ISSN 23499842(Online), Volume 1, Special Issue 1(NCFCSIT 15), March 2015
International Journal of Innovative Trends and Emerging Technologies

MINING BIGDATA IN HEALTHCARE

R.SNEHAA
PG Scholar, Department of CSE, Agni College of Technology, Chennai, snehaaagni@gmail.com

ABSTRACT: Big Data is a new term used to identify the datasets that due to their large size and complexity. Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. we cannot manage them with our current methodologies or data mining software tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity. Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and bio-medical sciences. With HACE theorem, when analyzing the inconsistent data such as sparse data, incomplete data, and Corrupted files, it becomes difficult to analyses and recover the inconsistent data. In the proposed approach , Clustering (Using Hierarchical Clustering), Classification (Using See5),Rules (Jrip) is used to develop an application called "Recommendation System for Pediatric Health Care". Here, the medical record is going to be analyzed and data mining techniques are going to be applied for that record to get the recommended output. Mining from Sparse, Uncertain, and Incomplete Data is perfectly analyzed using proposed mining techniques such as Clustering (Using Hierarchical Clustering), Classification (Using See5), Rules (Jrip).

Keywords: Big data, Clustering, Classification, Association rules.

I.INTRODUCTION:

Recent years have viewed a dramatic increase in our ability to collect data from various sensors, devices, in different formats, from independent or connected applications. This data flood has outpaced our capability to process, analyze, store and understand these datasets. Consider the Internet data. The web pages indexed by Google were around one million in 1998, but quickly reached 1 billion in 2000 and have already exceeded 1 trillion in 2008. This rapid expansion is accelerated by the dramatic increase in acceptance of social networking applications, such as Face book, Twitter, Weibo, etc., that allow users to create contents freely and amplify the already huge Web volume. Furthermore, with mobile phones becoming the sensory gateway to get real-time data on people from different aspects, the vast amount of data that mobile carrier can potentially process to improve our daily life has significantly outpaced our past CDR(call data record)-based processing for billing purposes only. It can be foreseen that Internet of things (IoT) applications will raise the scale of data to an unprecedented level. People and devices (from home coffee machines to cars, to buses, railway stations and airports) are all loosely connected. Trillions of such connected components will generate a huge data ocean, and valuable information must be discovered from the data to help improve quality of life and make our world a better place. In all these applications, we are facing significant challenges in leveraging the vast amount of data, including challenges in (1) system capabilities (2) algorithmic design (3) business models. In Big data mining, Doug Laney was the first one in talking about 3 V's in Big Data management:

1.Volume: there is more data than ever before, its size continues increasing, but not the percent of data that our tools can process

2.Variety: there are many different types of data, as text, sensor data, audio, video, graph, and more

3.Velocity: data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time .

Nowadays, there are two more V's:

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

5.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 Gartner summarizes this in their definition of Big Data in 2012 as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. There are many applications of Big Data, 1.Business: costumer personalization, churn detection

2.Technology: reducing process time from hours to seconds

3.Health: mining DNA of each person, to discover, monitor and improve health aspects of every one

4.Smart cities: cities focused on sustainable economic development and high quality of life, with wise management of natural resources. These applications will allow people to have better services, better costumer experiences, and also be healthier, as personal data will permit to prevent and detect illness much earlier than before.

1.1:EXISTING SYSTEM:

In the Existing System Big Data rapidly expanding in all science and engineering domains, including physical, biological and bio-medical sciences. This article presents a HACE theorem . Big Data Concerns a large volume of Data considering Many Rapidly Expanding Activities in which data integrated into a data-sever Resulting with Difficulties in mining. Due to heterogeneous environment in big data, the aggregation of data is withheld. Data Driven Model Only provides the Architecture for mining particular data thus lack in heterogeneous Maintenance thus becomes a challenge. Each Servers is fully functional relying on other servers thus integration of inconsistence data will migrate the servers. When the inconsistent data such as sparse data, incomplete data, and Corrupted files are analyzed, it becomes difficult to analyses and recover the inconsistent data. While mining the fusion data from various resources there is no preprocessing technologies or data filtering technologies. In Medium scale data mining tasks, data are typically large (and possibly distributed) and cannot be fit into the main memory of the server.

1.2 PROPOSED APPROACH:

In the proposed system Clustering (Using Hierarchical Clustering), Classification (Using See5), Rules (Jrip) is going to be used and pediatric health care medical record is going to be analyzed and data mining techniques are going to be applied for that record to get recommended output. Data mining Techniques like Clustering, classification and Rules to be generated for the pediatric medical record to get the treatment and drugs for the patient. Provides a certain amount of information and each server is able to fully function without necessarily relying on other servers. Mining from Sparse, Uncertain, and Incomplete Data is Perfectly analyzed using mining techniques such as Clustering (Using Hierarchical Clustering), Classification (Using See5),Rules (Jrip). By using CLUSTERING**(Hierarchical Clustering Algorithm)**, Smaller clusters are generated, which may be helpful for discovery. By using **JRIP Algorithm** it is Efficient and Rules

generated accurately. By using **SEE5 Algorithm** it increase its Speed, Memory usage is less, Smaller decision trees can be generated.

