to ad-hoc nature of queries, strongly correlative attributes are grouped together using a decision tree to share a common minimum support probability distribution. At the same time in order to predict the cluster for unseen attribute set, the decision tree may work as a classifier. In this paper we propose classification and clustering of unseen attribute set using decision tree to improve tuple reconstruction time.

1. INTRODUCTION

Decision Tree algorithm is efficient technique for classifying data and providing the accuracy in relatively short time. It acts as a decision support tool by identifying most suitable strategy to reach the goal. Decision trees have been used in different areas of computations. However, big data warehouse with large number of attributes in a relation may be a problem for decision tree classification, as unseen attributes are overlooked. To solve this problem, frequent attribute set generation algorithms may be used to cluster the attribute set, hence improving the accuracy of unseen attribute set. Although these frequent attribute set generation algorithms provide additional information about the domain, that knowledge is never used in decision tree classification. To improve the accuracy of decision trees, several approaches have been developed from different perspective. Clustering algorithms add greater flexibility, although result and time complexity may vary. The aim of clustering algorithms is to group the data according to similarity, hence it is an important tool for unsupervised learning [2, 3, 5].

Exploratory data analysis is a major outcome of successive clustering. Modern tuple reconstruction method DTFCA is typically utilized for clustering frequent attribute sets [1]. Since there is large number ad-hoc queries, some of the clusters may get too small amount of attributes. When a attribute set does not occur in the frequent attribute sets, the closest cluster is being found to classify it. The combination of clustering and decision trees offers more efficiency and accuracy in classification process. DTFCA approach, exploits decision tree to cluster frequently accessed attributes of a relation and hence reduces tuple reconstruction time [1].

In this paper we extended the DTFCA, for classifying the unseen attribute set. The paper is organized as follows. The related work to our approach is presented in Section 2. Section 3

DOI : 10.5121/ijdms.2013.5603

presents the extended version of DTFCA. The experimental environment and analysis is discussed in Section 4. Finally, the paper ends with concluding remarks in Section 5.

2. RELATED WORK

DTFCA tries to make cluster strategy based on minimum support analysis and classification, for selecting frequent attribute set from a list of frequent ad-hoc queries [1]. To improve the performance of clustering-based classification algorithm preprocessing methods are used [2]. The detail discussion to choose optimal cluster for efficient classification from entropy measures of dataset is presented in [3]. Literature reveals hybrid classifier system, which includes artificial intelligence techniques with decision trees. Fuzzy decision tree and genetic algorithms are frequently used in the literature to construct decision tree for data classification in various database applications [4]. The main idea is to convert the historic database into smaller clusters based on fuzzy decision rules, hence increases the accuracy.

To successfully solve clustering/classification, a method based on swarm optimization was proposed by [5]. This method combines the swarm optimization, rough set theory and Huang index algorithm. It achieves clustering and classification optimally. To improve the performance of K-means and Bisecting K-means, an algorithm "cooperative bisecting k-means" was proposed by [6]. A hybrid model for case based data clustering and fuzzy decision tree was proposed by [7]. For homogeneity the data set was pre-processed. A fuzzy decision tree and genetic algorithm are then applied to improve the accuracy. Genetically optimized cluster oriented soft decision tree to develop granules was proposed by [8]. Deployment is happened on synthetic and machine learning data sets to validate the decision tree. To improve the quality of results in classification/clustering tasks, hybrid system was explored by [9]. Despite clustering being a popular and commonly used technique that is applied to different fields, no works have been reported in which clustering is used to improve the accuracy of decision trees for unseen attribute set, to improve tuple reconstruction time in column-store. For this reason, this paper makes an original an important contribution.

3. EXTENDED DTFCA ALGORITHM

Our proposed algorithm is an extension to DTFCA [1], which is majorly focus on to classify unseen attribute set. Correlativity and minimum support are the parameters used to obtain good clusters of unseen attribute set, without the necessity to traverse full attribute set. It takes the accessed attribute list, attribute list in relation, minimum support, correlativity threshold as an input and output the cluster of unseen attribute set [Code snippet 1]. Table 1 defines the sets of attributes with correlativity for given minimum support for TPC-H dataset, without having to run the whole classifier for each attribute.

