ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(8), DOI: 10.17485/ijst/2016/v9i8/87846, February 2016* ISSN (Online) : 0974-5645

The Personalized Disease Predication Care from Harm using Big Data Analytics in Healthcare

J. Senthil Kumar[1*] and S. Appavu[2]

[1]Department of Information Technology, Bharath Niketan Engineering College, Aundipatti, Theni – 625531, Tamil Nadu, India; ssasenthils@gmail.com
[2]Department of Information Technology, K.L.N. College of Information Technology, Sivagangai, Madurai - 630612, Tamil Nadu, India; app_s@yahoo.com

Abstract

Background/Objectives: Nowadays, big data plays an important role in various areas such as industries, research, education, hospitals and etc., healthcare has its vitality in medical streams. **Methods/Statistical Analysis**: Healthcare is a data-rich industry. Executive databases embrace an incredible number of transactions for each patient treated. Though the healthcare industry has been a meadow, this change has the probable to be revolutionarily. It provides medical solutions for the different kinds of diseases. The manually maintained records are electronically stored in the database. **Findings:** A specialized tool disease recommendation system is used for entering personalised model health profile of the victims. This tool stumbles on entering large number of data and health profiles. It also increases the computational time, so this function in a timeframe for clinical use. **Improvements/Applications:** This paper begins by analyzing the performance limitation for personalized disease prediction contraption CARE (Collaborative Assessment and Recommendation Engine). CARE is analysis in two categories, they are Current CARE architecture and Parallel CARE architecture for performance benefits on big patient data.

Keywords: Big Data, CARE, Prediction Engine

1. Introduction

The healthcare industry has traditionally generated outsized amounts of data, driven by keeping record, acquiescence, dogmatic requirements and patient care. While most of the data are stored in hard replica form, the present trend express digitization of these huge amounts of data. Determined by the obligatory requirements and for the prospective to improve the worth of healthcare rescue by dipping the costs, these enormous quantities of data promise to support wide range of healthcare and medical functions. Those data are known as big data[1]. The promise is obtained formerly for the untouched intelligence and insights from data to address several recent and essential questions. Inside the health segment, it provides stakeholders with original insights that have the approach to advance personalized care recover patient outcomes and shun superfluous costs[2].

By classification, big data in healthcare transfer to electronic health data sets so huge and multifaceted that they are complicated (or impracticable) to manage with conventional software and/or hardware; nor can they be effortlessly managed with traditional or common data supervision tools and methods. Big data in healthcare is devastating not only because of its volume but also because of the assortment of data types and the swiftness at which it is to be managed[3].

Medical research has taken set for decades. It has provided what we as a people feel are some of the greatest modern accomplishments, from the discovery of bacteria and viruses to the increase of antibiotics. Nowadays, as the healthcare production begins its transition into the digital age it is easy to make out the happening as mere comings of age, merely the alteration of the medical communicates paper records into electronic form that is database. However, it provides so to a large extent more. This conversion has

* *Author for correspondence*

laid the establishment for another essential progression in the field of healthcare, the progression from preventative care into personalized treatment strategy[4]. It has been well recognized that early detection and treatment of many diseases is unswervingly simultaneous with improved health outcomes for the patient.

As a result, ordinary so called good health care programs have been implementing by numerous companies and care providers in order to encourage pre-emptive testing for certain conditions[5]. However, as the identification and treatment of these diseases are performed in the behaviour for multiple persons based chiefly on their current health circumstances, i.e. age, gender, race, prior lab results, etc. this form of care cascade closer to defensive medicine than personalized care[6]. While preceding studies have guided deterrent medicine treatment strategies by provided that historical probabilistic models based on the conclusions of patients who urbanized similar conditions, new predictive techniques can facilitate create personalized models of a patient's expectations health risks adapted to the individual's health data about persons[7].

In order to generate this personalized replica, data mining techniques have been useful to population-level health data cumulative beginning electronic healthcare records (EMR). While benchmark data mining such as clustering, decision trees and cohort analysis produced buoyant results, there was unfortunately a problem[8].