II.RELATED WORK

We selected four contributions that together shows very significant state-of-the-art research in Big Data Mining, and that provides a broad overview of the field and its forecast to the future. Other significant work in Big Data Mining can be found in the main conferences as KDD, ICDM, ECMLPKDD,or journals as "Data Mining and Knowledge Discovery" or "Machine Learning".

- Scaling Big Data Mining Infrastructure:

The Twitter Experience by Jimmy Lin and Dmitriy Ryaboy (Twitter,Inc.). This paper presents insights about Big Data mining infrastructures, and the experience of doing analytics at Twitter. It shows that due to the current state of the data mining tools, it is not straightforward to perform analytics. Most of the time is consumed in preparatory work to the application of data mining methods, and turning preliminary models into robust solutions.

- Mining Heterogeneous Information Networks:

A Structural Analysis Approach by Yizhou Sun (North-eastern University) and Jiawei Han (University of Illinois at Urbana-Champaign). This paper shows that mining heterogeneous information networks is a new and promising research frontier in Big Data mining research. It considers interconnected, multi-typed data, including the typical relational database data, as heterogeneous information networks. These semi-structured heterogeneous information, network models leverage the rich semantics of typed nodes and links in a network and can uncover surprisingly rich knowledge from interconnected data.

- Big Graph Mining:

Algorithms and discoveries by U Kang and Christos Faloutsos (Carnegie Mellon University). This paper presents an overview of mining big graphs, focusing in the use of the Pegasus tool, showing some findings in the Web Graph and Twitter social network. The paper gives inspirational future research directions for big graph mining.

- Mining Large Streams of User Data for Personalized Recommendations by Xavier Amatriain (Netix).

This paper presents some lessons learned with the Netix Prize, and discusses the recommender and personalization techniques used in Netix. It discusses recent important problems and future research directions. Section 4 contains an interesting discussion about if we need more data or better models to improve our learning methodology.

- Knowledge Based Cluster Ensemble for Cancer Discovery From Biomolecular Data:

Most of the existing works adopt single clustering algorithms to perform class discovery from biomolecular data. However, single clustering algorithms have limitations, which include a lack of robustness, stability, and accuracy.In this paper, we propose a new cluster ensemble approach called knowledge based cluster ensemble (KCE) which incorporates the prior knowledge of the data sets into the cluster ensemble framework.Specifically, KCE represents the prior knowledge of a data set in the form of pairwise constraints. Then, the spectral clustering algorithm (SC) is adopted to generate a set of clustering solutions.Next, KCE transforms pairwise constraints into confidence factors for these clustering solutions. After that, a consensus matrix is constructed by considering all the clustering solutions and their corresponding confidence factors. The final clustering result is obtained\ by partitioning the consensus matrix. Comparison with single clustering algorithms and conventional cluster ensemble approaches, knowledge based cluster ensemble approaches are more robust, stable and accurate.

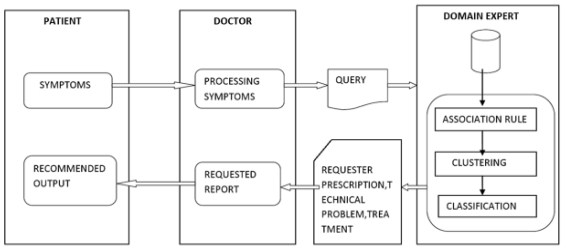
- Data Mining to Generate Adverse Drug Events Detection Rules:

Different kinds of outcomes are traced, and supervised rule induction
methods (decision trees and association rules) are used to discover ADE detection rules, with respect to time constraints. The rules are then filtered, validated, and reorganized by a committee of experts.The rules are described in a rule repository, and several statistics are automatically computed in every medical department, such as the confidence, relative risk, and median delay of outcome appearance. 236 validated ADE-detection rules are discovered; they enable to detect 27 different kinds of outcomes. The rules use a various number of conditions related to laboratory results, diseases, drug administration, and demographics. Some rules involve innovative conditions, such as drug discontinuations.

