Constructing Scalable local Distributed Decision Trees algorithm for heterogeneous data sources

Dr.E.Chandra

Research Supervisor and Director
Dept. of Computer Science, DJAME
Coimbatore, Tamilnadu, India

P.Ajitha

Research Scholar and Assistant Professor
Dept. of Computer Science, DJAME
Coimbatore, Tamilnadu, India.

Abstract—This papers proposes a new scalable and robust distributed algorithm for constructing distributed decision trees in peer-to-peer environment for the heterogeneous data sources. Computation and communication cost in the peer-to-peer environment is higher and also on chances of reduced accuracy and response time may be higher. Proposed algorithm scales good and also provides the best prediction model in the well known classification technique of distributed decision trees

Keywords- Distributed Decision Trees ,peer-to-peer environment, heterogeneous data

I. INTRODUCTION

Distributed Decision Trees (DDT) is one of the most popular classification techniques of predictive modeling. On considering the flow of tera bytes of information available in various formats and sources integrating the same may lead to inaccuracy of the data. Comparing with the sequential decision tree the distributed decision trees may be able to handle massive large data sets. Distributed Decision trees are basically used for handling complex predictions, easily translatable to human understandable form and also subsets of attributes its description in detail can be specified.

Data are in of various formats and sources. Integrating it in the centralized environment is not a quiet easy task. Computation task and efficient handling of the data may be lost. Knowledge acquisition systems are needed to perform the analysis and transmission of the data in the specified locations.

Peer-to-Peer(P2P) data mining environment considers the distributed[5] resources of data, computing and communication to gather the final optimal result. Large scale data analysis have become the new generation era of advancement and computing resources availability. Primary intention of P2P data mining is to gather the data without necessity to store in the centralized location. Further P2P harness the resources and able to make use of decision trees in each and every node. P2P considers the set of attributes across the networks and choose the attributes that has the closest matching criteria or attributes that match. Currently emerging and most widely efficient generating results is possible through P2P in data mining.

II. RELATED WORK

Distributed Data Mining is the mining of data from the large distributed resources across the different nodes, and also in various formats. Many of the existing algorithms deal with the Bayesian models, ID3 algorithms [2]. Main characteristics of P2P are to have high scalability, reliability and end-to-end communications. Scalable local algorithm was developed for distributed systems in multivariate regression where asynchronous, communication efficient scalable algorithm.

Caragea *et al.* [6] presented a decision tree induction algorithm for both horizontally and vertically distributed data. Crux of any decision tree algorithm is the use of an effective splitting criteria, the authors propose a method by which this criteria can be evaluated in a distributed fashion. Cargea et al[10] discusses the learning for heterogeneous and semantic data sources with sufficient statistics learning. Here the approach is based on the hypothesis generation.

Scalable Local algorithm for distributed decision trees in proposed where multivariate regression is used to generate efficient prediction. Kanishka et al[12] discussed the peer-to-peer environments using regression so that feedback mechanism is also considered to save the communication and computation cost.

Client Server and P2P both generates the same results but P2P is self organised and supports for Ad-hoc networks[5]. Distributed systems in relation with the P2P is also discussed well in Rudiger *et.al*. Distributed architecture is a called as Peer-to-Peer(P2P,P-to-P) if the participants share some of the processing power, storage capacity and the like. These resources can be shared either using providers or services. On the message passing, without any intermediary content it can be communication with the available resources.

III. PROPOSED ALGORITHM

To select the best attributes in the distributed environment for the decision trees is a necessary step towards the efficient generation of predictive model.

A. Proposed Distributed Decision trees Algorithm

P2P environments are quite interesting and efficient when the computation and communication cost are under consideration. DDT in P2P is an classification technique to share the distributed resources in the disseminated environment. The proposed algorithm constructs the scalable local algorithm in P2P environment so that it each and every time the new constraints in the decision trees are updated frequently.

Previously the algorithm proposed was based on the multivariate regression. But the algorithm proposed here is for DDT without the multivariate regression as it tends to monitor the models closely which may degrade the performance.

```
Input : D→database, n-nodes
Initialization:
Create a roof leaf and let set D← {n1,n2,…nn}
if root is the designated criteria
   Set nodes n={root}
Else
   Push the {root} to a queue
Send message to all communicating nodes
```

Fig 1 for initializing roots in all nodes

Fig1 describes the initialization of all the nodes and communicating the messages. Suppose there is any delay in the message passing.Following fig 2 suggests solution for that

```
If the {root} > message {γ}
Send message to D←{n1,n2,..nn} with delay of τ
   Add the D into the queue for the further processing
Else
  Limit only the attributes criteria’s satisfied
```

