ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Bagging and Distributed Association Rule Mining Techniques for Distributed Data Mining on Homogeneous Datasets

S.Urmela
Ph.D Scholar
Department of Computer Science,
Pondicherry University,
Puducherry, India

Dr.M.Nandhini
Assistant Professor
Department of Computer Science,
Pondicherry University,
Puducherry, India

Abstract: Data Mining (DM) is the process of extracting useful unknown information from data using patterns. Distributed Data Mining (DDM) evolved from DM in recent years to mine geographical distributed data. Mining can be performed either on geographical distributed data with same attribute or different attribute. This paper implements DDM based on bagging and Distributed Association Rule Mining on soy-bean, iris and contact-lens datasets having same attribute across geographical distributed sites.

Keywords: Distributed Data Mining; classifier approach; bagging; Distributed Association Rule Mining

I. INTRODUCTION

DDM is the process of extracting useful, unknown information from data which is geographically distributed and whose data will be having either same set of attributes or different set of attributes. DDM aims to minimize computation time and promises minimum memory cost[1].

DDM evolved from DM since the necessity of mining information from geographical distributed data came into existence to lead DM to next step. Inspired by the area in this paper we implemented bagging and Distributed Association Rule Mining for geographical distributed data with homogeneous attribute namely soy-bean, iris and contact-lens datasets.

A.O. Ogunde et al.(2015)[2] discussed a partition enhanced mining algorithm for DARM that involved agents for association rule mining. Agents will be assigning coordinating agents thereby requests received will be forwarded to coordinating agents which will determine the targeted geographical sites. Dr. C.Sunil Kumar et al.(2013)[3] discussed an Apriori algorithm for DARM that involved XML data distributed mining. Though there are certain cons in mining XML data, the proposed algorithm OARM (Optimal Association Rule Mining), involves parallel mining process thereby achieving promising computation time and minimal communication cost.

Philip K. Chan et al.(1999)[4] discussed credit card fraudulent detection by detecting fraud credit card transactions by maintaining frequent transactions patterns across distributed geographic sites. The method proposed is a scalable and efficient technique, thereby generating base-classifiers learning model. Meta-learning classifier approach is adopted; predictive learning models are derived from base classifiers. Several base-classifiers can operate in parallel with global level meta-classifier. Even highly-skewed data has been studied considered in this approach. Pros of this proposed work is efficient, scalability, and a cost- effective approach. Cons of the proposed work are implementation of adaptive technique.

Frank S.C. Tseng et al.(2010)[5] discussed boosting based DARM by data de-clustering which involves de-clustering data where data will be clustered into partitions. Round-robin assigning method will be followed for iterative assignment of datasets to each participating data-sites geographically distributed. Pros of the proposed work are minimal communication cost and reduced space complexity.

The organization of the paper is as follows: Section II discusses the related works of DDM based on bagging and association rule mining techniques. Section III depicts the implementation details of bagging and Distributed Association Rule Mining techniques of DDM. Section IV summarizes the paper.

I. RELATED WORKS

Related works in bagging and Association Rule Mining techniques is discussed here. In this paper we implemented bagging and Distributed Association Rule Mining technique of DDM on soy-bean, iris and contact-lens datasets.

A. Bagging

L. Breiman (1996)[6] discussed predictors for bagging explained method for multiple versions generation of predictor. It is useful for generation of aggregated predictor. Average of aggregation over distinct versions will be considered for numerical outcome prediction and s true voting is done when class prediction is performed. Multiple versions are formed while bootstrap replicating of learning set. Experimental results on real datasets by traditional techniques on selection of subset and regression tree shows that bagging gives promising accuracy.

J.R. Quinlan(1996)[7] discussed on bagging and boosting techniques which reported results of both bagging and boosting techniques to systems which frame decision trees and testing is done on datasets considered. Though

both techniques aim in improved prediction accuracy, boosting shows better performance. Cons of boosting is that it produces dataset modification in targeted datasets but combination of classifiers at local level can show a better performance in boosting and thereby forecasting improved prediction results.