III.SYSTEM DESIGN:

DIAGRAM:

Patient will give their Symptoms (Query) to the Doctor. Doctor will analyze the Query and give their medical report to the Domain Expert. Domain Expert will analyze the report and that report is compared with the Domain Expert Database (i.e. Big Data) and Treatment and Drugs and their Technical problems are prescribed for their disease. Some Data Mining techniques such as Rules, Classification, and Clustering are used to prescribe their problem, treatment and drugs.



IV.MODULE IDENTIFICATION:

- Data Ruling Module using J-rip technique.
- Hierarchical Clustering Module for the Optimized result.
- Finding the Decision Module for classification Using See5.

V.MODULE DESCRIPTION:

- Data Ruling Module using J-rip technique:

1. Building stage- The description length (DL) of the rule set.

1.1. Grow phase:

Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain.

1.2. Prune phase:

Incrementally prune each rule and allow the pruning of any final sequences of the antecedents.

2. Optimization stage:

After generating the initial ruleset {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (TP+TN)/(P+N).Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the ruleset.After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

3. Delete the rules from the rule set that would increase the DL of the whole rule set if it were in it and add resultant rule set to RS.

- Hierarchical Clustering Module for the Optimized result:

Start with the points as individual clusters. At each step, merge the closest pair of clusters until only one cluster left.

- Finding the Decision for classification Using See5:

Decision Trees are constructed using See5 Algorithm. This Algorithm is used to construct a smaller decision tree. So that we can easily identify the recommended output.

VI.PROPOSED METHODOLOGY

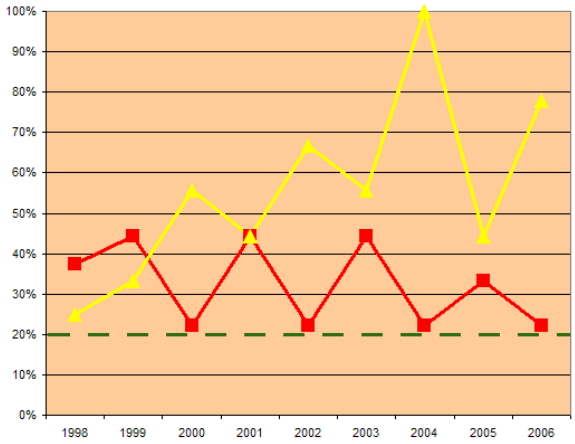
Patient will give their Symptoms (Query) to the Doctor. Doctor will analyze the Query and give their medical report to the Domain Expert. Domain Expert will analyze the report and that report is compared with the Domain Expert Database (i.e. Big Data) and Treatment and Drugs are prescribed for their disease. Some Data Mining techniques such as Rules(JRIP ALGORITHM), Classification(SEE5 ALGORITHM), and Clustering(HIERARCHICAL CLUSTERING ALGORITHM) are used to prescribe their problem, treatment and drugs. These Algorithm Can be run in Weka Tool, which is binded with the Neat Beans (Java Framework).

VII. PROPOSED ALGORITHM:

In the proposed system data mining techniques such as clustering, decision tree and decision rules are going to be carried out for the application pediatric healthcare clinical records (Big Data).For clustering ,Hierarchical Clustering Algorithm is to be used. For Decision Tree, Enhanced See5 Algorithm is going to be proposed. For Rules, Enhanced Jrip Algorithm is going to be proposed. By using CLUSTERING **(Hierarchical Clustering Algorithm)**, Smaller clusters are generated, which may be helpful for discovery. By using **JRIP Algorithm** it is Efficient and Rules generated accurately. By using **SEE5 Algorithm** it increase its Speed, Memory usage is less, smaller decision trees can be generated.

VIII.RESULT AND DISCUSSION:

From the result we can See that Data mining techniques such as clustering, classification and Association rule mining is efficient than the HACE theorem to get an accurate result for the above application because inconsistent datas like sparse,incomplete and corrupted files are easily analyzed here.



IX.CONCLUSION:

Big Data is a new term which is used to identify the datasets that due to their large size and complexity. It concerns large-volume, complex, growing data sets with multiple, autonomous sources. we cannot manage them with our current methodologies or data mining software tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity. Initial pool of preoperative data elements (prior to feature selection) comprised of four categories: demographics, hemodynamics, laboratory values, and medications used for making decision trees by using SEE5 Algorithm. IF(condition_1 &. . .& condition) THEN outcome is used for rule generation by JRIP Algorithm. knowledge based cluster ensemble approach works well in data sets. Clusters are formed by using hierarchical clustering.

X.REFERENCES:

]. Qi Liu, Enhong Chen "A Cocktail Approach for Travel Package Recommendation" IEEE *IEEE Trans. Knowl. Data Eng*, Vol. 26, NO. 2, February 2014

[2]. Q. Liu, Y. Ge, Z. Li, H. Xiong, and E. Chen, "Personalized Travel Package Recommendation," Proc. IEEE 11th Int'l Conf. Data Mining (ICDM '11), pp. 407-416, 2011.

