International Research Journal of Computer Science (IRJCS) ISSN: 2393-9842
Issue 04, Volume 3 (April 2016) www.irjcs.com

FEATURE CLUSTERING USING SUBSELECTION ALGORITHM IN BIG DATA USING FIDOOP

C.Surekha*
Department of CSE,
Kingston Engineering College,
Vellore,Tamil Nadu

S.Vijayalakshmi*
Department of CSE,
Kingston engineering College,
Vellore,Tamil Nadu

Natteshan N.V.S
Assistant professor, CSE,
Kingston Engineering College,
Vellore,Tamil Nadu

Abstract— Big data processing is a high demand area which imposes a heavy burden on computation, communication, storage in data centers, which incurs considerable operational cost to data center provider. So minimizing cost has become an issue for the upcoming big data. Different from conventional cloud service one of the main feature of the big data service is the tight coupling between data and computation, as computation task can be conducted only when the corresponding data are available. As a result, three factors that is communicational cost, computational cost, operational cost effects the expenditure cost of data centers. So in order to minimize the cost clustering is used. Clustering groups a selected objects into classes of similar objects. Feature Selection Removes Irrelevant Features- it occurs in the batch processing (scheduling algorithm) Redundant Features its occurs in the cluster formation (data-centric algorithm) joint-optimization– 2 steps Features divided into clusters(subsets) MST Cluster representatives are selected Efficient, Effective, Independent. Based on these criteria, a feature clustering based on selection algorithm is proposed and experimentally evaluated for a sample cancer dataset. This work finds the effective attributes used and removes redundancy.

Key words-Big data, clustering, Feature selection, Scheduling, Subset selection.

I. INTRODUCTION

Big data usually includes data set with sizes beyond the ability of commonly used software tools to capture, manage and process data within a tolerable elapsed time. Its size is constantly moving target as of 2012 ranging from a few Dozen of terabyte to many petabytes of data "massively parallel software running on tens, hundreds, or even thousands of servers".
Advantages:

- *Big data is timely, accessible, trustworthy, relevant and secure.*
- *Reduced maintenance cost*
- *Big data tools allow us to identify the threats that we face internally.*
- *It keeps data safe.*
- *Helps in keeping relevant datas.*

A. OBJECTIVE OF THE WORK

The objective of the project is subset selection to achieve fast retrieval. Compressed storage and load balancing is achieved by calculating T-relevance. We are going to compute F-correlation and construct a Minimum Spanning Tree to improve speed and accuracy. The cost reduced by applying Kruskal's algorithm. Efficiency and effectiveness of fast clustering algorithm are evaluated. ARFF (Attribute Relation File Format) is used to create attribute table for the CSV (Comma Separated Values) datas which are given as input. The table is divided into sub tables based on last attribute. Entropy considers all data values. Conditional entropy considers data values based on conditions and gain is the difference between entropy and conditional entropy. It is calculated based on standard deviation to remove redundancy. MST is constructed using kruskal's algorithm and the cost is reduced based on the shortest path. The similar datas are clustered using clustering algorithm. The final data set is stored in HDFS (Hadoop Distributed File System) using map reduce technique. The performance after redundancy removal is evaluated.

The overall organization of this paper is as follows, Section II describes the related work, section III gives a description of the system, Section IV provides a insight of the experimental evaluation and result and section V concludes the work.