Pooja Shrivastava et al.(2014)[8] presented the accuracy analyzing on forest-fire database for UCI machine learning by bagging technique. The work is done with bagging technique on WEKA tool with forest-fire dataset. The work showed better precision and recall values compared to traditional techniques.

Mrinal Pandey et al.(2014)[9] presented a comparative study on ensemble techniques for students' academic performance modeling which examined accuracy of ensemble techniques for predicting students' academic performance for 4 years engineering graduate program. Traditional ensemble techniques namely bagging, AdaBoost, Rotation Forest and Random Forest have been used to construct and combine different number of ensembles.

These four algorithms have been compared for 10 base classifiers. Bagging shows a better ensemble classifier result for predicting student performance for students' performance modeling.

BadrHssina et al.(2014)[10] discussed comparative study of C4.5 and decision tree. The work presents algorithm comparison of C4.5 and ID3 along with a comparison work of CART and C5.0. On comparison of algorithms, C4.5 shows better prediction performance.

B. Association Rule Mining

Paresh Tanna et al.(2014)[11] discussed Apriori algorithm on WEKA tool for frequent pattern mining. It presents sample algorithm usage by WEKA tool and finally discusses applications of association rule mining.

Charanjeet Kaur(2013)[12] discussed a survey work of association rule mining. It presents survey of works on Apriori algorithm done by several researchers. The work result can vary depending on datasets alteration and candidate variation.

Divya Bansal et al.(2014)[13] discussed Apriori algorithm on Tumultuous Crimes Concerning Women which elaborates use of Apriori algorithm by WEKA tool and experimental results shows that general Apriori algorithm forecast better performance than predictive Apriori algorithm. Further association rule mining helps in identifying victims based on age group, accused age group, stranger, etc, thereby helping in improving the deterioration condition of crime against women.

Stephen M. Kang'e the et al.(2011)[14] discussed on extraction of patterns for diagnosing on Electronic Media by association rule mining. Apriori algorithm for mining associations rule in patients Electronic Media Records is done and shows association rule which can be useful in generating probabilistic statements like: "If patient is undergoing treatment T, then there is 0.5 probability value that they are diagnosed with disease D".

Umesh Kumar Pandey(2013)[15] discussed DM for class-room teaching, thereby 7 association rules are generated and results show that mix-medium concept is more preferred than hindi-medium or english-medium alone. Next section focuses on experimental implementation of bagging and DARM with datasets considered along with experimental results.

I. EXPERIMENTAL IMPLEMENTATION

Bagging was implemented on soy-bean and iris datasets and DARM on contact-lens and soy-bean datasets. The 2 homogeneous classifier approach was implemented on Java language with Eclipse Mars. IDE on WEKA API version 3.7.

A. Datasets description

IRIS dataset: The iris dataset consists of 150 instances, 4 attributes and 3 classes. The classes are various types of the iris plant. All attributes are numeric except for the class which is nominal. No missing values are present.

SOY-BEAN dataset: The soybean data set consists of 683 instances, 35 attributes and 19 classes. The classes are various types of soybean diseases. The attributes are observations on the plants together with some climatic variables. All attributes are nominal. Some missing values were filled-in by their modal values. Plants were measured on 35 attributes and there are 19 disease categories like stem-canker, root rotting, bacterial infections, etc.

CONTACT-LENS dataset: This dataset allows optician to prescribe lens by parameters like information about a patient's age, tear production rate, whether the patient is suffering from astigmatism or not. It allows optician to decide whether to prescribe patient either hard, soft contact lens or no contact lens at all.

