ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

A review of Big Data Clustering Information Retrieval in Health Informatics

M.Thanaraj[1],G.Sujatha[2], P.Punitha Ponmalar[3]

Associate professor
Department of Computer Science
[1]Madurai kamaraj University, Madurai,India
thangarajmku@Yahoo.com
[2] Sri Meenakshi Govt College for Women, Madurai, India
Sujisekar05@rediffmail.com
*Second Company
[3]Sri Meenakshi Govt College for Women, Madurai, India
p_punithanadraj74@yahoo.co.in

***Abstract*— Big data is generating a lot of hype in every industry including healthcare. Healthcare organizations produce and collect large volumes of information from variety of sources. Raw data from healthcare organizations are voluminous and heterogeneous. They need to be collected and stored in the organized forms, and their integration enables forming of hospital information system. The basic aim of health informatics is to take in real world medical data from all levels of human existence for better understanding of medicine and medical practices. A number of use cases in healthcare are well suited for a big data solution. Big data indexing techniques add real value to healthcare analytics in the future. Predictive analytics and Real-time alerting are the major issues of big data.**

This paper on Big Data Clustering Information Retrieval in Health Informatics (BCIRH) provides countless possibilities for hidden pattern retrieval from health care informatics. These patterns can be used by physicians to determine diagnoses, prognoses and treatments for patients in healthcare organizations.

***Keywords*— Big Data, Health Care, Information Retrieval, Predictive analytics, Data Profiling**

I. Introduction

Big Data is a term which is used to describe massive amount of data generating from digital sources or the internet. From the past few years data is exponentially growing due to the use of connected devices such as smart phone's, tablets, laptops and desktop computer. This generated data volume is so vast and overwhelming which makes complex to process and analyze using traditional software systems consuming more time.

Doug Laney [3] discussed about 3 V's in Big Data management are shown in fig 1.

Volume : Volume is the amount of data generated by organizations or individuals. Today the size of data are increased to Exa bytes. The grand scale and rise of data outstrips traditional store and analysis techniques [4].

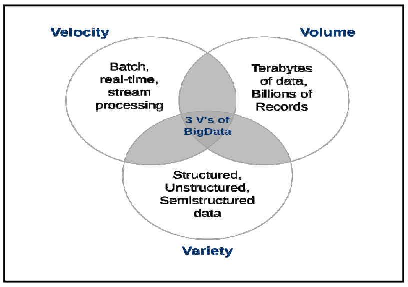
Variety : Variety makes health care data really big. Big data comes from a great variety of sources such as internal, external, social and behavioural and generally has in three types: structured, semi structured and unstructured. Structured data inserts a data warehouse already tagged and easily sorted but unstructured data is random and difficult to analyze. Semi-structured data does not conform to fixed fields but contains tags to separate data elements [4][13].

Velocity: Velocity describes the rate at which data is generated, captured and shared [4][16]. Healthcare Data can be generated from two sources: humans, or sensors. With a few exceptions like diagnostic imaging and intensive care monitoring, most of the data used in healthcare is entered by people, which effectively limits the rate at which healthcare organizations can generate data.

Nowadays, there are 2 more V's:

Variability: There are changes in the structure of the data and how users want to interpret that data.

Value: Business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach [6].


Velocity
Volume
Batch, real-time, stream processing
Terabytes of data, Billions of Records
3 V's of BigData
Structured, Unstructured, Semistructured data
Variety

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Fig 1: The 3V-s of Big Data

Healthcare predictive analytics today requires the processing of big data relating to hospital and patients administrative data, clinical and non-clinical data including patient demographics, disease diagnoses and procedures, patient charges, medical health records, discharge status. This big data needs to be processed and analyzed to extract knowledge for decision-making and cost-saving.

Big Data analytics provide a set of tools that can be applied to detect patterns, classifications, hospital transfers, and mortality.

.