Fig 2 for the delay communication

{γ} specifies the time that is computed[6] in the distributed environment. Both the above fig1 and fig 2 specifies the sending the message from one node designated as root nodes to all the branch nodes. The fig 3 describes the Branch node receiving and processing further nodes.

```
Input: messages from other nodes
On Branch(τ,n1,n2,..nn)
Send Branch messages { γ1, γ2,.., γn} with delay τ .
Pop from the queue into γ
If  passive
       Enqueue γ
If not passive
       Call Branch
       Add τ
Dequeue τ, γ
```

Fig 3 for the branch node

Fig 3 proposes the branch node algorithm which is called recursively and partitioned for vertical data i.e heterogeneous data. Each and every time the recursive call will send the messages with the specific criteria is met.

All the above proposed algorithm discusses the message passing in the nodes. For classification in P2P environment, another new algorithm is proposed which minimises the cost and reduce the misclassification errors that may occur.

B. P2P Algorithm

This algorithm takes input as peer , passes to it neighbors and set of data

Input: set of variables k, neighbors N_k
Output : criterion attribute A*
For every A* initialize the inputs $X_1,X_2,..X_i$.
Denote this instances $Q_i,..Q_k$ where the change in the instances are $Q_0^i\Delta_k$ and $Q_n^i\Delta_{k_n}$
Further for all N_k and Q_k the instances are located and agreed upon.

Fig 4 Algorithm for P2P

Fig 4 describes instances and how the exchange of information passed through nodes with the instances.

- For $A^i \in \{A^i,..A^n\}$
- if not $Q_0^i\Delta_k < \theta$ and $Q_n^i\Delta_{k_n} > n$ call Branch($Q_n^i\Delta_{k_n}$)
- if not $Q_n^i\Delta_{k_n} > \lambda$ and $Q_0^i\Delta_k < \theta$ call Branch

Fig 5 Algorithm for A* selection

Fig 5 specifies the attributes selection of the pivot criteria met. When these attributes are selected the nodes of the exchange transmission can be done without further mitigated delay. Transmission of the messages are enqueued and dequeued for the heterogeneous data sources further all types and format can be taken into account.

Input: $Q_0^i\Delta_k$, A^i, γ_i,τ and D.
Output: $Q_n^i\Delta_{k_n}$ if $i < \theta$, 1 otherwise
Initialization:
Initialize nodes of root and Branch
if MessageRecvdFrom {root} with $n_1,n_2,..n_n$ then
Q_j ,D, $|X_i,\dots X_n|$
$Q_{j,i} \leftarrow Q$;
$|Q_{j,i}| \leftarrow |X_n|$;
Update branch and nodes;
end
if γ_i, τ_i *or* A^i *changes* **then**
forall Qj $\in X_i,X_{i1},......X_n$ **do**
if LastMsgSent > N_n *time units ago* **then**
if D = γ **then**
$X_{i,j} \leftarrow |Q_{\Delta i}| Q_{\Delta k} - |Q_{j,i}| X_{j,i}$
$|Q_{\Delta i}| - |Xj,i|$;
$|X_{i,j}| \leftarrow |Q_{\Delta K}| - |X_{j,i}|$;
end
if $A_{i,j} = \theta$ *or* $X_{i,j} > Q_\Delta$ **then**
Set $X_{i,j}$ and $|X_{i,j}|$ such that $A^*_{i,j}$ and $Q_{\Delta\ i,j}$ ε R;
end
SendMessage
$Q_{\Delta k}$,$X_{i,j}$, $|X_{i,j}|$,D
LastMsgSent $\leftarrow Q_\Delta$;
Update all nodes;
end
else Wait till nodes with A* units are communicated and then check again;
end

Fig 6 Algorithm for p2p with scalable handling

fig 6 describes the overall algorithm in terms of the message exchanging and communicating in the disseminated environment in consider to delays and minimisation cost.

IV. COMMUNICATIONAL COMPLEXITY

The communication complexity of computing a predictive modelling of the proposed algorithm is depends on the degree of the number of Message passed or exchanged in the data points i.e. |D|.

The task of computing can be reduced to computing the certain attributes A* with the pivot value consideration. The dimensionality of $D_1,D_2,..D_n$ can be communicated by only with the exchanging of the data points and the nodes passing of $n_1,n_2,..n_n$.Therefore the total communication complexity is $O(\log n^2)$, which is independent of the size of the dataset |D|. The efficiency of the converge cast process is due to the fact that $n \leq |D|$. Hence there can be significant savings in terms of communication cost based on the criteria's met and attributes selection in the P2P.

V. EXPERIMENTAL EVALUTION

The communication complexity of the model can quite reduce the cost and also minimises misclassification in the distributed decision trees constructions. Further the delay of the messages also considered into the communication cost.

