Collaborative Space Analysis for Doctor-Patient Workflow for Hypertension

Nihar Patil
Department of Computer Science
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14623
nrp9155@cs.rit.edu

Abstract—The communication between doctor and patient is crucial nowadays. Doctor has very less time to devote to its patient. This may create serious issues in health of the patient. To improve their communication, we proposed a smart platform, which uses data mining to reduce the cost of medication, with the better time utilization of doctor. Our system co-ordinates between doctor and patient and continuous monitors their lifestyle. Based on their feedback, we recompute the clusters. These clusters help doctors to have consolidated view of a patient. Due to our platform, both doctor and patient always have similar information about patient's health. We have compared the cluster formed with the state of the art K-means algorithm.

I. INTRODUCTION

Technology advancement in healthcare impacts our lifestyle significantly. Many wearable devices help us to track and improve our health. Makers of healthcare wearable devices provide insights to better ones life. Using data mining techniques on historical information available from such devices, we can predict future valuable decisions. Collaborative analysis, multiple systems working together to build a product or a service, can be considered as an improvement in the workflow. Amongst the different ways to improve overall workflow, optimizing the smaller workflows first, helps in boosting the efficiency drastically. These collaboration improvements will help to provide better health treatment to patients, efficient utilization of doctors time and reduce the cost of healthcare.

Due to increasing health awareness among patients, patient-centered healthcare software applications have become prevalent. However, these applications are solely based on users initiative and self-awareness, with no involvement of a doctor. To improve the communication between doctors and patients, the collaborative analysis will play a crucial role in a variety of contexts. To facilitate this gap, we need to focus on software application development which involves doctors and is controlled by doctors. To demonstrate the effectiveness of collaboration, we choose to analyze patient's hypertension records as the initial step.

The Goal is to develop a collaborative system which helps the doctor to provide better medications. Fig 1 shows an example of collaborative space analysis proposed by Eikey, E., et al. [6]. It can be achieved by providing a smart platform to facilitate Doctor-Patient information exchange about Hypertension data. This data is collected using a personal micro-sensor device. Our approach is to cluster the patients hypertension data using state of the art partitioning clustering algorithm K-means [1]. Though, K-means algorithm is simple and has linear complexity, it can converge to local minima since initial selection of cluster centroids is random. Moreover, it cannot be used for categorical data. As system scales, we are expecting to have categorical variables. With taking this into consideration, another algorithm ROCK (RObust Clustering using linKs) proposed by Sudipto Guha et al. [2], will be more suitable and robust. This uses similarity links, which could be nonmetric, which helps to correlate categorical data points. In case of K-means, Euclidean distance cannot be calculated for categorical attributes. Performing comparative study of these two algorithms will help us to form better clusters. These clusters help the doctor to view consolidated information of the patients belonging to the same cluster. Patients cluster changes with their lifestyle. With this information, the doctor could advise the patient to change his/her lifestyle to control their hypertension based on an ideal cluster already formed. This improves the communication between the doctor and patients. A system can also generate automated emergency alerts in case of critical hypertension readings. After clustering data, a classifier can be built to analyze the patterns of clusters to predict the hypertension class variable. Clusters formed can help to build a classifier, which can predict the future possibility of having hypertension condition. To test the effectiveness of the system, the doctors and patients will use the system and monitor the formulated clusters.

II. RELATED WORK

There are several systems has been developed to make efficient use of EHR. CoMed system for doctors, allows to
share patient records and to communicate with each other on the Internet[2]. It is only a web-based solution, which provides flexibility, extensibility and location transparency of patient database. It uses O2 dbms, CORBA servers and JAVA applets for User interface. This system is informative and helpful in medication, but it does not involve patients in its process.

Since technology advances, smart-phone based applications are evolved. It uses a three-tier clinical information systems design model[3]. It is more generic system. This is more flexible, maintainable and reusable compare to CoMed, since it uses service oriented architecture(SOA). Using SOA, we can create small micro-services for each small module. These micro-services can be extended to desktop or mobile applications easily. These applications are mainly patient-centric. Involvement of doctors is negligible.

### III. Algorithms

#### A. K Means

The aim of K-means algorithm is to divide M points in N dimensions into K clusters so that the within cluster sum of squared error is minimized. This algorithm produces K centroids which help in segregating a dataset to meaningful groups or categories. The process also helps in assigning a new incoming datapoint to an unsupervised cluster. The Euclidean distance is calculated from each point to the randomly selected initial clusters. This will provide the base values for comparing clusters. The advantages of using this algorithm is that it is simple to implement. Thus, decreasing the overhead of computations. K-means algorithm is one of the state-of-the-art algorithms, and it forms a baseline for the clusters formed and also for the intra and inter-cluster distance.

