ISSN: 2321-7782 (Online)
Volume 3, Issue 10, October 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

The Study of Efficiency of IMA Using ACF

Ashish Saxena[1]
Ph.D. Scholar Shepherd Institute of Engg, &Technology, SHIATS(AAI-DU) , Allahabad, India

Gurmit Singh[2]
Prof Emeritus , Deptt of Computer Sc. and Information Tech, SSET, SHIATS (AAI- DU)NAINI ,Allahabad -211007, India

Abstract: In the case of large datawarehouse having large incremental dataset the support of algorithm should be high for increasing the efficiency of DELTA an extra pass is added using count function which increases the efficiency of IMA [incremental data mining algorithm] is analyzed using ACF.

Key words: Count, Updatecount, Getfrequent, Apriorifunction, ACF, Upper Confidence, lower Confidence.

I. INTRODUCTION

In the 20th Century as the size of databases storing information increased, the concept of a large unorganized set of huge data came into existence, which is often called a data warehouse. To mine the data from the unorganized stored data, the need of mining algorithms arose. This led to the development of fast, more efficient, reliable and less time consuming algorithms. Data mining, the extraction of hidden predictive information from large data warehouses, is a powerful new technology with a greater potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive and knowledge-driven decisions. Data mining techniques are the result of a long process of research and product development. This evolution began when business data were first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through the data in real time. Incremental Data Mining based algorithm DELTA was not very efficient. Besides this, it is time consuming and also utilizes large memory space .Keeping in view of the above problems, the need of developing of faster, more-reliable and less memory- time consuming algorithm arises for the case of incremental data mining. The emphasis is laid on developing an algorithm which mines data with more precision and in lesser time as compared to the existing DELTA algorithm. The developed algorithm must be capable of discovering association rules more efficiently in the case of incremental data-mining in lesser time as size of incremental data warehouse increases rapidly, which increases the efficiency of data mining. There is thus a need to develop more efficient algorithm for incremental data mining.

II. REVIEW OF LITERATURE

In the 20th Century the size of data warehouse increased due to large number of events involved in industry. For this the process of extracting valuable information for huge time variant unorganized data has increased. This led to the development of fast data mining algorithms. These enabled fast data mining from large online warehouses. Various studies have been carried out by the investigators in the area of data mining .

Hosseini et.al., (2010) worked on Cluster analysis using data mining approach to develop and analyses large data of customers using association rules as a case of incremental data mining. ***Chen et.al. (2011)*** worked on Incremental learning from stream data and used the concept of tree like structure. ***Mark et al. (2011)*** worked on a financial data mining model for extracting customer behavior. He used the concept of clustering as a case of incremental data mining using large warehouse of aims at developing an intelligent Financial Data Mining Model (FDMM) for extracting customer behavior in the financial industry, so as to increase the availability of decision support data and hence increase customer satisfaction. The proposed financial model first clusters the customers into several sectors, and then finds the correlation among these sectors. It was noted

that better customer segmentation can increase the ability to identify targeted customers, therefore extracting useful rules for specific clusters can provide an insight into customers buying behavior and marketing implications. To validate the feasibility of the proposed model, a simple dataset collected from a financial company in Hong Kong. The simulation experiments show that the proposed method not only can improve the workflow of a financial company, but also deepen understanding of investment behavior. Thus, a corporation is able to customize the most suitable products and services for customers on the basis of the rules extracted. ***Sung et al.(2011)*** worked on forecasting changes in Korea Composite Stock Price Index (KOSPI) using association rules. They used the association rule mining for obtaining golden association rules over large data warehouse.

Lin et al.(2012) worked on incremental mining of a short Chinese text using incremental clustering algorithm based on weighted semantics and naive bayes , which worked on a cluster based approach. The availability of large quantity of text documents from the world wide web and business document management systems has made the dynamic separation of texts into new categories as a very important task for every business intelligence systems. But, the presented text clustering algorithms still suffer from problems of practical applicability. In order to improve the performance of document clustering, ontologies are useful. Ontology is nothing but the conceptualization of a domain into an individual identifiable format, but machine-readable format containing entities, attributes, relationships and axioms. By analyzing all types of techniques for document clustering, a clustering technique depending on Genetic Algorithm (GA) is determined to be better as GA is a global convergence technique and has the ability of determining the most suitable cluster centers without difficulties. They proposed a new document clustering scheme with fuzzy ontology based genetic clustering is proposed. They experimental results reveal that the proposed approach increases the accuracy to a large extent and the clustering time is also highly reduced.

