© 2018 IJSRST | Volume 4 | Issue 5 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Big Data Clustering Using Heuristic Data Intensive Computing and Self Organizing Maps

K Nagamani, K Sunitha

[1]Master of Science (CS), Department of Computer Science, RIIMS College, SV University, Tirupati, Andhra Pradesh, India

[2]Associate Professor, Department of Computer Science, RIIMS College, SV University, Tirupati, Andhra Pradesh, India

ABSTRACT

Traditional data clustering algorithms are having pitfalls while discovering efficient clusters. As the data base size increases dynamically and the dramatic changes in the use of data, will shows adequate results on clustering performance. Transforming the massive amounts of data into knowledge will leverage the organization performance to the maximum. Scientific and business organization would benefit from utilizing big data. However, there are many challenges in dealing with big data such as storage, transfer, management and manipulation of big data. Many techniques are required to explore the hidden and transitive patterns inside the big data which have limitations in terms of hardware and software implementation. Through this, a unified framework is presented for big data clustering using a Heuristic data intensive computing (HDIC) and Self-Organizing Maps (SOM). It is implemented on an N-node HDIC clusters, driven by a wide range of data sets created using IBM synthetic data generator and real time data sets taken from UCI. This is significantly implemented to improve the performance of the big data clustering on the existing approaches.

Keywords: HDIC, Self organizing maps, big data, clustering and IBM database generator.

I. INTRODUCTION

Big data is a term that describes the large volume of data both structured and unstructured that inundates a business on a day-to-day basis .But it's not the amount of data that's important .It's what organizations do with the data that matters .Big data can be analyzed for insights that lead to better decisions and strategic business moves .While the term "big data" is relatively new, the act of gathering and storing large amounts of information for eventual analysis is ages old. The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three V's:

Volume: Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden.

Velocity: Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.

Variety: Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

At SAS, consider two additional dimensions when it comes to big data:

Variability: In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data.

Complexity: Today's data comes from multiple sources, which makes it difficult to link, match, cleanse and transform data across systems. However, it's necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

CLUSTERING

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

II. PROBLEM DEFINITION

Let P be a set of points in the plane. A partitioning of P into k disjoint (possibly empty) sets $C_1, C_2, \ldots, C_k$ is called a clustering, and the individual sets C_i are called its clusters. In cluster analysis, the points represent properties (data) of real-world objects, and the aim is usually to collect "similar" objects (points which are close to each other) in the same cluster, and to put objects which are very "different" into different clusters. The definition of "similarity" of objects is crucial for every clustering process. In a general setup, let W be some weight function that assigns a real weight to any set of finite point sets $C_1, C_2, \ldots, C_k$ in the plane (Examples for W are the maximum diameter of all C_i or the sum of the circumferences of the convex hulls of all C_i or the distances between all pairs of points in the same point set). Intuitively, W is a measure of the quality of the clustering $C_1, C_2, \ldots, C_k$. Then the planar clustering problem for W is defined as follows.

INSTANCE: A set P of m points in the plane; integers k, nl and nu; a rational number d. **QUESTION:** Is there a clustering for P into k sets $C_1, C_2, \ldots, C_k$ such that $nl \le |C_i| \le nu$ and such that $W(C_1, C_2, \ldots, C_k) \le d$ holds?

Clearly, this problem could be defined in higher dimensions, but confine our interest to the plane. Usually, not all of the numbers k, nl and nu are specified; sometimes they are specified but not as part of the input. Sometimes there are additional restrictions on the clusters (e.g. the convex hulls are required to be pair wise disjoints).

2.1 EXISTING SYSTEM

Data size has increased dynamically with the advent of today's technology in many sectors such as Manufacturing, Business, and Science and Web applications. Most of the data are structured, and some data are semi-structured while others are unstructured and mix with different types of data such as documents, records, images and videos. Resources of data are from Web applications, which produce a very big volume of data.

Limitations in Existing System

Unsupervised algorithm lack of mechanism that enables automatic data distribution, load balancing and fault tolerance on large computing clusters.

The main disadvantage of this algorithm is every time a new data is arrived, it reconstructs the clusters.

Takes more time and memory space in existing methods, the clusters will contain a less number of features. Difficulty to get accurate clusters due to the data base scalability.

2.2 PROPOSED SYSTEM

The proposed framework for big data manipulation is shown in the Figure. It uses Heuristic Data Intensive Computing (HDIC) for data processing and Self-

Organizing Maps (SOM) algorithm is used for data clustering. The proposed framework starts with various data sources merged into a master database.