II. Related Work

Fahad et al. [5] present a survey of existing clustering algorithms of different categories. In [2] the authors focus on the most popular and most used algorithms in the literature like k-means, they presents some comparative work of these algorithms. Another recent research [11] presents a general view of data mining algorithms and platforms that can be used in the field of Big Data by discussing different challenges and characteristics. Paper [4] discusses some of Big Data mining algorithms to find the most appropriate among them using a comprehensive comparison. Others in [14] are discussed classification algorithms and how it is used in statistics and apply them to specific databases. Researchers in [7] present a review of some old algorithms that can handle large data set as Nearest Neighbor Search, Decision Tree and Neural Network. In [8], Laney et al. discuss different data management techniques. They present an overview of different categories of data mining. The scalability of the parallel k-means algorithm has also been demonstrated [10]. In [12] proposes a classification algorithm for Big Data based on feature selection. Cui [3] addresses the Big Data processing problem using the K-means algorithm that proposes a new model of treatment with Map Reduce to eliminate iteration dependency and achieve high performance.

Our study covers all the above techniques. It deals with different categories of data mining clustering algorithms and discusses their advantages and disadvantages.

III. Proposed Big Data Clustering Information Retrieval in Health Informatics (BCIRH)

The BCIRH consists of four important levels that carry out Data Ingestion, Data Analytics, Information Analysis and Information Consumption. Information technologies in healthcare have enabled the creation of electronic patient records obtained from monitoring of the patient visits. Flow model of Big data shown in Fig 2.

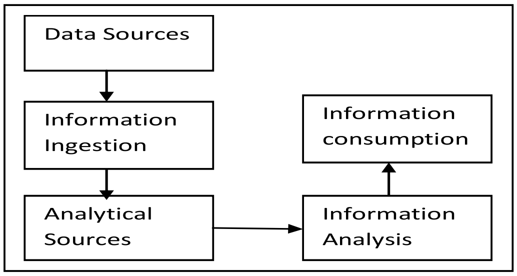


Fig 2: Flow model of BCIRH

The information includes patient demographics, records on the treatment progress, details of examination, prescribed drugs, previous medical history, lab results, Social Media, Biometric Data, Electronic Medical Record (EMR),etc. The large volumes of data collected from different sources are fed into the BCIRH..
The different level of BCIRH is explained here with the supporting components in Fig 3.

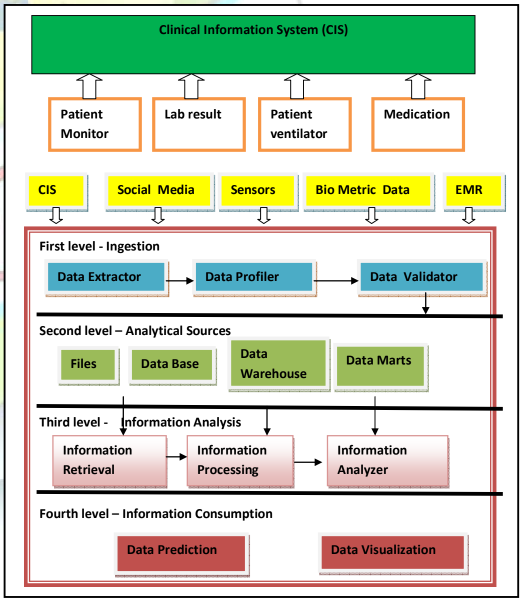


Fig 3 Architecture of BCIRH

A. First Level : Ingestion

Data ingestion is the process of obtaining, importing, and processing data for later use or storage in a database. This process often involves altering individual files by editing their content and/or formatting them to fit into a larger document. Medical Data Ingestion is carried out with Data Extractor, Data Profiler and Data Validator. Data Extractor extracts data from different sources and the preprocessing work is carried out with Data Profiler which is then passed to Data Validator for application dependent validation.