However due to the incessant computations of centroids, the algorithm may fixate in a local minimum. Centroids also get distorted if there are outliers present in the dataset.

The challenges in our real-time dataset is that new datapoints may be added for forming clusters as per doctors criterion. This may be lead to more outliers in the dataset and subsequently wrong clusters might be formed. To overcome with the scalability issue, we have proposed ROCK algorithm.

#### B. ROCK (Robust Clustering using Links)

1) **Neighbors**: A datapoints neighbor are those points that are present nearest to the current datapoint. A datapoints neighbor is calculated using Euclidean distance. ROCK considers more than one neighbor while deciding cluster of a data point. The lesser the Euclidean distance, closer is the datapoint to the current datapoint.

2) **Link**: Clustering points based on closeness is not strong enough to distinguish two not so well separated clusters. For example, if there are two datapoints who have more than a threshold level of neighbors This is due to the fact that it is possible that two neighbors can be part of different clusters. To resolve this issue, if a datapoint and its neighbor are considered in a cluster only if they share large number of neighbors. Link is defined as number of common neighbors between the data point and its neighbor. This adopts the global approach for clustering and avoid local maxima occurred in case of means.

ROCK uses multiple neighbor while computing the cluster centroid, which is unlike the K-Means approach. It also considers links between neighbors. If a datapoint does not share a common neighbor, then it could be considered as an outlier. In short, ROCK overcomes all the shortcomings of K-means. Although the complexity of computing neighbors and links is expensive, it is traded off by the increase in accuracy of cluster formulation.

### IV. System Components

Fig 2 shows the System’s work-flow. Initially user registers on the platform. Users enter their profile. The system will cluster the patient based on already formed clusters. It also monitors blood pressure level of the patient with each reading and generates an alert to either doctor or patient or both. Depending on alert type. System will also suggest the changes to the patient, which will help them to improve their hypertension condition. Based on this, we recompute the clusters. Doctor will get notified.

#### A. Data Collection

1) Developed a user interface to collect users lifestyle feed
2) Developed a secured data gathering from a blood pressure device associated with the patient.

#### B. Data Processing / Clustering

1) Cluster data using ROCK algorithm
2) Cluster the dataset every day based on the users lifestyle attribute.
3) Clusters will be recalculated based on the updated users lifestyle attribute.
4) Monitor the changes in the patients cluster.
5) Notify to appropriate user based on changes.

#### C. Alert Generation

1) Alert Types: Normal: <<Patient>> Congratulations! Your blood pressure reading indicates a normal range, continue healthy lifestyle to keep blood pressure normal. <<Doctor, no Alert needed>> Pre-hypertension: <<Patient>> Your blood pressure reading indicates a pre-Hypertension range, you should take another reading, you have an increased risk of future hypertension. Lifestyle modifications and regularly monitor blood pressure should be considered. <<Doctor, no Alert needed>> Hypertension 0: <<Patient>> Recheck your blood pressure to make sure the reading is correct, if your BP readings continue in the hypertension range you should contact your Doctor for further treatment.
2) Hypertension 1: <<Patient>> Recheck your blood pressure to make sure the reading is correct, if your BP readings remain high (Stage 1 hypertension) you should contact your Doctor as soon as possible for further treatment. <<Doctor>> Patient Stage 1 Hypertension Alert - calendar appointment

3) Hypertension 2: <<Patient>> Recheck your blood pressure to make sure the reading is correct, if your BP readings remain high (Stage 2 hypertension), contact your Doctor immediately to explain the situation and make an appointment as soon as possible. <<Doctor>> Patient Stage 2 Hypertension Alert - calendar appointment

4) Emergency: <<Patient>> Recheck your blood pressure immediately to make sure the reading is correct. If your BP readings remain high, immediately go to urgent care, emergency dept., or call 911, please do not drive. <<Doctor>> Urgent Patient Alert - Stage 3

5) Hypertension-pager/phone message >> Depending on the blood pressure reading, alert gets associated with it. And if alert is greater than some threshold then it sends alert to doctor and patient with the precautionary measures.

D. Lifestyle suggestion feedback

Patients dashboard is fed with information based on their lifestyle. These suggestions, if followed, help the patient improve their lifestyle. Thereby changing their cluster. For e.g. if the patient reduces smoking amount and it is registered for a week, then we automatically update the smoking attribute for the patient and recompute the cluster. This change in cluster is notified to the doctor and patient.