Heuristic:

A heuristic technique often called simply a heuristic, is any approach to problem solving, learning, or discovery that employs a practical method not guaranteed to be optimal or perfect, but sufficient for the immediate goals. Where finding an optimal solution is impossible or impractical, heuristic methods can be used to speed up the process of finding a satisfactory solution. Heuristics can be mental shortcuts that ease the cognitive load of making a decision. Examples of this method include using a rule of thumb, an educated guess, an intuitive judgment, stereotyping, profiling, or common sense.

This heuristic is used to convert one form of data to another form of data i.e., for example K-Means algorithm support only for discrete data. If input is numerical it doesn't work, so in that time heuristic can convert numerical data to discrete data. Data Intensive Computing is used to do parallel processing of large data.

Self-Organizing Map (SOM):

The Self-Organizing Map was been proven useful in many applications One of the most popular neural network models. It belongs to the category of competitive learning networks. Based on unsupervised learning, which means that no human intervention is needed during the learning and that little need to be known about the characteristics of the input data.

Use the SOM for clustering data without knowing the class memberships of the input data. The SOM can be used to detect features inherent to the problem and thus has also been called SOFM, the Self-Organizing Feature Map. It provides a topology preserving mapping from the high dimensional space to map units. Map units, or neurons, usually form a two dimensional lattice and thus the mapping is a mapping from high dimensional space onto a plane.

III. SYSTEM ANALYSIS: MODULES

3.1 IBM Database Generator

IBM Database Generator is used to generate test data from scratch or from existing data. It will have the parameters like average transaction size, length, correlation, number of attributes etc. Test data can be generated in a variety of formats, including SQL or XML.

3.2 Heuristic Data Intensive Computing (HDIC)

Heuristic means set of rules that are used to increase the probability of solving a problem. Heuristic is used to give optimal solution. It can provide a feasibility to convert one form of data into another form of data. Data Intensive Computing is a class of parallel computing applications which use a data parallel approach to process large volumes of data typically terabytes or peta bytes in size and typically referred to as a Big Data. This Heuristic data intensive computing is applied for uniform data. Here unwanted data or duplicate data is removed by using this technique.

3.3 Clustering using Self Organizing Maps (SOM)

The Self-Organizing Map (SOM) is one most popular neural network model. It is a unsupervised learning which means that no human interaction is needed during the learning.SOM is used for clustering data without knowing the class memberships of input data. The SOM can be used to detect features inherent to the problem and thus has also been called SOFM, the Self-Organizing Feature Map.

It provides a topology preserving mapping from the high dimensional space to map units. Map units, or neurons, usually form a two-dimensional lattice and thus the mapping is a mapping from high dimensional space onto a plane. The property of topology

preserving means that the mapping preserves the relative distance between the points. Points that are near each other in the input space are mapped to nearby map units in the SOM.

The SOM can thus serve as a cluster analyzing tool of high-dimensional data. Also, the SOM has the capability to generalize. Generalization capability means that the network can recognize or characterize inputs it has never encountered before. A new input is assimilated with the map unit it is mapped to.

3.4 Cluster Visualization

Cluster Visualization renders your cluster data as an interactive map allowing you to see a quick overview of your cluster sets and quickly drill into each cluster set to view sub clusters and conceptually-related clusters. If there is any possibility to do sub cluster, then sub clustering is done.

3.5 Building Knowledge Base

A knowledge base in Data Quality Services (DQS) is a repository of knowledge about your data that enables you to understand your data and maintain its integrity. A knowledge base consists of domains, each of which represents the data in a data field.

IV. SYSTEM IMPLEMENTATION

4.1 Orange tool:

Orange library is a hierarchically-organized toolbox of data mining components. The low-level procedures at the bottom of the hierarchy, like data filtering, probability assessment and feature scoring, are assembled into higher-level algorithms, such as classification tree learning. This allows developers to easily add new functionality at any level and fuse it with the existing code. The main branches of the component hierarchy are: data management and pre-processing for data input and output, data filtering and sampling, imputation, feature manipulation (discretization, continuization, normalization, scaling and scoring), and feature selection.

Classification with implementations of various supervised machine learning algorithms (trees, forests, instance-based and Bayesian approaches, rule induction), borrowing from some well-known external libraries such as LIBSVM (Chang and Lin, 2011).