[3].G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005

[4]. H. J. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," *Inf. Sci.*, vol. 178, no. 1, pp. 37–51, Jan. 2008.

[5]. D. Agarwal and B. Chen, "fLDA: Matrix Factorization through Latent Dirichlet Allocation," Proc. Third ACM Int'l Conf. Web Searchand Data Mining (WSDM '10), pp. 91-100, 2010.

[6]. A. Mccallum, X. Wang, and A. Corrada-Emmanuel, "Topic and Role Discovery in Social Networks with Experiments on

Enron and Academic Email," J. Artificial Intelligence Research, vol. 30,pp. 249-272, 2007.
[7]. Q. Liu, E. Chen, H. Xiong, C. Ding, and J. Chen, "Enhancing Collaborative Filtering by User Interests Expansion via Personalized Ranking," IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 42, no. 1, pp. 218-233, Feb. 2012.
[8]. R. Pan et al., "One-Class Collaborative Filtering," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08), pp. 502-511, 2008.
[9]. P. Lops, M. Gemmis, and G. Semeraro, "Content-Based Recommender Systems: State of the Art and Trends," Recommender Systems Handbook, chapter 3, pp. 73-105, 2010.
[10]. D. Agarwal and B. Chen, "fLDA: Matrix Factorization through Latent Dirichlet Allocation," Proc. Third ACM Int'l Conf. Web Search and Data Mining (WSDM '10), pp. 91-100, 2010.
[11]. J. Herlocker, J. Konstan, L. Terveen, and J. Riedl, "Evaluating Collaborative Filtering Recommender Systems," ACM Trans. Information Systems, vol. 22, no. 1, pp. 5-53, 2004.
[12]. C. Wang and D. Blei, "Collaborative Topic Modeling for Recommending Scientific Articles," Proc. ACM 17th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 448-456, 2011.
[13]. X. Wu, J. Li, and S. Neo, "Personalized Multimedia Web Summarizer for Tourist," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 1025-1026, 2008.
[14] .J. Wu, H. Xiong, and J. Chen, "Adapting the Right Measures for KMeans Clustering," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 877-886, 2009
[15]. G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6, pp. 734-749, June 2005.
[16]. D. Agarwal and B. Chen, "fLDA: Matrix Factorization through Latent Dirichlet Allocation," Proc. Third ACM Int'l Conf. Web Search and Data Mining (WSDM '10), pp. 91-100, 2010.
[17] T.L. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proc. Nat'l Academy of Sciences USA*, vol. 101, pp. 5228-5235, 2004.
[18] Q. Hao et al., "Equip Tourists with Knowledge Mined from Travelogues," *Proc. 19th Int'l Conf. World Wide Web (WWW '10)*, pp. 401-410, 2010.
[19] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl, "Evaluating Collaborative Filtering Recommender Systems," *ACM Trans. Information Systems*, vol. 22, no. 1, pp. 5-53, 2004.
[20] A. Jameson and B. Smyth, "Recommendation to Groups," *The Adaptive Web*, vol. 4321, pp. 596-627, 2007.
[21] Y. Koren, "Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model," *Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08)*, pp. 426-434, 2008.
[22] P. Lops, M. Gemmis, and G. Semeraro, "Content-Based Recommender Systems: State of the Art and Trends," *Recommender Systems Handbook*, chapter 3, pp. 73-105, 2010.
[23] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. Berkeley Symp. Math. Statistics and Probability (BSMSP)*, vol. 1, pp. 281-297, 1967.
[24] A. Mccallum, X. Wang, and A. Corrada-Emmanuel, "Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email," *J. Artificial Intelligence Research*, vol. 30, pp. 249-272, 2007.
[25] R. Pan et al., "One-Class Collaborative Filtering," *Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08)*, pp. 502-511, 2008.
[26] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proc. ACM Conf. Computer Supported Cooperative Work (CSCW '94)*, pp. 175-186, 1994.
[27] F. Ricci, D. Cavada, N. Mirzadeh, and N. Venturini, "Case-Based Travel Recommendations," *Destination Recommendation Systems: Behavioural Foundations and Applications*, chapter 6, pp. 67-93, 2006.
[28] F. Ricci et al., "DieToRecs: A Case-Based Travel Advisory System," *Destination Recommendation Systems: Behavioural Foundations and Applications*, chapter 14, pp. 227-239, 2006.
[29] F. Ricci and Q. Nguyen, "Mobyrek: A Conversational Recommender System for On-the-Move Travelers," *Destination Recommendation Systems: Behavioural Foundations and Applications*, chapter 17, pp. 281-294, 2006.