III. MATERIAL AND METHODS

The DELTA used for mining have only three passes, however , IMA uses four passes, The IMA in the 1st pass reads previous database and then increments *d(database element)* using the *updatecount function*, by this some item sets may become frequent in N(Negative border database element) .The infrequent itemsets are denoted in *Fknown*. These itemsets are extracted using the function *Getfrequent* .The infrequent itemsets are calculated by Infrequent =(FdbUNdb)- $F_{known.}$.In the 2nd pass of the algorithm these itemsets are used for pruning. If some itemsets do not move from Ndb to Fknown, then the negative border of Ndb of Fknown is computed by using apriori function. Itemsets in Nknown with unknown counts are stored in N^u thus the remaining counts are all infrequent. Any itemset in N^u and their extension are computed. If there are no elements in N^u their extensions. Any itemset which is not frequent in db cannot be frequent in DBUdb. In the 3rd pass all possible extensions of F_{known} which could be in form ((FDBUdb)U (NDBUdb))and store them in set count C. This is done by computing the layers of negative borders closure of Fknown .It is expected that all the other remaining layers can be generated together since the number of the two itemsets in Fknown is typically much smaller than the total number of 2-itemsets pairs. Initially C is reset to zero using the function of reset count. Then at every stage of computation of closure, those itemsets that are in Infrequent and Infrequent db are removed so that none of the extensions are generated. Itemsets from F_{known} and N^u are removed from C. In this pass the counts within db of the remaining itemsets C are computed. The 4th pass scans the set of minsup *| D U d | and returns F_{DBudb}(frequent itemsets) and N_{DBudb}(Negative Border) values, which gives a unique value of support .This value of support is used to evaluate the performance of the algorithm.

```
12.  count= renew count(d,c);// IIIrd extra proposed pass Starts
13.   for C>0
14.  C=C+F known;
15.   Reset count (C );
16.  C= C+ Negative border(C);
```


17. C= C- (InfrequentU $Infrequent_{db}$)

18. C= C- (F_{known} +N^u)

19. if C >0 then updatecount(db,C);

20. ScanDB= Getfrequent (CUN^U, minsup *|db|);

The standard synthetic data warehouse T40I10d100K is used as incremental data warehouse. The optimization is performed only in routines that access the database and do not effect the structure of incremental mining algorithm. The second optimization is performed before each pass over the increment of previous database. The successor of items that are not from candidate are totally removed from the arrays of successor that are earlier computed during the first optimization.

IV. FINDING

The ACF analysis of DELTA and IMA was done on SPSS16.0 we get the following result. In the tabular form as well as in the graphs generated by SPSS 16.00 software.

Table 1 ACF analysis of DELTA obtained in SPSS16.00

Syntax		ACF VARIABLES=DELTA /NOLOG /MXAUTO 16 /SERROR=IND /PACF.
Time Series Settings (TSET)	Amount of Output	PRINT = DEFAULT
	Saving New Variables	NEWVAR = NONE
	Maximum Number of Lags in Autocorrelation or Partial Autocorrelation Plots	MXAUTO = 16
	Maximum Number of Lags Per Cross-Correlation Plots	MXCROSS = 7
	Maximum Number of New Variables Generated Per Procedure	MXNEWVAR = 60
	Maximum Number of New Cases Per Procedure	MXPREDICT = 1000
	Treatment of User-Missing Values	MISSING = EXCLUDE
	Confidence Interval Percentage Value	CIN = 95
	Tolerance for Entering Variables in Regression Equations	TOLER = .0001
	Maximum Iterative Parameter Change	CNVERGE = .001
	Method of Calculating Std. Errors for Autocorrelations	ACFSE = IND
	Length of Seasonal Period	Unspecified
	Variable Whose Values Label Observations in Plots	Unspecified
	Equations Include	CONSTANT

(a)

Model Description		
Model Name		MOD_5
Series Name	1	DELTA
Transformation		None
Non-Seasonal Differencing		0
Seasonal Differencing		0
Length of Seasonal Period		No periodicity
Maximum Number of Lags		16
Process Assumed for Calculating the Standard Errors of the Autocorrelations		Independence(white noise)a
Display and Plot		All lags
Applying the model specifications from MOD_5		
a. Not applicable for calculating the standard errors of the partial autocorrelations.		

