ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

ACO Swarm Search Feature Selection for Data stream Mining in Big Data

Shivani Harde, Vaishali Sahare

Scholar Student, Dept. of CSE ,G. H. Raisoni Institute of Engineering and Technology for Womens, Nagpur , India

Scholar Student, Dept. of CSE ,G. H. Raisoni Institute of Engineering and Technology for Womens,Nagpur , India

ABSTRACT: Big data is a term for large datasets use for analysis to make beneficial decision and strategic move. But it has many technical challenges that also confront by both academic research communities and commercial IT deployment. Data streams and the curse of dimensionality are founded to be the root sources of Big Data. The commonly used procedure for data sourced from data streams is continuously making batch based model and inducing algorithms which is infeasible for real-time data mining. An optimal feature subset which is derived by mining over high dimensional data search space grows exponentially in size which leads to an intractable demand in computation. In order to solve this problem which is based on high dimensionality and streaming format in data feeds in big data , light weight feature selection is indicated which will particularly concentrate on mining data on fly , by using ant colony optimization (ACO) type of swarm search which can achieve enhanced analytical accuracy.

KEYWORDS: Feature selection , Big Data , Swarm intelligence, Classification , Ant Colony Optimization

I. INTRODUCTION

Recently Big Data is lot of attention but at the same time it faces three problematic issues that are: Velocity problem where data generated is larger and grows continuously ; Variety problem relates to captured data which is of different type and may be structured , unstructured or semi-structured that makes data preprocessing and integration difficult ; Volume problem relates to data which continuously arrives in streams of data where to create a real value rate of change must be done quickly. In existence of such these three challenges traditional data mining approach will not fulfill the demand of analytic efficiency because in traditional data mining approach which require full data set and every time when new data arrive the traditional induction method need to be re-run. In oppose, the new algorithm known as data stream mining method. The algorithm is able to induce classification model from bottom up approach and incrementally update itself without reloading previous data. in both types of data stream mining algorithms feature selection attempts to select the subpart of most frequent features excluding irrelevant and repeated features in order to improve accuracy.

However recently reported proposed methods are limited to the following restrictions in their designs: (1) fixed size of the result and feature set is assumed. due to which user may not know the upper limit of subset. (2) due to the principle of removing redundancy the feature set become minimal. (3) the feature selection methods are designed for some specific classifier and optimizer. Although an extreme and exhaustive computing used for finding the accurate feature subset , this is quite difficult for data streams which are usually in high dimensionality and large amount.

This paper proposed the idea of new light weight feature selection called swarm search with ant colony optimization with the goal of finding correct combination of classification algorithm and the feature select algorithm on the fly accurately.

II. RELATED WORK

A. Traditional and incremental model learning methods

In traditional classification, top-down supervised learning is followed where to construct classification model a full dataset is used by recursively partitioning the data to form mapping relations. These models are develop based on stationary datasets. This model needs to be update each time the new sample arrives. The traditional model have a good performance on data which is stationary without updating new changes. But in dynamic streaming processing data streams would have to be frequently updated. Therefore a new algorithm known as incremental classification algorithm has been proposed [4].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

B. Incremental learning algorithms

There are two main types of algorithms where design for incremental learning: functional based and incremental based learning. Kstar and updatable naïve bayesare two most popular functional based incremental learning algorithm.

Kstar is "instance based learner using an entropic distance measure" which learns per instance incrementally by using similarity function which measures the entropic distance between test instance and other instance. But because of the large amount of summation over all possible paths ,Kstar require longer processing time is been proposed [3].

Updatable naïve bayes algorithm is designed on the assumption that there is strong independence between features. This assumption is beneficial in such a way that it require small amount of data to estimate the similarity and dissimilarity of the features.

HOT is the algorithm based on decision tree. It produce some optional tree branches and rules with low accuracy are replace by optional one at the same time because of the construction of the optional tree branches learning speed get slowed.

C. Feature selection by swarm search

A algorithm which is one of the contemporary type of feature selection algorithm designed specially for choosing optimal subset from a huge search space is called swarm search feature selection (SS-FS) model. In this algorithm Initially random selection of feature subset is done from which iteration starts and continues further to improve the accuracy of classification model is been proposed [2].FS-APSO [8] proposed an algorithm that search the space by adjusting individual agents known as particles. In PSO swarm search there are two major components : a stochastic component and a deterministic component , among which stochastic based search strategy is used where instead of testing on individual feature subset , the multiple search agents work parallel. But this type of parallel testing create confusion in selection of feature subset.