B. Bagging

Pseudocode: Bagging

Bagging(D,T)
Input:
D – Datasets targeted
T – no. of training sets considered
Lb – base-model
Hm – base model generated
S – bootstrap set
$h_{fin}(x)$ – final generated model

For each t = 1,2,...,T
D_t = Sample_with_replacement(D, |D|)
H_t = $L_b(D_t)$
Return $h_{fin}(x) = \text{argmax}_{y \in Y}$? I ($h_t(x)$=y)
Sample_with_replacement (D, N)
S = {}
For i = 1,2,...,N
r = random_integer(1,N)
Add D[r] to S
Return S

Table 3.1 Bagging on IRIS dataset

Parameters	J48	Bagging
Correctly classified instances	98%	98.67%
Incorrectly classified instances	2%	1.33%
Precision	0.980	0.987
F-measure	0.980	0.987
Kappa statistics	0.97	0.98

Table 3.2 Bagging on SOYBEAN dataset

Parameters	J48	Bagging
Correctly classified instances	96.33%	97.36%
Incorrectly classified instances	3.66%	2.63%
Precision	0.965	0.974
F-measure	0.962	0.974
Kappa statistics	0.97	0.97

Table 3.1 and table 3.2 depicts result of bagging and J48 on soybean (683 instances) and iris (150 instances) datasets. Precision value generated from correctly classified instances and incorrectly classified instances shows a better prediction result in comparison with J48 decision tree of 0.987 and 0.974 for iris and soybean datasets. Similarly, incorrectly classified instances rate is minimized in bagging technique comparing with J48 technique of nearly 1.33% from 2% in iris dataset and 2.63% from 3.66% in soybean dataset. Likewise, F-measure value which is the ratio of Precision and Recall depicts improved value of 0.987 and 0.974 in comparison with J48 value of 0.980 and 0.962. Kappa statistics value shows an improved value.

```
J48 pruned tree
***************
Plant – growth = norm
 Leafspot-size = lt – 1/8
  Canker-lesion = dna
   Leafspots-marg = w-s-marg
    Roots = norm: bacterial-blight (16.68/0.68)
    Roots = rotted: bacterial-pustule (5.21/0.21)
    Roots = galls-cysts: bacterial-blight (0.0)
  Leafspots-marg = no-w-s-marg: bacterial-pustule (18.77/0.77)
  Leafspots-marg = dna: bacterial-pustule (0.0)
  Canker-lesion = brown: bacterial-pustule (0.0)
  Canker-lesion = dk-brown-blk: bacterial-pustule (0.0)
  Canker-lesion = tan: purple-seed-stain (10.0)
 Leafspot-size = gt-1/8
  Mold-growth = absent
   Fruit-pods = norm
    Precip = lt-norm: phyllosticta-leaf-spot (4.0)
    Precip = norm

Summary:
*****************************************************************
Correctly classified instances         665    97.3646 %
Incorrectly classified instances        18     2.6345 %
Kappa Statistics                         0.9711
Mean Absolute Error                      0.0099
Root Mean Squared Error                  0.0536
Relative Absolute Error                 10.2993 %
Root relative squared error             24.4427 %
Coverage of cases (0.95 level)         100 %
Mean rel. Region size (0.95 level)      8.7617 %
Total number of instances              683
```

Fig 3.1 Bagging algorithm result on SOYBEAN dataset

```
J48 pruned tree
***************
Petalwidth <= 0.6: Iris-setosa (48.0)
Petalwidth > 0.6
 Petalwidth <= 1.6
  Petallength <= 5: Iris-versicolor (53.0)
  Petallength > 5: Iris-virginica (4.0/1.0)
 Petalwidth > 1.6: Iris-virginica (45.0)

Number of Leaves  :     4
Size of the tree  :     7

Summary:
*****************************************************
Correctly classified instances         148     98.6667 %
Incorrectly classified instances         2      1.3333 %
Kappa Statistics                         0.98
Mean Absolute Error                      0.0206
Root Mean Squared Error                  0.0893
Relative Absolute Error                  4.6302 %
Root relative squared error             18.9393 %
Coverage of cases (0.95 level)         100 %
Total number of instances              150
```