As with document records, all additional medical encounter by a patient resulted in supplementary data added to their electronic health record, and the extent of data soon exceeded the capability of benchmark data processing techniques. In response, original data dispensation techniques and architectures are individual created, such as Yahoo's Hadoop, Google's MapReduce, etc.,[9]

These techniques utilize the concepts of task segmentation, parallel and distributed computing in instruct to assuage some of the computational load from a solitary machine, along with allow for drastically improved runtimes for parallelizable tasks[10]. Due to the time critical nature of medical circumstances, the utility of a few model created is directly proportional to the time betrothed to create it. As such we must focus on preparation time of a model instruct to allow personalized healthcare models to be perverted within a beneficial timeframe[11]. Amid the most worth mentioning examples from promising Electronic Medical Records (EMR) based technology which used by database, is the disease prediction replica. These replicas exploit a patient's personal healthcare data in instruct to status the likelihood of the personage obtaining specific diseases[12].

One such scheme came commencement the University of Notre Dame in the appearance of a disease prediction technique called CARE. In this paper architecture are mainly analysed on parallel architectures. The CARE architecture in its current state is tremendously accurate, with an implementation already being qualified for clinical use[13]. However regardless of CARE's effectiveness, one of the architecture's foundational features, the capability to instruct hazard models from people level healthcare records, has the probable to become one of the greatest performance weaknesses. The CARE architecture utilizes massive amounts of individual healthcare encounters in order to erect a detailed correspondence model for a detailed personage and collaborative filtering is intrinsically a computationally architecture.

These specifics mutual with the ever-increasing amount of Electronic Medical Records (EMR) encounter data current in hospital databases produce a foremost operational concern. This paper will focus on the primary issue of CARE's resembling parallel and distrusted usability in a clinical location. It will instigate by identifying the limitations of the current CARE architecture[14], and aspire to provide a deposit of optimal performance parameters. Next some of the accuracy limitations will be deal with through the conception of a single patient adaptation of CARE.

Finally, this work will demonstrate a parallel distributed movement's implementation of the CARE architecture. This performance will tackle issues with together execution moment and disease cure while attempting to provide near industry level performance.

2. Related Work

The CARE architecture was the initial of its kind, receiving data from the hospital. However, to date numerous former diseases suggestion systems have been created. While these systems utilize many changed machine learning and data mining techniques command to construct their recommendations, each still potentially suffers on the confidence of elevated volume datasets. Classically these systems are reduced into two major categories, assembly use of a patient's phenotypic silhouette, or their therapeutic disease and family recitation as the preparation position of disease occurrences. Amid the

generally widely known is the system HARM. Similarly, to CARE, HARM is a personalized disease counsels system, but with collaborative filtering HARM utilizes a more composite mathematical replica based on association rules. Though as with CARE, and numerous erstwhile systems mentioned beyond, the authors of HARM do not converse the probable for parallelization or distributed work out in their paper[15].

Conversely, already it has been well recognized that distributed computing can afford significant development in runtime for computationally steep systems. Collaborative filtering techniques resembling those CARE have been used rarely in online artefact inference arrangement.

Conversely, their purpose to disease prediction is relatively new. This preparation has been fetch about by a fundamental transfer in how we assume about diseases. Recently there has been a focal point on modelling diseases as a complex rather than secluded instances, allowing for the utilization of numerous networking-modelling techniques. However, healthcare in succession is mostly clandestine, and the difficulties associated with of bring and covering huge scale healthcare data platforms have been a quantity of the major obstacles preventing procedure such as these starting widespread adoption.

There exists some earlier work evaluating seclusion when using collaborative filtering techniques on circulated data sets, for instance the work is done[16]. This paper details the apprehension of passing in the region of sensitive in sequence just to achieve calculations on the data. Nevertheless, in his performance, Berkvosky details a technique for subset data range in order to exceed a negligible amount of identifiable information to the classification. The clarification planned in our paper aspire to collect the scheme one step further, and rather than distribute a nominal data subset for working out, distribute the working out to each data site. Further, this paper also aims to address seclusion concerns by transmitting merely the effect of calculations larger than the network. The architecture provided in this paper is added similar to the effort described in division Map Reduce problem, everywhere the data are summarized at every nodule and then these synopsis results are returned to the supplicant.

Additional work associated to the concept of personalized distributed data is described[17] Lathia details a method for creating a custom parallel grade based on haphazard instances to guard the privacy of data. This likewise data could then be passed about exclusive of fear of instructive personalized information.