CONCLUSION

This paper proposed the efficient, scalable algorithm in local P2P environment with the essence in reducing the computation and communication cost. Based on the delay of the transmission of the messages further data exchange can be decided. Thus a scalable and robust algorithm is constructed for DDT in P2P environment.

REFERENCES

1. S. Datta, K. Bhaduri, C. Giannella, R. Wolff, and H. Kargupta, "Distributed Data Mining in Peer-to-Peer Networks," IEEE Internet Computing Special Issue on Distributed Data Mining, vol. 10, no. 4, pp. 18–26, 2006.
2. Kanishka Bhaduri, Ran Wolff, Chris Giannella,Hillol Kargupta, "Distributed Decision Tree Induction in Peer-to-Peer Systems", Journal Statistical Analysis and Data Mining, vol 1,issue 2,2008.
3. Kolweyh, " Towards Next-Generation Peer-to-Peer Systems" ,University of Bremen.
4. Hillol Kargupta, "Distributed Data Mining in Peer-to-Peer Networks: Local Algorithms, Privacy Issues, and Games". University of Maryland, Baltimore County and AGNIK.
5. Rüdiger Schollmeier, München, "A Definition of *Peer-to-Peer* Networking for the Classification of *Peer-to-Peer* Architectures and Applications".
6. D. Caragea, A. Silvescu, and V. Honavar, "A Framework for Learning from Distributed Data Using Sufficient Statistics and Its Application to Learning Decision Trees," *International Journal of Hybrid Intelligent Systems*, vol. 1, no. 1-2, pp.80–89, 2004.
7.C. Giannella, K. Liu, T. Olsen, and H. Kargupta, "Communication Efficient Construction of Deicision Trees Over Heterogeneously Distributed Data," in Proceedings of ICDM'04, Brighton, UK, 2004, pp. 67–74.
8. S. Merugu and J. Ghosh, "A Distributed Learning Framework for Heterogeneous Data Sources," in Proceedings of KDD'05, 2005, pp. 208–217.
9.D. Krivitski, A. Schuster, and R. Wolff, "A Local Facility Location Algorithm for Large-Scale Distributed Systems," Journal of Grid Computing, vol. 5, no. 4, pp. 361–378, 2007.
10. Doina Caragea, "Learning classifiers from distributed, semantically heterogeneous, autonomous data sources",Doctral Thesis.
11.H. Kargupta and K. Sivakumar, Existential Pleasures of Distributed Data Mining. Data Mining: Next Generation Challenges and Future Directions. AAAI/MIT press, 2004.
12. Kanishka Bhaduri_, Hillol Kargupta , " A Scalable Local Algorithm for Distributed Multivariate Regression", SIAM Data Mining Conference 2008.

AUTHORS PROFILE

Dr.E.Chandra received her B.Sc., from Bharathiar University, Coimbatore in 1992 and received M.Sc., from Avinashilingam University ,Coimbatore in 1994. She obtained her M.Phil., in the area of Neural Networks from Bharathiar University, in 1999. She obtained her PhD degree in the area of Speech recognition system from Alagappa University Karikudi in 2007. She has totally 15 yrs of experience in teaching including 6 months in the industry. Presently she is working as Director, Department of Computer Applications in D. J. Academy for Managerial Excellence, Coimbatore. She has published more than 30 research papers in National, International Journals and Conferences in India and abroad. She has guided more than 20 M.Phil., Research Scholars. Currently 3 M.Phil Scholars and 8 Ph.D Scholars are working under her guidance. She has delivered lectures to various Colleges. She is a Board of studies member of various Institutions. Her research interest lies in the area of Data Mining, Artificial Intelligence, Neural Networks, Speech Recognition Systems, Fuzzy Logic and Machine Learning Techniques. She is an active and Life member of CSI, Society of Statistics and Computer Applications. Currently she is Management Committee member of CSI Coimbatore Chapter.

P.Ajitha,, received her B.Com from Bharathiar University, Coimbatore in 1998 and received MCA from Bharathiadsan University, Trichy in 2001. She obtained her M.Phil in the area of Data Mining in 2004. She has 9 years of experience in teaching and 3 months of industrial experience. Currently, she is working as Assistant Professor, Department Of Computer Applications, D.J.Academy for Managerial Excellence, Coimbatore and pursuing her Ph.D in Bharathiar University,Coimbatore. She has presented more than 10 research papers in National and International Conferences and published a paper in an International Journal. Her Research Interest lies in Distributed Data Mining, Machine Learning and Artificial Intelligence. She is Life a member of CSI, life member of Institute of Advanced Scientific Research and also a member IASCIT.