Profiling evaluates the actual content, structure and quality of the data by exploring relationships that exist between value collections both within and across data sets. Profiling and validation is carried out with the help of variety of clustering algorithms. Clustering in Health care data is required to identify the existing patterns which are not clear in first glance. The properties of big data pose some challenge against adopting traditional clustering methods:

Type of dataset (Variety): May contain both numeric and categorical attributes. Clustering algorithms work effectively either on purely numeric data or on categorical data, most of them perform poorly on mixed categorical and numerical data types.

Size of dataset (Volume): The size of the dataset has effect on clustering time – efficiency and the clustering quality.

Handling outliers/ noisy data: Data from Medical applications suffers from noisy data which pertains to faults and misreported readings from wearable health sensors.

Time Complexity: Most of the clustering methods must be repeated several times to improve the clustering quality.

Stability: Stability corresponds to the ability of an algorithm to generate the same partition of the data irrespective of the order in which the data are presented to the algorithm.

Cluster shape: A good clustering algorithm should be able to handle real data and their wide variety of data types, which will produce clusters of arbitrary shape.

In the Health care Iinformation Retrieval (IR) field, cluster analysis has been used to create groups of patients/diseases the goal of benefiting the efficiency and effectiveness of retrieval. Cluster analysis is an unsupervised study method without training data.

1. *K-Means: A Centroid - Based Technique :*

The k -means algorithm defines the centroid of a cluster as the mean value of the points within the cluster. First, it randomly selects k of the objects in D, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the Euclidean distance between the object and the cluster mean. The k-means algorithm then iteratively improves the within - cluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration clusters formed in the current round are the same as those formed in the previous round. The k - means procedure along with algorithm is given below.

Algorithm K-means:

Input = K:
The number of clusters = D: A dataset containing n objects
Output = A set of K clusters

Method:
(1) Arbitrarily choose K - objects from D as the initial cluster centers
(2) Repeat
(3) Re-assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster
(4) Update the cluster means, i.e .calculate the mean value of the objects for each clusters
(5) Until no changer

The time complexity of the k-means algorithm is O(nkt),where n is the total number of objects, k is the number of clusters, and t is the number of iterations. Normally, $k \ll n$ and $t \ll n$. Therefore, the method is relatively scalable and efficient in processing large data sets.

2. *D-stream Clustering Algorithm :*

D-Stream has an online component and an offline component. For a data stream, at each time step, the online component of D - Stream continuously reads a new data record, place the multidimensional data into a corresponding discretized density grid in the multi-dimensional space, and update the characteristic vector of the density grid. The density grid and characteristic vector are to be described in detail later. The offline component dynamically adjusts the clusters every gap time steps, where gap is an integer parameter. After the first gap, the algorithm generates the initial cluster. Then, the algorithm periodically removes sporadic grids and regulates the clusters.

3. *CLARA :*

CLARA (CLustering LARge Applications) relies on the sampling approach to handle large data sets. Instead of finding medoids for the entire data set, CLARA draws a small sample from the data set and applies the PAM algorithm to generate an optimal set of medoids for the sample. The quality of resulting medoids is measured by the average dissimilarity between every object in the entire data set D and the medoid of its cluster, defined as the following cost function:

$$Cost(M,D) = \frac{\sum_{i=1}^{n} dissimilarity(O_i, rep(M,O_i))}{n}$$

where M is a set of selected medoids, dissimilarity(Oi, Oj) is the dissimilarity between objects Oi and Oj, and rep(M, Oi) returns a medoid in M which is closest to Oi.

To alleviate sampling bias, CLARA repeats the sampling and clustering process a pre-defined number of times and subsequently selects as the final clustering result the set of medoids with the minimal cost. Assume q to be the number of samplings.

4. *Fuzzy C-Means* :

In the K-means algorithm, each vector is classified as belonging to a single cluster (hard clustering), and the centroids are updated based on the classified samples. In a variation of this approach known as fuzzy c-means all vectors have a degree of membership for each cluster, and the respective centroids are calculated based on these membership degrees.