III. ANT COLONY OPTIMIZATION

Ant colony optimization is totally based on the behavior of real ants. As individual ants are not capable to solve complex problems. In contrast , the collective or group of ants are capable to solve complex tasks. Some basic idea proposed in [6] followed by ACO strategy are:

- Population of ants performs search
- Solution construction is incremental
- Stigmatized information is important for probabilistic solution component.
- There is indirect communication between the ants based on the pheromone deposited on the path.

A. Why Ant Colony Optimization

The reason for using ACO algorithm is that simple agents (ants) cooperate with one another to achieve an rising , collective behavior for the system as whole , producing a system capable of finding high-quality solution with a large search space and incrementally modifies the solution for the target problem.

B. Goals of Proposed Algorithm:

- To solve the problem of High dimensionality and streaming format of data- It is generally known that big data manifested 3V challenges velocity , variety and volume. Develop a data stream mining method which will potentially handle the 3V challenges.
- To improve analytical accuracy of pre-processing time- To tackle the problem related to traditional data stream mining of re-running and constructing model by inducing fresh data consumes time. Present incremental methods are with average accuracy. Proposed method will improve the present accuracy of data stream mining on the fly.
- To design and implement incremental method on Dynamic data- The proposed algorithm will be capable of inducing a classification. Each pass of data streams triggers the model incrementally update itself without the need of reloading any previously seen data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

IV. FEATURE SELECTION BY ANT COLONY OPTIMIZATION

In the proposed method a new light weight feature selection is proposed which is designed particularly for the data stream on the fly by using Ant Colony Optimization (ACO). This type of swarm search that will achieve improved analytical accuracy in minimized processing time. New feature selection algorithm will test the big data from e-commerce with high degree of dimensionality and streaming format.

Operation of swarm search feature selection is as follows:

1. Start with a random selection of feature subset .
2. Search better feature subset to refine the accuracy of classification model
3. To find out usefulness of candidate feature subset wrapped classifier will be used as fitness evaluator.
4. Searching for candidate feature subset by optimization function will be done in stochastic manner as we need perform data stream mining in big data.
5. Here stochastic based swarm search will perform in parallel manner to find the optimal feature subset.
6. To shorten the search process we will speed-up in the initialization step by using swarm search ACO.

V. CONCLUSION

In big data analytics , there are some computational challenges in data mining due to high dimensionality and the streaming behavior of incoming data. As big data generate fresh data all the time it requires incremental computational approach that will be able to handle large scale of data dynamically. In this paper we proposed the novel light weight feature selection method by using swarm search and ant colony optimization which will be useful for data stream mining.

REFERENCES

1. Wei Fan, Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2, pp.1-5
2. S. Fong, X.S. Yang, S. Deb, Swarm Search for Feature Selection in Classi-fication, The 2nd International Conference on Big Data Science and En-gineering (BDSE 2013), 2013, 3-5 Dec. 2013
3. John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, pp.108-114, 1995
4. Aggarwal, Charu C., ed. Data streams: models and algorithms. Vol. 31.Springer, 2007.
5. Simon Fong, Suash Deb, Xin-She Yang, Jinyan Li, "Metaheuristic Swarm Search for Feature Selection in Life Science Classification", IEEE IT Professional Magazine, August 2014, Volume 16, Issue 4, pp.24-29.
6. Rafael S. Parpinelli, Heitor S. Lopes, and Alex A. Freitas "Data Mining with an Ant Colony Optimization Algorithm" CEFET-PR, CPGEI, Av. Sete de Setembro, 3165, Curitiba - PR, 80230-901, Brazil
7. Mallios, N, Papageorgiou, E., Vassilakopoulos, M. "Ant Colony Optimization and Data Mining: Techniques and Trends" , IEEE P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC) international conference (2010).
8. Fong, S.; Wong, R. ;Vasilakos, A. "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data" Services Computing, IEEE Transactions on (Volume:PP , Issue: 99) (2015):1939-1374.
9. SampadaLovalekar "Big Data: An Emerging Trend In Future" , (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 538-541

BIOGRAPHY

Vaishali N. Sahareis Assistant Professor of CSE Department , G. H Raisoni institute of Engineering and Technology for women , Nagpur , India. She received Master of technology inComputer science and engineering degree from Nagpur university. Her research interests are Computer Networks (wireless Networks), data mining etc.