3. Proposed Work

3.1 Current CARE Architecture

The current CARE architecture shown in Figure 1 and is reasonably basic. The crucial steps for the algorithm are exhaustive below.

Current CARE begins with character presenting a deposit of diseases. This set is the accumulation of diseases larger than their personal medical narration.

The individual's disease correspondence is then compared to all erstwhile patients in the provider's existing record and a primary filtering is done.

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} i \times j = \{\text{matches i and j: no problems}\}\ i \neq j \text{ not matched}$$

Then, this filtering partitions the total dataset to embrace only person's patients with whom the current test patient has a quantity of disease resemblance, as collaborative filtering will acquiesce no promote between two persons who do not have several diseases in familiar.

Collaborative filtering is then the stage on this filtered dataset.

Finally, a probabilistic ranking of diseases for the behaviour is returned.

3.2 Data

The data used in this scrutiny is the similar dataset developed within the CARE research. The information consists of anonymous Medicare declare records collected by the hospitalities. There are approximately 8 hundred individual patients, accounting for just over 100 hospitals appointment, and include a total of 900 unique disease codes. Each record represents a solitary hospital visit and is comprised of a patient ID.

Quantity and equipped 10 personality diagnoses from the shatter. The analysis codes are distinct by ICD-9-CM, published by the World Health Organization (WHO)[18]. Through the ICD-9-CM code each disease is given an inimitable code, which can be up to 5 typeset long. These codes may embrace information of the circumstance, for instance the anatomical location.

Conversely, these fine-grained details are not required for the CARE architecture, and as a product the 5 digit analysis codes can be malformed to a 3-digit