Whereas the K-means algorithm computes the average of the vectors in a cluster as the center, fuzzy c-means finds the center as a weighted average of all points, using the membership probabilities for each point as weights. Vectors with a high probability of belonging to the class have larger weights, and more influence on the centroid.

As with K-means clustering, the process of assigning vectors to centroids and updating the centroids is repeated until convergence is reached.

5. *Hierarchical Clustering* **:**

Hierarchical clustering creates a hierarchical tree of similarities between the vectors, called a dendrogram. The usual implementation is based on agglomerative clustering, which initializes the algorithm by assigning each vector to its own separate cluster and defining the distances between each cluster based on either a distance metric (e.g., Euclidean) or similarity (e.g., correlation). Next, the algorithm merges the two nearest clusters and updates all the distances to the newly formed cluster *via* some linkage method, and this is repeated until there is only one cluster left that contains all the vectors. Three of the most common ways to update the distances are with *single*, *complete* or *average* linkages.

6. Self Organizing Maps *:*

By applying self organizing maps (SOM) to the data, clusters can be defined by points on a grid adjusted to the data. Usually the algorithm uses a 2-dimensional grid in a higher dimensional space, but for clustering it is typical to use a 1-dimensional grid.

SOM clustering is very useful in data visualization since the spacial representation of the grid, facilitated by its low dimensionality, reveals a great amount of information on the data.

7. *Gaussian Mixture Model :*

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system.

GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model.

8. *Generalized K-Harmonic Means* **:**

K-Harmonic means has a built-in dynamic weighting function which tries in boosting the data objects which are not nearer in the current iteration. In this approach, weighting function is automatically changed after each iteration. The K-Harmonic means is insensitive to the data set's initialization. First the initializations are made then the convergence factor is being calculated which enhances the efficiency. Harmonic value will always tend to be minimum and thus behaves as min () function rather than as avg () function. K-Harmonic means are generally been used in many real-world datasets. The dynamic weighting approach here makes the algorithm being built robust to inefficient initializations and noise in the data sets. The various important clustering techniques were discussed in Table1.

B. Second Level: Analytical Sources:

The profiled big data can be stored in different types of sources for the purpose of analytics like flat data file, relational data, data warehouse, data marts, Nosql data, object-oriented data and hierarchical data.

C.Third Level: Information Analysis:

The size of the data increases the time required and efficiency reduces. Soft computing techniques used in health care to improve the quality of data by reducing imprecision and uncertainty employing approximate reasoning and logic, in order to achieve tractability, robustness and low cost solutions. The learning process of Big data is carried out in this level with the soft computing techniques like Neural Network, Fuzzy, Genetic or with the tools like SPSS, R, SAS. Artificial Neural Networks (ANN) can be used for prediction accuracy and Genetic Algorithms to optimize the error function. As a result of this analysis an intelligent solution for the healthcare is presented to the next level [15]

TABLE I
COMPARATIVE ASPECTS OF CENTROID BASED ALGORITHMS

S.No	Sensitive to noise	Outlier	Structure-centric	Minimize Intra Cluster Variance	Complexity
K-Means	Very High	Very Sensitive	Yes	No	$O(n^{d-1} \log n)$
K-Medoids	Optimum	Sensitive	Yes	No	$O(K(n-K)^2)$
CLARA	Optimum	Kick-off to study	No	Yes	$O(ks^2 +k(n-k))$
CLARANS	Very Low	Deals with outliers	No	Yes	$O(n^2)$
k-Harmonic means	High	Sensitive	Yes	No	$O(n^2 \log n)$
Fuzzy c-means	Optimum	Kick-off to study	Yes	No	$O(n^{1-v}\ lo\ n)$

1. Artificial neural networks

ANN is an analytical technique that is formed on the basis of superior learning processes in the human brain. As the human brain is capable to, after the learning process, draw assumptions based on previous observations, neural networks are also capable to predict changes and events in the system after the process of learning. Neural networks are groups of connected input/output units where each connection has its own weight. The learning process is performed by balancing the net on the basis of relations that exist between elements in the examples. Based on the importance of cause and effect between certain data, stronger or weaker connections between "neurons" are being formed.