Fig 3.2 Bagging algorithm result on IRIS dataset

Fig 3.1 depicts bagging algorithm on soybean dataset for which correctly classified instances is 665 and incorrectly classified instances is 18. The attributes are observations on plants along with some climatic variables. Based on 35 variables 19 disease categories are detected. There are 19 disease categories like stem-canker, root rotting, bacterial infections, etc. Say, from fig 1. If canker-lesion = brown,

then the leaves of iris plant is detected to be affected by bacterial-pustule. Disease affected on iris plant is measured by class attributes.

Fig 3.2 depicts bagging algorithm on iris dataset for which correctly classified instances is 148 and incorrectly classified instances is 2. The attribute petalwidth and petallength determines classification of class iris-setosa, iris-versicolor and iris-virginica. If petalwidth value is < = 0.6 then identified class is iris-setosa, if petalwidth value is < = 1.6, petallength value is < = 5, then identified class is iris-versicolor, if petallength value is > 5, then identified class is iris-virginica and if petalwidth value is > 1.6, then identified class is iris-virginica. Thus 4 classes will be framed by bagging technique.

Apriori
****
Minimum support : 0.2 (5 instances)
Number of cycles performed : 16
Generated sets of large itemsets:
Size of set of large itemsets (L1) : 11
Size of set of large itemsets (L2) : 21
Size of set of large itemsets (L3) : 06

Best rules found:
1. Tear-prod-rate = reduced → contact-lenses = none
2. Spectacle-prescrip = myope tear-prod-rate = reduced → contact-lenses = none
3. Spectacle-prescrip = hypermetrope tear-prod-rate = reduced → contact-lenses = none
4. Astigmatism = no tear-prod-rate = reduced → contact-lenses = none
5. Astigmatism = yes tear-prod-rate = reduced → contact-lenses = none
6. contact-lenses = soft → astigmatism = no
7. contact-lenses = soft → tear-prod-rate = normal
8. tear-prod-rate = normal contact-lenses = soft → astigmatism = no
9. astigmatism = no contact-lenses = soft → tear-prod-rate = normal
contact-lenses = soft astigmatism = no → tear-prod-rate = normal

Fig 3.3 DARM algorithm result on CONTACT-LENS dataset

C. DARM

DARM helps to identify the relationship between data-items in a given dataset. Apriori algorithm has been used in this technique to extract frequent item-sets and hence generate association rule for datasets. The key idea of Apriori algorithm is to make iterative pass over the entire dataset. It further employs breadth-first search technique where item-sets are explored. From the first pass, item-count is formulated and collects those items which satisfy minimum support value. Thus L1 is formed. Further, L2 is formed from L1 and so on.

Fig. 3.3 depicts DARM on contact-lens dataset where minimum support value is 10% (0.2) thereby L1 value is 11 (initial breadth-first pass of item-count), L2 value is 21 (dependent on L1) and L3 value is 06 (dependent on L2).

Further from the item-sets generated, association rules are framed by which opticians can prescribe if a patient be given hard-contact lens, soft-contact lens or no contact lens at all from the tear-production rate, patient age and whether person is suffering from astigmatism or not.

Fig. 3.4 depicts DARM on iris dataset where minimum support value is 10% (0.8) thereby L1 value is 06 (initial breadth-first pass of item-count), L2 value is 06 (dependent on L1) and L3 value is 02 (dependent on L2).

Further from the item-sets generated, association rules are framed by which a plant is affected either by sclerotia or mycelium or not is decided based on discoloration, leaves abnormality, etc.