CODE SNIPPET 1

```
function   unseenattributeset(String S)
{
/*Function to determine unseen attribute set*/
Input :   a collection of accesses, correlativity, Minimum support
Output: clusters of unseen attribute set.
for each access in A      //Processing an element per iteration
{
compute  unclassified attribute set N for correlativity < threshold correlativity and for given
minimum support
for i=0 to N-1
{
string ResT;
ResT= A[i].getName(); //Getting the item name of tree node for attribute
return
}
}
```

4. EXPERIMENT DETAILS

The objective of the experiment is to compare the execution time of existing tuple reconstruction method with extended DTFCA for unseen attribute set on column-stores DBMS.

4.1. Experimental Environment

Experiments are conducted on 2.20 GHz Intel® Core™2 Duo Processor, 2M Cache, 1G Memory, 5400 RPM Hard Drive, Monet DB, a column oriented database and Windows® XP operating system.

4.2. Experimental Data

TPC-H data set is used as the experiment data set, which is generated by the data generator. Given a TPC-H Schema, fourteen different queries are accessed 140 times during a given window period.

Table1: Access Frequency Matrix

Attr Que	S_supplierkey	P_partkey	O_orderdate	l_orderkey	c_custkey	n_nationkey	Correlativity (%) (For Minimum Support 35%)	Correlativity (%) (For Minimum Support 42%)
	Acc_Fq	Acc_fq	Acc_Fq	Acc_Fq	Acc_Fq	Acc_Fq		
2	1	1	0	0	0	1	50	66
3	0	0	1	1	1	0	50	33
4	0	0	1	1	0	0	0	33
5	1	0	1	1	1	1	83	100
6	0	0	0	0	0	0	0	0
7	0	0	0	1	1	1	50	33
8	1	1	1	1	1	1	100	100
9	1	1	0	1	0	1	66	100

10	0	0	1	1	1	1	66	66
12	0	0	0	0	0	0	0	0
16	1	0	0	0	0	0	16	33
17	0	1	0	0	0	0	16	0
19	0	1	0	0	0	0	16	0
21	1	0	0	1	0	1	50	100

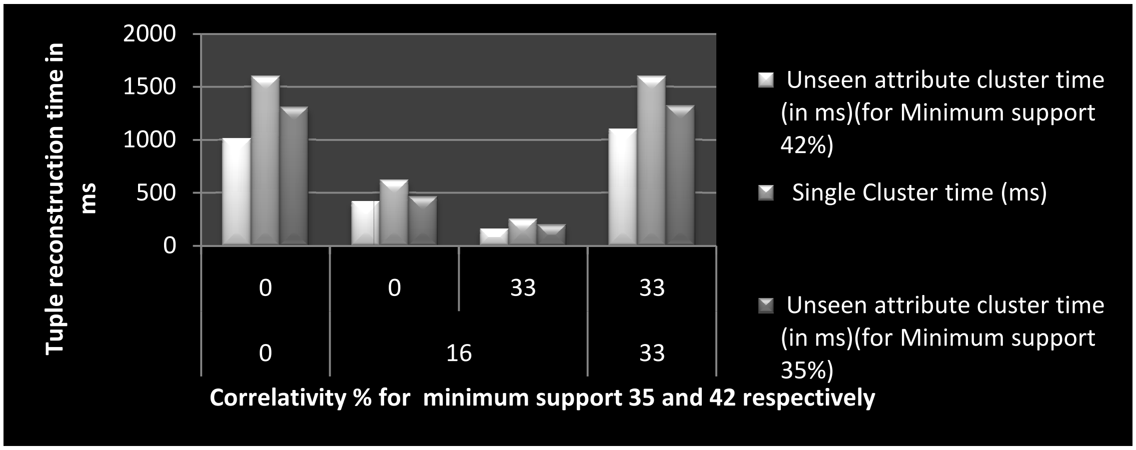

Figure 1: Tuple reconstruction time for unseen attribute set

4.3 Experimental Analysis

Six attributes with access frequency are presented in Table 1. Two parameters namely minimum support and correlativity have used for classification and clustering respectively. We can observe that tuple reconstruction time for unseen attribute set is greatly improved by 80% through classification and clustering. Repeating the experiments by changing the class of minimum support instead of the number of clusters yields the results are shown in Figure 1. In most cases, tuple reconstruction time is directly proportionate to minimum support and correlativity.