(b)

Case Processing Summary		DELTA
Series Length		24
Number of Missing Values	User-Missing	0
	System-Missing	0
Number of Valid Values		24
Number of Computable First Lags		23

(c)

Autocorrelations					
Series:DELTA					
Lag	Autocorrelation	Std. Errora	Box-Ljung Statistic		
			Value	df	Sig.b
1	.377	.192	3.864	1	.049
2	.259	.188	5.760	2	.056
3	.036	.179	6.836	4	.145
a. The underlying process assumed is independence (white noise).					
b. Based on the asymptotic chi-square approximation.					

(d)

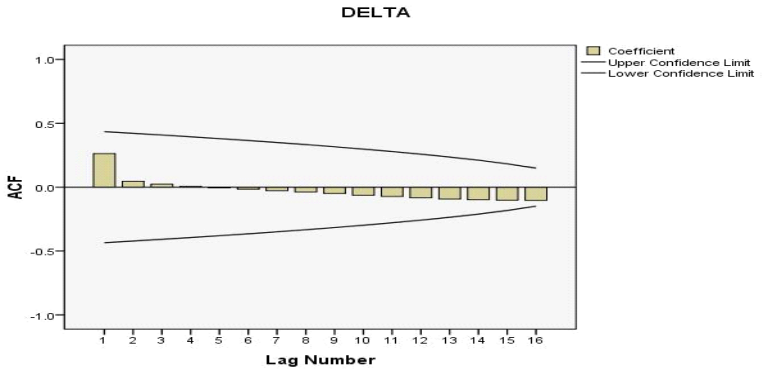


Fig 1: ACF of DELTA

The above figure of ACF of DELTA is generated on SPSS 16.0.It is clear that the DELTA has more deviation from normal there by having low performance .The value of upper confidence limit is less and the lower confidence limit increases as the lag number increases .It causes decrease in the efficiency of mining as the dataset increase in incremental data mining, from the figure 1 it is evident that DELTA efficiency decreases as the lag number increases.

Case Processing Summary		IMA
Series Length		24
Number of Missing Values	User-Missing	0
	System-Missing	0
Number of Valid Values		24
Number of Computable First Lags		23

Table 2 ACF analysis of IMA on SPSS16.00

Output Created		2013-07-25T10:34:38.781
Comments		
Input	Data	C:\Documents and Settings\ASHISH SAXENA\Desktop\FINALAMRESHPRINT16102011\ALGO1\2012pd.sav
	Maximum Number of New Cases Per Procedure	MXPREDICT = 1000
	Treatment of User-Missing Values	MISSING = EXCLUDE
	Confidence Interval Percentage Value	CIN = 95
	Tolerance for Entering Variables in Regression Equations	TOLER = .0001
	Maximum Iterative Parameter Change	CNVERGE = .001
	Method of Calculating Std. Errors for Autocorrelations	ACFSE = IND
	Length of Seasonal Period	Unspecified
	Variable Whose Values Label Observations in Plots	Unspecified
	Equations Include	CONSTANT

Model Description		
Model Name		MOD_6
Series Name	1	IMA
Transformation		None
Length of Seasonal Period		No periodicity
Maximum Number of Lags		16
Process Assumed for Calculating the Standard Errors of the Autocorrelations		Independence(white noise)a
Display and Plot		All lags

Table 3 ACF analysis Using Box-Ljung Statistic of IMA on SPSS16.00

Autocorrelations					
Series:IMA					
Lag	Autocorrelation	Std. Errora	Box-Ljung Statistic		
			Value	df	Sig.b
1	.721	.192	14.107	1	.000
2	.476	.188	20.521	2	.000
3	.282	.183	22.879	3	.000
4	.054	.179	22.968	4	.000
5	-.018	.174	22.979	5	.000
6	-.024	.170	22.999	6	.001
7	-.055	.165	23.112	7	.002
a. The underlying process assumed is independence (white noise).					
b. Based on the asymptotic chi-square approximation.					