II. RELATED WORK

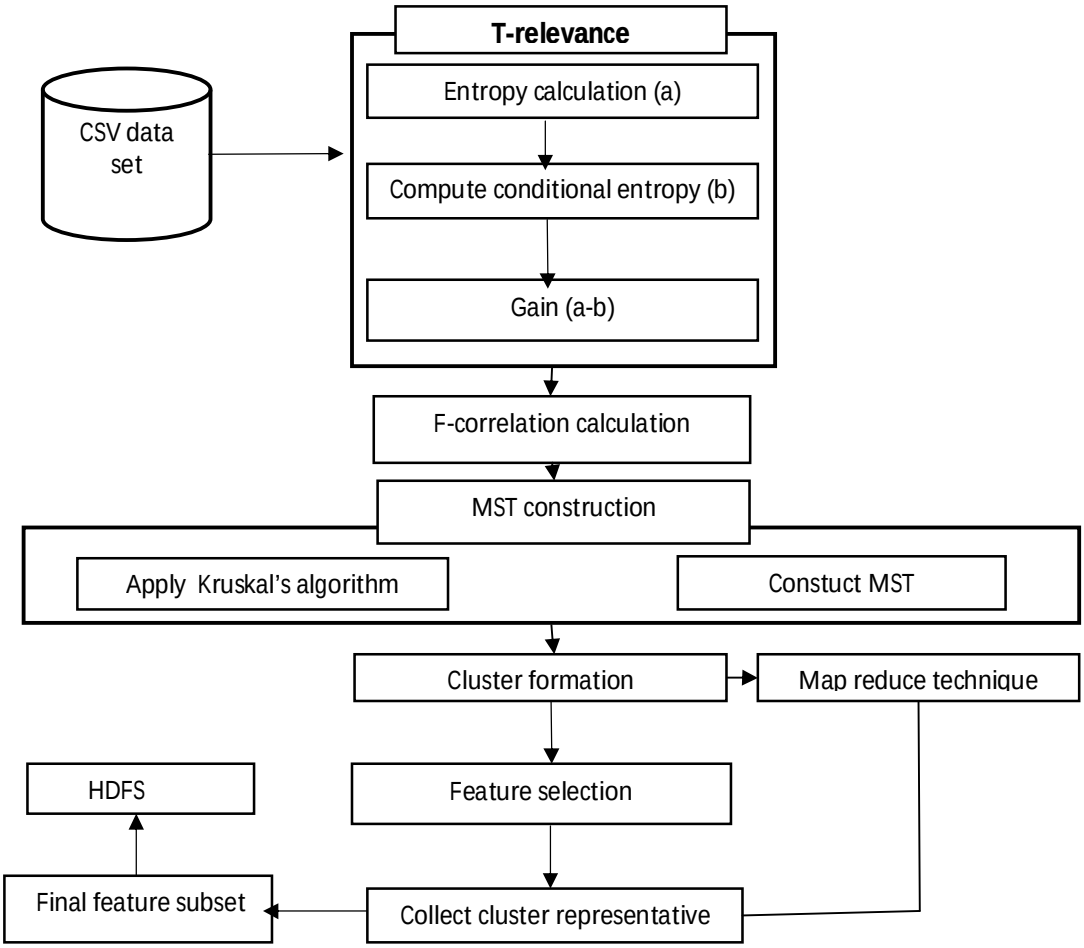
This section deals with the related work performed in evaluating the A FEATURES CLUSTERING BASED ON SUBSELECTION.

Yuh-Jiuan Tsay(2009) describes an efficient method, the frequent item ultrametric tree(FIUT) for mining frequent item sets in a database. FIUT uses a special frequent item ultrametric tree (FIUT) structure to enhance its efficiency in obtaining frequent itemset.

Jan Neerbek (2012) describes that frequent itemset mining finds frequently occurring itemsets in transactional data, which is applied to diverse problems such as decision support, selective marketing, financial forecast and medical diagnosis. They also propose that the cloud computation has a utility service, allow us to crunch large mining problems.

Yanyan Shen(2013) describes that K nearest neighbor join(KNN join) designed to find K nearest neighbors from a data set S for every object in another dataset R, is a primitive operation widely adopted by many data mining applications. As a combination of the K nearest neighbor query and the joint operation, KNN join is an expensive operation.In this paper they investigated how to perform KNN join using map reduce which is a well accepted framework for data intension applications over clusters of computers. In brief, the mappers cluster objects into groups. The reducers perform the KNN join on each group of object separately. They designed an effective mapping mechanism that exploits pruning rules for distance filtering and hence reduces both the shuffling and computational costs.

III. A CLUSTERING OF FEATURES BASED ON SUBSELECTION ALGORITHM IN BIG DATA USING FIDOOP



This section gives the detail about the system design and the steps used.

A. Steps performed:

1. *Load Data and Classify*
2. *Information Gain Computation*
3. *T-Relevance Calculation*
4. *F-Correlation Calculation*
5. *MST Construction*
6. *Cluster Formation using Big data*
7. *Data Stored in HDFS*

The below figure describes the overall architecture diagram where the users provide their data set and from their input a ARFF is applied and the attributes are identified then a T-relevance and F-correlation will be computed and then the minimum cost is computed by using the MST construction and then the effective attributes are identified.