5. Conclusions

In this paper we have presented an extended version of DTFCA[1]. Classifying the unseen attribute set according to correlativity limits the scope for attribute search during tuple reconstruction in column-stores, hence improves tuple reconstruction time. However, more experiments are needed to improve the accuracy of results.

References

[1] Tejaswini Apte, Dr. Maya Ingle, Dr. A.K.Goyal “Decision Tree Clustering: A Column-Stores Tuple Reconstruction, Computer Science and Information Technology Vol3 Number 8 ,(2013), 295–303.

[2] Wang, J.-S., and Chiang, J.-C. “An Efficient Data Preprocessing Procedure for Support Vector Clustering”; Journal of Universal Computer Science 15, (2009),705–721.

[3] Kajdanowicz, T., Plamowski, S., and Kazienko, P. “Training set selection using entropy based distance”; In IEEE Jordan Conference On Applied Electrical Engineering and Computing Technologies (AEECT), (2011), 1 –5.

[4] Pei-Chann, C., Chin-Yuan, F. and Wei-Yuan, D. “A CBR-based fuzzy decision tree approach for database classification”; Expert Systems with Applications, 37, (2011), 214–225.

[5] Kuang, Y.H. “A hybrid particle swarm optimization approach for clustering and classification of datasets”; Knowledge-Based Systems, 24, (2011), 420–426.

[6] Kashef, R. and Kamel, M.S. “Enhanced bisecting k-means clustering using intermediate cooperation”; Pattern Recognition, 42, (2009), 2557-2569.

[7] Chin-Yuan, F., Pei-Chann, C., Jyun-Jie, L. and Hsieh, J.C. “A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification”; Applied Soft Computing, 11, (2011), 632–644

[8] Shukla, S.K. and Tiwari, M.K. “Soft decision trees: A genetically optimized cluster oriented approach”; Expert Systems with Applications, 36, (2009), 551–563.

[9] Kajdanowicz, T., and Kazienko, P. “Hybrid Repayment Prediction for Debt Portfolio”; In Computational Collective Intelligence, Semantic Web, Social Networks and Multiagent Systems, (2009), 850–857.

AUTHORS

Tejaswini Apte received her M.C.A(CSE) from Banasthali Vidyapith Rajasthan. She is pursuing research in the area of Machine Learning from DAU, Indore,(M.P.),INDIA.. Presently she is working as an ASSISTANT PROFESSOR in Symbiosis Institute of Computer Studies and Research at Pune. She has 7 papers published in International/National Journals and Conferences. Her areas of interest include Databases, Data Warehouse and Query Optimization.
Mrs. Tejaswini Apte has a professional membership of ACM

Maya Ingle did her Ph.D in Computer Science from DAU, Indore (M.P.) , M.Tech (CS) from IIT, Kharagpur, INDIA, Post Graduate Diploma in Automatic Computation, University of Roorkee, INDIA, M.Sc. (Statistics) from DAU, Indore (M.P.) INDIA. She is presently working as PROFESSOR, School of Computer Science and Information Technology, DAU, Indore (M.P.) INDIA. She has over 100 research papers published in various International / National Journals and Conferences. Her areas of interest include Software Engineering, Statistical, Natural Language Processing, Usability Engineering, Agile computing, Natural Language Processing, Object Oriented Software Engineering. interest include Software Engineering, Statistical Natural Language Processing, Usability Engineering, Agile computing, Natural Language Processing, Object Oriented Software Engineering.
Dr. Maya Ingle has a lifetime membership of ISTE, CSI, IETE. She has received best teaching Award by All India Association of Information Technology in Nov 2007.

Arvind Goyal did his Ph.D. in Physiscs from Bhopal University,Bhopal. (M.P.), M.C.M from DAU, Indore, M.Sc.(Maths) from U.P.
He is presently working as SENIOR PROGRAMMER, School of Computer Science and Information Technology, DAU, Indore (M.P.). He has over 10 research papers published in various International/National Journals and Conferences. His areas of interest include databases and data structure. Dr. Arvind Goyal has lifetime membership of All India Association of Education Technology