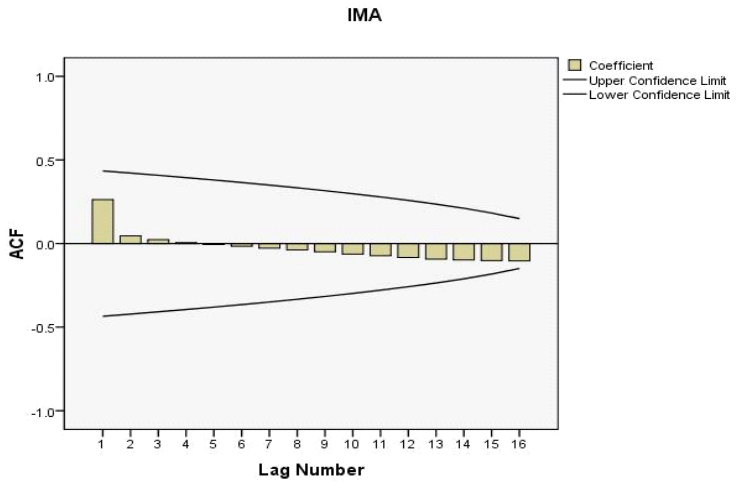


Fig 2 ACF of IMA

From the ACF it is clear that IMA is having less deviation from normal so , IMA has more efficient then as compared with DELTA. The value of upper confidence level for DELTA is higher as the autocorrelation function lag number increases .The lower confidence level gradually decreases as the lag number increases. Thus, it is evident from the figure 2 the IMA holds higher performance than that of DELTA.

V. Summary and Conclusion

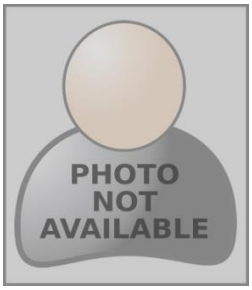
From the above analysis it is obvious that IMA is more efficient than that of DELTA when the dataset increases as IMA is having a small deviation from normal whereas in the case of DELTA deviation is large.

References

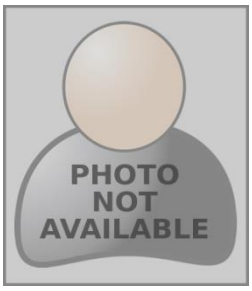
1. Hosseini, S.M.S., Maleki, A., Gholamian, M.R., (2010) "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty", Expert Systems with Applications, Vol. 37, no.7, 5259-5264p, 2010
2. H. He, S. Chen, K. Li, X. Xu,(2011) Incremental learning from stream data, IEEE Transactions on Neural Networks, Vol.22(12), 1901-1914p,2011
3. Mark K.Y. Mak , George T.S. Ho and S.L. Ting (2011) " A Financial Data Mining Model for Extracting Customer Behavior" ,International journal of engineering , business management., 2011, Vol. 3, No. 3,59-72p,2011
4. Sung, H.N., So, Y. S, (2011) " Forecasting changes in Korea Composite Stock Price Index (KOSPI) using association rules", Expert Systems with Applications, vol. 38, no. 7, 9046–9049p,2011

5. P. Lin, Z. Lin, B. Kuang, P. Huang,(2012) A Short Chinese Text Incremental Clustering Algorithm Based on Weighted Semantics and Naive Bayes, Journal of Computational Information Systems, 2012, 4257- 4268p,2012

AUTHOR(S) PROFILE



Ashish Saxena, is Ph.D Scholar SSET ALLAHBAD. He has published many research paper in the field of Incremental Data Mining. He has done M.Sc [Computer Science] from GBPUA&T, PANTNAGAR. He is life member of I.S.T.E, New Delhi



Prof (Col. Gurmit singh), competed MTech from J.N.U, NEW DELHI, Currently Emeritus Fellow at SSET, SHIATS, ALLAHABAD.