Clustering using Self Organizing Maps (SOM)

The Self-Organizing Map (SOM) is one most popular neural network model. It is a unsupervised learning which means that no human interaction is needed during the learning.SOM is used for clustering data without knowing the class memberships of input data. The SOM can be used to detect features inherent to the problem and thus has also been called SOFM, the Self-Organizing Feature Map.

Algorithm for SOM:

- ✓ Step 1: Initialize the weights from M inputs to the N output units to small random values.
- ✓ Step 2: Present a new input **a.**
- ✓ Step 3: Compute the distance *di* between the input and the weight.
- ✓ Step 4: Select the output unit k with minimum distance.
- ✓ Step 5: Update weight to node k and its neighbors.
- ✓ Step 6: Repeat Steps 2 through 5 for all inputs several times.

4.2 R-Programming:

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R is freely available under the GNU General Public License, and pre-compiled binary versions are

provided for various operating systems like Linux, Windows and Mac. R is free software distributed under a GNU-style copy left, and an official part of the GNU project called GNU S.

Design of the R system:

The primary R system is available from the Comprehensive R Archive Network[15], also known as CRAN. CRAN also hosts many add-on packages that can be used to extend the functionality of R.

The R system is divided into 2 conceptual parts:

1. The "base" R system that you download from CRAN: Linux Windows Mac Source Code

2. Everything else.

R functionality is divided into a number of packages.

• The "base" R system contains, among other things, the base package which is required to run R and contains the most fundamental functions.

• The other packages contained in the "base" system include utils, stats, datasets, graphics, gr Devices, grid, methods, tools, parallel, compiler, splines, tcltk, stats4. There are also "Recommended" packages: boot, class, cluster, code tools, foreign, KernSmooth, lattice, mgcv, nlme, rpart, survival, MASS, spatial, nnet, Matrix.

When you download a fresh installation of R from CRAN, you get all of the above, which represents a substantial amount of functionality. However, there are many other packages available:

• There are over 4000 packages on CRAN that have been developed by users and programmers around the world.

• There are also many packages associated with the Bioconductor project.

• People often make packages available on their personal websites; there is no reliable way to keep track of how many packages are available in this fashion.

• There are a number of packages being developed on repositories like GitHub and BitBucket but there is no reliable listing of all these packages.

R programming packages:

R will download the package from CRAN, so you'll need to be connected to the internet. Once you have a package installed, you can make its contents available to use in your current R session by running. There are thousands of helpful R packages for you to use, but navigating them all can be a challenge. They used each of these, and found them to be outstanding – they even written some of them. But you don't have to take our word for it, these packages are also some of the top most downloaded R packages.

Advantages of R:

- ✓ R is free and open source software, allowing anyone to use and, importantly, to modify it.
- ✓ R is a programming language and environment developed for statistical analysis by practicing statisticians and researchers.
- ✓ The graphical capabilities of R are outstanding, providing a fully programmable graphics language that surpasses most other statistical and graphical packages.
- ✓ R has no license restrictions, so run it anywhere and at any time, and even sell it under the conditions of the license.
- ✓ R has over 4800 packages available from multiple repositories specializing in topics like econometrics, data mining, spatial analysis, and bio-informatics.

The programming features available in R

R has data structures (vectors, matrices, arrays, data frames) that users can operate on through functions for performing statistical analyzes and creating graphs. R has a wide variety of data types including scalars, vectors. Importing data into R is fairly simple such as a comma delimited text file, Excel, SPSS, SAS, Stata, system, etc. R's programming features include database input, exporting data, viewing data, variable labels, missing data, etc. According to R Language features (2015), R programming language has two superior essential features that cover all contemporary

applications being useful and productive in analyzing big datasets as shown below:

- ✓ Input/output: text, .csv, binary urls, XML, mysql, ODBC, Oracle, etc.
- ✓ Object-oriented programming: C, Java, Perl, Python, parallel programming, etc.
- ✓ Distributed computing: Amazon EC2 compatibility
- ✓ Included R Packages: base, compiler, datasets, etc.; MASS, cluster, lattice, etc.

Describe how the analytics of R are suited for Big Data:

Big Data analytics that processes data into information and particularly knowledge has emerged as a contemporary business trend in industry and academics .Many consultants, scientists, and researchers pay attention to Big Data because it contains meaningful information. Various applications such as healthcare, security, medicine, politics, etc. can use the information to solve data-related problems in society.

V. CONCLUSION

A framework for the clustering of big data using Heuristic Data Intensive computing and Self-Organizing Maps algorithm has been proposed. The Heuristic concept is to give optimal solution while SOM algorithm is for the clustering of big data. There is a need to be able to collect this data, analyze and visualize and process in a parallel manner in a distributed way.