Artificial neural networks are ideal for multiprocessor systems, where a large number of operations are performed in parallel. Artificial Neural Networks (ANN), a supervised learning data mining approach, is selected as the classifier as it is sensitive to non-linear input values and improves the prediction efficiency. The pre processed data, worked on: as per the proposed pre processing framework is input to the prediction model in fig 4.

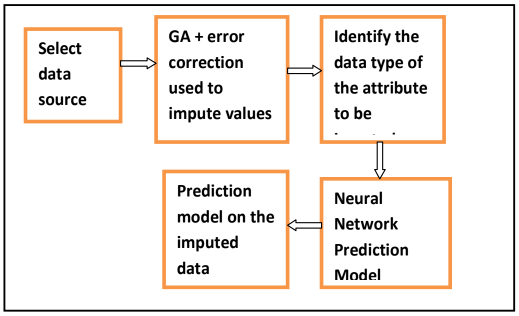

Select data source
GA + error correction used to impute values
Identify the data type of the attribute to be
Neural Network Prediction Model
Prediction model on the imputed data

Fig. 4 Data Prediction Model

2. Decision tree

It is a graphical representation of the relations that exist between the data in the database. It is used for data classification. The result is displayed as a tree, hence the name of this technique. Decision trees are mainly used in the classification and prediction. It is a simple and a powerful way of representing knowledge. The models obtained from the decision tree are represented as a tree structure. The instances are classified by sorting them down the tree from the root node to some leaf node. The nodes are branching based on if-then condition. Tree view is a clear and easy to understand, decision tree algorithms are significantly faster than neural networks and their learning is of shorter duration. Decision tree is a tree where each (non-terminal) node represents a test or decision on the item of information that is listed for consideration. The choice of a particular industry depends on the outcome of the test. In order to classify the data, process is starting from the root node and following the argument down until it reaches the final node, at which time a decision is made. Decision tree can also be interpreted as a special form of a rule set, which is characterized by its hierarchical organization of rules.

3. Genetic algorithms

It is based on the principle of genetic modification, mutation and natural selection. These are algorithmic optimization strategies inspired by the principles observed in natural evolution. The genetic algorithm creates a number of random solutions to the problem. All these solutions may not be good, a group of solutions can be skipped entirely, and it can come down to the overlapping solutions. Poor solutions are discarded, and the good ones retained. A good solution is then being hybridized, and then the whole process is repeated. Finally, similar to the process of natural selection, only the best solutions remain. So, from the set of potential solutions to the problems that compete with each other, the best solutions are chosen and combined with each other in order to obtain a universal solution from the set of solutions that will become better and better, similar to the process of evolution of organisms. Genetic algorithms are used in data mining to formulate hypotheses about the dependencies between variables in the form of association rules or other internal formalism. The disadvantage of this method is that it requires an enormous amount of processing power and it is too slow for trivial issues. Since evolutionary computation is a robust and parallel search algorithm, it can be used in data mining to find interesting knowledge in noisy environment.