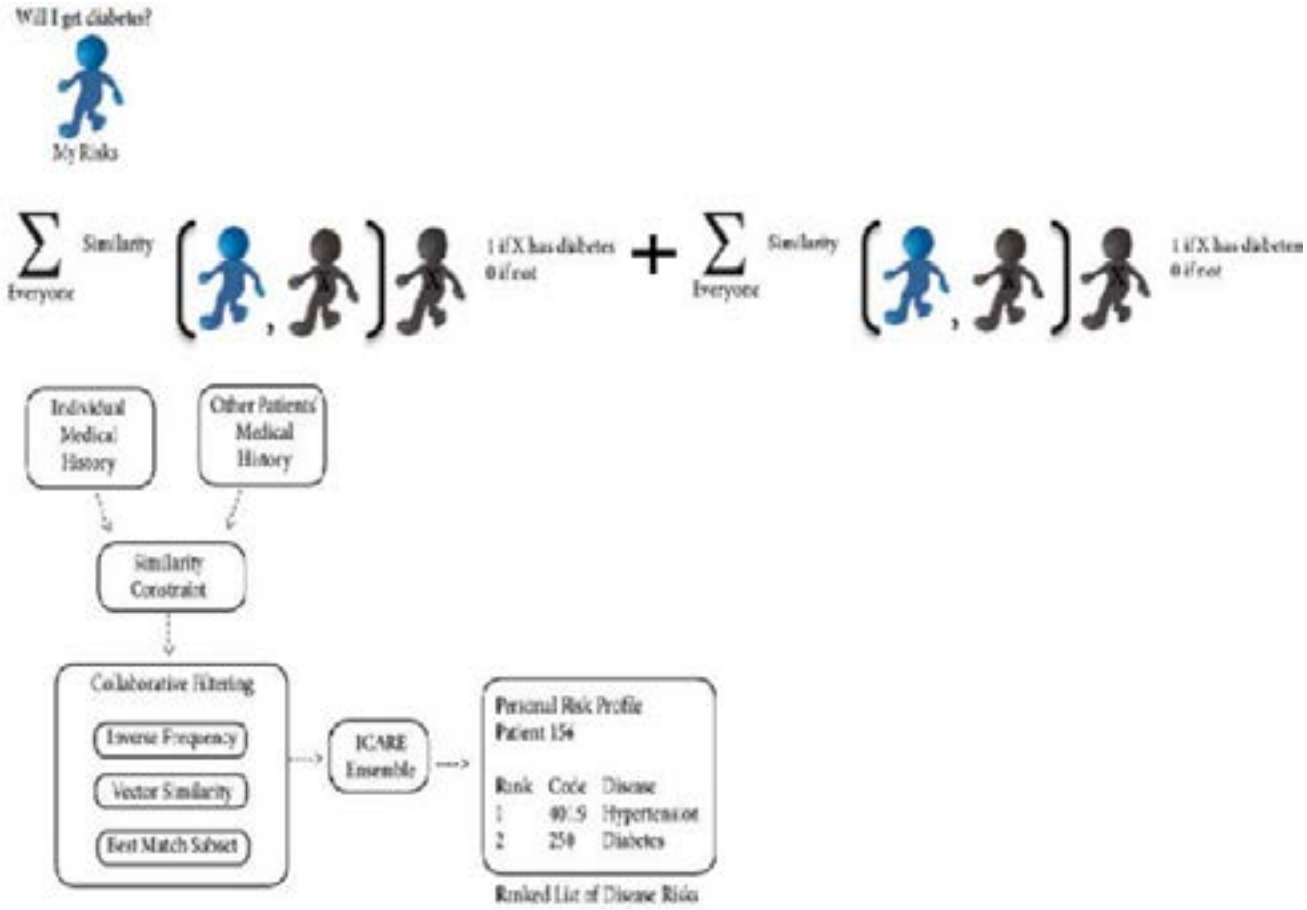

Figure 1. Current CARE Implementation.

generalization of the diseases. For model codes 461.0 and 461.1 can be collapsed into the generic analysis code 461. The accuracy of this oversimplification is recognized within the CARE paper, and by itself will be used departure forward in this work as well[19].

It is important to note that a disease may be analysed to a personality numerous times during their medical course. Conversely, as multiple diseases are not constructive when comparing patient's disease sets, only single diseases are essential for recommendations. Figure 1 shows that the average number of unique diseases converges to approximately 7 per patient larger than the full dataset[20]. This significance will be used whilst identifying outliers from the erratically particular patients, helping to diminish the bias between datasets and execution time.

3.3 Parallel CARE Architecture Implementation

In creating a single user implementation of CARE it shows that there is a possibility to improve the execution time, the computation desires of collaborative filtering boundary the maximal performance gains that can be achieved by the current CARE architecture. After evaluating both the current implementation and improved single patient CARE architectures, it is clear that the fundamental CARE architecture need to be changed to obtain any further performance improvements [20]. In order to understand how the CARE architecture could be a benefit from optimizations, such as parallel execution, it was first important to understand where internally did CARE stall.

This paper has previously shown that current CARE is CPU bound, moving further it is important to define where this occurs. In order to answer this CPU, bound

components were broken down into the individual functions as a percent of total runtime (Table 1). The CPU bounding is dominated by one function, Best Match.

Table 1. Function Breakdown of Parallel CARE Execution

Total	Best Match	Load Patient	Load Disease
5	55.28	16.56	17.55
10	88.74	6.99	5.23
25	97.89	1.22	0.67
50	99.54	0.52	0.24
100	99.62	0.32	0.04
250	99.88	0.15	0.01

Table 2. Component Breakdown of Best Match Function

	Percent of Time Per Function		
Total	Vector Similarity	Merge Visits	System Calls
25	31.78	9.32	27.90
50	21.69	8.05	39.26
100	23.24	7.22	38.54
250	32.13	6.52	49.36
Average	27.21	7.778	38.77
SD	5.22	2.79	1.22

Taking this further the Best Match function was broken down to analyze exactly what was causing the bottleneck (Table 2). Note that due to the short execution time and sparse nature of the disease classification, analyzing datasets controlling less than 10 patients creates highly variable and non-convergent results. Thus all datasets below 10 patients be barred from this evaluation.

According to the table, the percent of time spent in each component of the function remains unchanged as a product of number of patients in the dataset. This result lends itself well to the potential benefits of parallelization as it shows that even though the architecture has an exponential runtime the amount of time spent in each function is stable.

4. Conclusion

To recap, this paper has shown the performance limitations of the current CARE architecture. While some claim that an overnight batch execution is sufficient, as it can process a large patient dataset with a high degree of accuracy, this method is non-viable for medical usage. Big data providers such as Face book utilize similar batch events to help with data processing, but the information generated does not have the safety-critical nature of healthcare data. In the event that a disease is incorrectly recorded, a patient may have to wait up to 24 hours to receive updated disease risks. This turnaround time may be unacceptable, especially for time critical units.