IV. EXPERIMENTAL EVALUATION

Experiments were conducted by using some cancer datasets tabulated below and the number of effective attributes and in effective attributes is found out and is graphically represented

.

TABLE I
DATA SET USED

DATA SET ID	DATASET NAME	NO OF ATTRIBUTES
D_001	BRAIN TUMOR	16
D_002	LUNG CANCER	16
D_003	BREAST CANCER	14
D_004	ABDOMEN CANCER	12
D_005	SKIN CANCER	10

Table I describes the data sets used in the experiments and the number of attributes used in that dataset.

TABLE –II
INFORMATION GAIN, T-RELEVANCE COMPUTATION FOR A SAMPLE DATASET

ATTRIBUTES	INFORMATION GAIN	T-RELEVANCE
A1	57	2
A2	60.544	2
A3	300.542	2
A4	29.1	2
A5	59.25	2
A6	70.465	2
A7	81.27	2
A8	121.25	2
A9	135.24	1
A10	255.46	1
A11	39.05	1
A12	49.10	1
A13	59.11	1
A14	47.02	1
A15	33.39	1

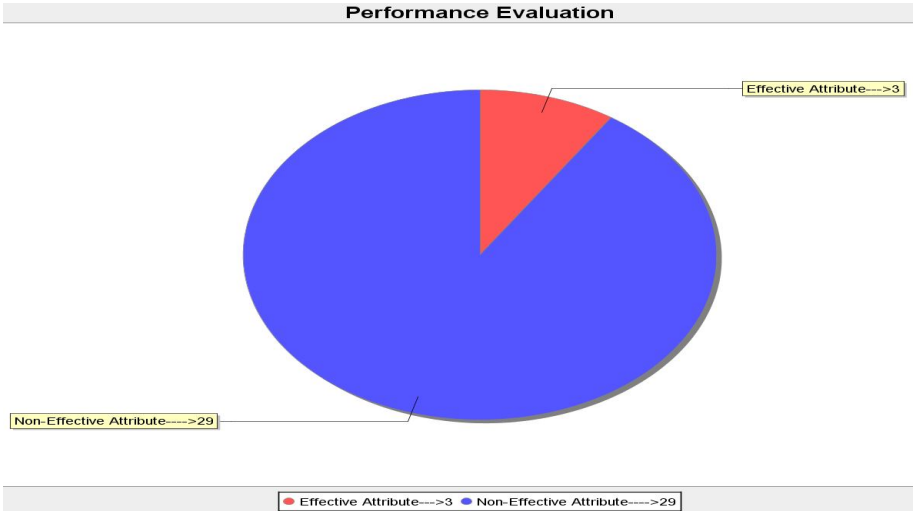


Fig. 2 Effective attribute computation

The above pie graph representation describes about the number of effective attributes computed in this process and the measure of ineffective attributes computed in the research work.

V. CONCLUSION AND FUTURE WORK

Thus the a clustering feature based on subselection algorithm in big data using fidoop is done and it effectively removes redundant data in tables and the effective attributes are computed by using the Minimal spanning tree. Removes both irrelevant and redundant attributes, Cost is reduced Fast retrieval of data from database No data loss and corruption, Uses minimum spanning tree concept for fast elimination of unnecessary edge. The method of performing the elimination of unnecessary or costly edge can be done using some other algorithm like dijkstras algorithm.

REFERENCES

[1]. Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:25 NO:1 YEAR 2013.

[2]. Yaling Xun, Jifu Zhang, and Xiao Qin, "FiDoop: Parallel Mining of Frequent Itemsets Using MapReduce", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, Issue 3, pp. 313-325, March 2016.

[3]. R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, 1993.

[4]. R. Agrawal and J. C. Shafer, "Parallel mining of association rules," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 6, pp. 962–969, Dec. 1996.

[5]. Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.

[6]. Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537- 550, 1994

[7]. Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Selection through Clustering, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005

[8]. Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009.

[9]. Guyon I. and Elisseeff A., An introduction to variable and feature selection, Journal of Machine Learning Research, 3, pp 1157-1182, 2003.

[10]. Arauzo-Azofra.A, Benitez.J.M, and Castro J.L, (2004) "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109.

[11]. Almuallim.H and Dietterich.T.G,(1992)"Algorithms for Identifying Relevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45.

[12]. Almuallim.H and Dietterich.T.G,(1994) "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305.