Using a Genetic Algorithm a local unconstrained minimum is found x, to the objective function, fitness function.

x = GA(fitness function, nvars) (1)

nvars is the dimension (number of design variables) of fitness function. The objective function, fitness function, accepts a vector x of size 1-by-nvars, and returns a scalar evaluated at x. The predicted value for the incorrect or missing value is corrected using this optimization through genetic algorithms. The steps involved in this process is

1. Select data-subset without any missing values out
2. Build regression model to impute missing values
3. Compute error function for predicted values
4. Optimize error function using Genetic Algorithm
5 .Impute value correction
6. Completed data set, imputing continuous values

4. Nearest neighbour method

It is a technique that is also used for data classification. Unlike other techniques, there is no learning process to create a model. The data used for learning is in fact a model. When the new data shows up, the algorithm analyzes all the data in the database to find a subset of instances that are the best fit and based on that it is able to predict the outcome. The study conducted on the application of nearest neighbour method on benchmark data set to detect efficiency in the diagnosis of heart diseases, produced the results that application of this method achieved an accuracy of 97.4% which is a higher percentage than any other published study on the same set of data.

C. Fourth Level – Information Consumption:

The predicted results are presented to the use cases by means of some visualization tools.

IV. CONCLUSIONS

Big Data demands a revolutionary change in research methodology and in tools to be employed. The overall goal of this process is to extract information from a large data set and transform it into an understandable form. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). The combination of data mining method and soft computing tools like ANN, GA, FL, play a significantly important role in this due to their inherent capabilities of dealing with imprecision and uncertainty.

REFERENCES

[1] F. Bu, Z. Chen, Q. Zhang, and X. Wang, “Incomplete Big Data Clustering Algorithm Using Feature Selection and Partial Distance,” In Digital Home (ICDH), 5th International Conference on. IEEE, p. 263-266, 2014.

[2] A.BEN AYED, M.BEN HALIMA and M. ALIMI, “Survey on clustering methods: Towards fuzzy clustering for Big Data,” In Soft Computing and Pattern Recognition (SoCPaR), 6th International Conference of. IEEE, p. 331-336, 2014.

[3] X. Cui, P. Zhu, X. Yang, K. Li, and C. Ji, “Optimized Big Data K-means clustering using MapReduce,” The Journal of Supercomputing, vol. 70, no3,p.1249-1259,2014.

[4] C. Eaton, D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Mc Graw-Hill Companies, 978-0-07-179053-6, 2012

[5] Fahad, Adil, et al. "A survey of clustering algorithms for big data: Taxonomy and empirical analysis." *Emerging Topics in Computing, IEEE Transactions on* 2.3 (2014): 267-279.

[6] W. Fan, A Bifet, "Mining big data: current status, and forecast to the future", ACM SIGKDD Explorations Newsletter, Vol. 14, pp.1-5, 2013.

[7] D. Jianqiang, W. Fei and Y. Bo, “Accelerating BIRCH for clustering large scale streaming data using CUDA dynamic parallelism,” In Intelligent Data Engineering and Automated Learning–IDEAL 2013. Springer Berlin Heidelberg, p. 409-416, 2013.

[8] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001

[9] S. Madden, "From Databases to Big Data", IEEE Internet Computing, June 2012, v.16, pp.4-6

[10] H. S. Nagesh, S. Goil, and A. Choudhary, “A scalable parallel subspace clustering algorithm for massive data sets,” In Parallel Processing, 2000. Proceedings. International Conference on. IEEE, p. 477-484, 2000

[11] A. Sherin, S. Uma, K.Saranya and M. Saranya Vani ”Survey On Big Data Mining Platforms, Algorithms And Challenges”. International Journal of Computer Science & Engineering Technology,Vol. 5 No, 2014.

[12] S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, “Big Data Clustering: A Review,” In Computational Science and Its Applications–ICCSA 2014. Springer International Publishing, p. 707-720. 2014.

[13] S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011

[14] R. XU and D. WUNSCH, “Survey of clustering algorithms,” Neural Networks, IEEE Transactions, vol. 16, no 3, p. 645-678, 2005.

[15] V.Bhat, H., Rao, P. G., Shenoy, P. D., Venugopal, K. R., Patnaik, L. M. :An Efficient Prediction Model for Diabetic Database using Soft Computing Techniques. LNCS, vol. 5908/2009, pp. 328-335. Springer Heilelberg (2009). *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997