In order to solve the issue of computation time this paper has outlined two distinct methods. First is the single patient version of Current CARE architecture, which can be utilized to perform disease risk rankings on-demand with a fairly high degree of accuracy. This method is intended to be utilized in the case above where updated rankings must be regenerated due to error, or for a new patient who was not present in the database when the last batch job was run.

The second method is a Parallel CARE Implementation of the CARE architecture with this current care implementation. This implementation can be used to generate on- demand rankings for a single patient with a high degree of accuracy, or executed as a nightly batch job on significantly larger patient sets for large practices or hospitals.

In future, CARE architecture can be used in analysis of Distributed CARE and also for finding evolution metric extension to maintain patient's database with fewer data.

5. References

1. AbuKhousa E, Campbell P. Predictive data mining to support clinical decisions: an overview of heart disease prediction systems. 2012 international conference on Innovations in information technology (IIT); Abu Dhabi; 2012. p. 267–72.
2. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data- mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of clinical epidemiology. 2013; 66(4):398–407.
3. Sasirekha D, Punitha A. A comprehensive analysis on associative classification in medical datasets. Indian Journal of Science and Technology. 2015; 8(33):1–9. doi: 10.17485/ijst/2015/v8i33/80081.
4. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics. 2008; 77(2):81–97.
5. Bui P, Rajan D, Abdul-Wahid B, Izaguirre J, Thain D. Work queue+ python: A framework for scalable scientific ensemble applications. Workshop on python for high perfor-

mance and scientific computing at sc11; 2011.
6. Kickbusch I, Payne L. Twenty –first century health promotion: the public health revolution meets the wellness revolution. Health Promotion International. 2003; 18(4):275–78.
7. Kim KW, Park WJ, Park ST. A study on plan to improve illegal parking using big data. Indian Journal of Science and Technology. 2015 Sep; 8(21):1–5. doi: 10.17485/ijst/2015/v8i21/78274.
8. Goil S, Choudhary A. High performance OLAP and data mining on parallel computers. Data Mining and Knowledge Discovery. 1997; 1(4):391–17.
9. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics. 2012; 13(6):395–405.
10. McCormick TH, Rudin C, Madigan D. Bayesian hierarchical rule modeling for predicting medical conditions. The Annals of Applied Statistics. 2012; 6(2):652–68.
11. Feldman K, Chawla NV. Scaling personalized healthcare with big data. 2nd International Conference on Big Data and Analytics in Healthcare; Singapore; 2014. p. 1–14.
12. Kumar SJ, Ramprasath N. A scrutiny on current and parallel big data analytics in health care. International Journal of Emerging Technology in Computer Science & Electronics. 2015; 12(4):244–47.
13. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. Journal of General Internal Medicine. 2013; 28(3):660–65.
14. Dhamodaran S, Sachin KR, Kumar R. Big data implementation of natural disaster monitoring and alerting system in real time social network using Hadoop technology. Indian Journal of Science and Technology. 2015 Sep; 8(22):1–4. doi no:10.17485/ijst/2015/v8i22/79102.
15. Davis DA, Chawla NV, Christakis NA, Barabasi A-L. Time to care: a col- laborative engine for practical disease prediction. Data Mining and Knowledge Discovery. 2010; 20(3):388–15.
16. Berkovsky S, Eytani Y, Kuflik T, Ricci F. Enhancing privacy and preserving accuracy of a distributed collaborative filtering. Proceedings of the 2007 ACM conference on recommender systems. 2007. p. 9–16.
17. Lathia N, Hailes S, Capra L. Private distributed collaborative filtering using estimated concordance measures. Proceedings of the 2007 ACM conference on recommender systems; 2007. p. 1–8.
18. Dean J, Ghemawat S. Mapreduce: simplified data processing on large clusters. Communications of the ACM. 2008; 51(1):107–13.
19. Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B, Hartwell L. The case for early detection. Nature Reviews Cancer. 2003; 3(4):243–52.
20. Lard LR, Visser H, Speyer I, vander Horst-Bruinsma IE, Zwinderman AH, Breedveld FC, Hazes JM. Early versus delayed treatment in patients with recent-onset rheumatoid arthritis: comparison of two cohorts who received different treatment strategies. The American Journal of Medicine. 2001; 111(6):446–51.