SOM algorithm has many advantages to be used in big data mining because it has the ability to scale with the size of the data set, prior knowledge of the number of expected clusters is not needed and easy to integrate with clusters ensemble model. Big data analysis opens the door for many research areas and one of the most important areas is the data security.

VI. REFERENCES

[1]. Agneeswaran, V. S. (2012). Big-data-theoretical, engineering and analytics perspective. In S.Srinivasa & V. Bhatnagar (Eds.), Big Data Analytics SE-2Berlin, Germany: Springer-Verlag.

[2]. Brzezniak, M., Meyer, N., Flouris, M., Lachaiz, R. & Bilas, A. (2008). Analysis of grid storage element architectures: high-end fiber-channel vs. emerging cluster-based networked storage. In M. Brzezniak, N. Meyer, M. Flouris, R. Lachaiz & A. Bilas (Eds.), Grid middleware and services SE , US: Springer.

[3]. Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S. & Zhou, X. (2013). Big data challenge: a data management perspective. Frontiers of Computer Science.

[4]. Das, S., Abraham, A. & Konar, A. (2009). Metaheuristic pattern clustering-an overview. Metaheuristic Clustering, Berlin, Germany: Springer-Verlag.

[5]. Deneubourg, J.L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C. & Chretien, L. (1990). The dynamics of collective sorting robot like ants and ant like robots. Proceedings of the 1stInternational Conference on Simulation of Adaptive Behaviour: From Animals to Animats.

[6]. Hall, L.O. (2013). Exploring big data with scalable soft clustering. In R. Kruse, M. R. Berthold, C.Moewes, M.Á. Gil, P. Grzegorzewski & O. Hryniewicz (Eds.), Synergies of Soft Computing andStatistics for Intelligent Data Analysis, Berlin, Germany: Springer-Verlag.

[7]. Kim, B. (2012). A classifier for big data. In G. Lee, D. Howard, D. slezak & Y. Hong (Eds.), Convergence and Hybrid Information Technology, Berlin: Germany: Springer-Verlag.

[8]. Madheswari, A.N. & Banu, R.S.D.W. (2011). Communication aware co-scheduling for parallel jobscheduling in cluster computing. In

A. Abraham, J. Lloret Mauri, J. Buford, J. Suzuki & S.Thampi (Eds.), Advances in Computing and Communications, Berlin, Germany:Springer.

[9]. Qin, X. (2012). Making use of the big data: next generation of algorithm trading. In J. Lei, F. Wang, H.Deng & D. Miao (Eds.), Artificial Intelligence and Computational Intelligence, Berlin,Germany: Springer-Verlag.

[10]. Strehl, A. & Ghosh, J. (2002). Cluster ensembles-a knowledge reuse framework for combiningmultiple partitions. Journal of Machine Learning Research.

[11]. A. Fahad, N. Alshatri and Z. Tari, "A Survey of Clustering Algorithms for Big Data: Taxonomy", IEEE Transactions on Emerging Topics in Computing 2014.

[12]. Btissam Zerhari, Ayoub Ait Lahcen and Salma Mouline, "Big Data Clustering: Algorithms and Challenges", International Conference on Big Data, Cloud and Applications BDCA'15 , At Tetuan, Morocco , conference paper may 2015.

[13]. Apurva Juyal Dr. O. P. Gupta,"A Review on Clustering Techniques in Data Mining",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 7, July 2014.

[14]. Keshavanse, Meena Sharma,"Clustering methods for Big data analysis",International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 3, March 2015.

[15]. S.M. Junaid, K.V. Bhosle," Overview of Clustering Techniques", International Journal of Advanced Research in Computer Science and Software Engineering,Volume 4, Issue 11, November 2014.

[16]. DongkuanXu and YingjieTian, "A Comprehensive Survey of Clustering Algorithms", Annals of Data Science, Springer-Verlag Berlin Heidelberg August 20.

[17]. C. YADAV, S. WANG, et M. KUMAR, "Algorithm and approaches to handle large Data-A Survey," International Journal of computer science and network, vol 2, issue 3, 2013.

[18]. Manish Kumar Kakhani, Sweeti Kakhani and S.R. Biradar, "Research Issues in Big Data Analytics", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 8, August 2013.

[19]. Justin Samuel, Koundinya RVP, KothaSashidhar and C.R. Bharathi, A Survey on Big Data and its Research Challenges, ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 8, May 2015.