Apriori

****
Minimum support : 0.8 (546 instances)
Minimum metric <confidence> : 0.9
Number of cycles performed : 4

Generated sets of large itemsets:
Size of set of large itemsets (L1) : 06
Size of set of large itemsets (L2) : 06
Size of set of large itemsets (L3) : 02

Best rules found:
1. int-discolor = none → sclerotia = absent
2. mycelium = absent int-discolor = none → sclerotia = absent
3. leaves = abnorm sclerotia = absent → mycelium = absent
4. sclerotia = absent → mycelium = absent
5. int-discolor = none → mycelium = absent
6. int-discolor = none sclerotia = absent → mycelium = absent
7. int-discolor = none mycelium = absent → sclerotia = absent
8. leaf-malf = absent → mycelium = absent
9. mycelium = absent → sclerotia = absent
leaves = abnorm mycelium = absent → sclerotia = absent

Fig 3.4 DARM algorithm result on IRIS dataset

IV. REFERENCES

1. Ms. Vinaya Sawant and Dr. Ketan Shah, “A review of Distributed Data Mining using agents”, International Journal of Advanced Technology & Engineering Research (IJATER), Volume 3, Issue 5, September 2013, pp. 27-33.
2. A.O. Ogunde, O. Folorunso, A.S. Sodiya, “A partition enhanced mining algorithm for distributed association rule mining systems”, Egyptian Informatics Journal, Volume 16, Issue 3, November 2015, pp. 297-307.
3. Dr. C.Sunil Kumar, P.N.Santosh Kumar & Dr. C.Venugopal, “An Apriori Algorithm in Distributed Data Mining System”, Global Journal of Computer Science and Technology Software & Data Engineering, Volume 13, Issue 12, 2013.
4. Philip K. Chan, Wei Fan, Andreas L. Prodromidis, Salvatore J. Stolfo, “Distributed Data Mining in Credit Card Fraud Detection”, IEEE Intelligent Systems, December 1999, pp. 67-74.
5. Frank S.C. Tseng, Yen-Hung Kuo, Yueh-Min Huang, “Toward boosting distributed association rule mining by data de-clustering”, Journal of Information Sciences, Volume 180, Issue 22, November 2010, pp. 4263-4289.
6. Leo Breiman, “Bagging Predictors”, Springer Journal of Machine Learning, Volume 24, Issue 2, August 1996, pp. 123-140.

7. J. R. Quinlan, "Bagging, Boosting, and C4.5", In Proceedings of the Thirteenth National Conference on Artificial Intelligence, 1996.
8. Pooja Shrivastava, Manoj Shukla, "Uses the Bagging Algorithm of classification method with weka tool for prediction technique", In Proceedings of 16th IRF International Conference, October 2014, Chennai, India, pp. 23-27.
9. Mrinal Pandey, S. Taruna, "A Comparative Study of Ensemble Methods for Student's Performance Modeling", International Journal of Computer Applications, Volume 103 Issue 8, October 2014, pp. 26-32.
10. Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, Mohammed Erritali, "A comparative study of decision tree ID3 and C4.5", International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Advances in Vehicular Ad Hoc Networking and Applications 2014.
11. Paresh Tanna, Dr. Yogesh Ghodasara, "Using Apriori with WEKA for Frequent Pattern Mining", International Journal of Engineering Trends and Technology(IJETT), Volume 12, Issue 3, Jun 2014.
12. Divya Bansal, Lekha Bhambha, "Execution of Apriori Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, Sep 2013.
13. Stephen M. Kangethe, Peter W. Wagacha, "Extracting Diagnosis Patterns in Electronic Medical Records using Association Rule Mining", International Journal of Computer Applications, Volume 108, Issue 15, Dec 2014.
14. Umesh Kumar Pandey, "A Data Mining View on Class Room Teaching Language", International Journal of Computer Science Issues, Volume 8, Issue 2, March 2011.
15. Komal khurana, Mrs. Simple Sharma, "A Comparative Analysis of Association Rule Mining Algorithms", International Journal of Scientific and Research Publications, Volume 3, Issue 5, May 2013.