E. Security

1) Health care data must be compliance with HIPAA standards.
2) Security has involved at multiple levels.
3) User Authentication and authorization.
4) This involves basic user authentication. Data is encrypted using SHA256 encryption.
5) Sessions mentions for 30 mins. Session expires after 15 mins of inactivity.
6) Store patients profile and health data with HIPAA complaint storage.
7) Implementing HIPAA complaint is very challenging and a time-consuming task and out of the scope of this project. I have used Google Health Storage to store the data. It provides HIPAA complaint data storage.
8) Fetch data from micro devices securely.
9) User must provide access to his/her health information. To collect it securely, manufacturing vendors mostly uses two types of authentication namely OAuth1 and OAuth2. Based on different vendors, we choose appropriate authentication type and APIs and users approval.

F. Database Schema

1) Data has been classified into three levels.
2) Orange data: Publicly accessible. It includes data of life feed advises. This is stored in the sql database with the content.
3) Red Data: Authorized person accessible Also stored in SQL database.
4) Black Data: HIPAA complaint data Stored in Google HIPAA compliance data storage.
5) To fetch data from different data sources, developed a generic persistence framework.

G. Deployment: Application has two levels of deployment

Development stage - Development stage performs all the unit tests, code coverage test, functional tests. If all tests pass, then it builds project successfully. Jenkins is used for continuous deployment. Production Stage. The website can be accessed using http://ciashypertension.rit.edu:8080/healthMonitor/

V. COMPUTATIONS

To evaluate cluster quality, the most common approach is to calculate the sum of squared error distance. As it gives the variation in each cluster. The value with minimum SSE has best clusters. In the given formulae, x represents the data point which needs to be clustered. $m_i$ is centroid of cluster $C_i$ and k represents the number of clusters.

\[ \sum_{i=1}^{k} \sum_{x \in C_i} \text{dist}^2 (m_i, x) \]

VI. RESULTS AND DISCUSSION

In this section we will showcase how ROCK handles the outliers and also insertion of a new datapoint/attribute.

A. Cluster 0: Health conscious young adults

This cluster represents the age group of patients in their early 30s. Patients who smoke less frequently, having moderate drinking habits, normal salt intake, higher calorie consumption and higher amount of exercise. We can say that the patients in this cluster, are mostly young adult who are cautious about their health and with a casual interest in party. This deduction is based on the drinking habit which is on a moderate amount.
B. Cluster 1: Young Adults with moderate physical activities, drinking and smoking habits

Fig 3 cluster represents the age group of early 40s. Having little higher BMI which can also be reflected from their diet and salt intake. Smoking and drinking levels are moderate which adversely affects the BP. The increase of normal BP value pushes the record to a cluster containing values from 133 and 81 for Systolic and Diastolic respectively.

C. Cluster 2: Young Adults with no physical activity and high calorie intake

Fig 4 cluster represents the age group of late 30s. Patients in this cluster, is similar to cluster 0, but the blood pressure value is on higher side due high intake of salt and high calorie diet. BMI is also on higher side. This increases the average value of the cluster.

D. Cluster 3: Older Adults with heavy drinking habits

Fig 5 cluster represents the mid-50s patients. They have high BMI, moderate smoking, high drinking, normal salt and diet intake. The blood pressure is mainly affected due to age and lack of exercise in this cluster.

E. Cluster 4: Older adults with moderate physical activities, drinking and smoking habits

Fig 6 cluster represents the age group of late 60s. Patients have moderate BMI, with rare smoking and drinking habits. Normal diet and salt intake with occasional exercise. The blood pressure range is like cluster 3. But it is mostly affected due to age.

F. Scenarios of Addition of a synthetic Outlier

Both the algorithms formed similar clusters. This can be verified by checking its SSE values. To test how both the algorithms handle outliers, I have added an outlier profile. Profile has following attribute values:
Age 69
BMI- 23
Smoking 0
Drinking 0
Salt Intake 1.2
Diet 1.6 kcal/day
Exercise 60 mins
Diastolic 82
Systolic 130

This profile is an outlier because patient does regular exercise, maintains diet and salt habits. The patient does not smoke or drink. This profile should not be part of any of the above cluster as we dont have any cluster which matches this description. The patient is old and health conscious and such a cluster has not been formed at all.

Due to addition of this outlier, the value of cluster 3 has changed a lot in case of k-means but remained invariant in case of ROCK. Slight changes in values are possible due to the random nature of the algorithm. This proves ROCK handles Outlier effectively compared with K-means.

Fig 7 shows the values of SSE for 5 clusters. Initially SSE values were approximately same for both the algorithms. As we add outliers in the project, the SSE value in case of ROCK increases significantly. This proves the ROCK algorithm handles outliers effectively. It helps to cluster the health data.
VII. CONCLUSION

The data mining process helps to improve the communication between doctor and patient as it provides consolidated view to the doctor and the patient received periodic feedback from a doctor. Based on these feedback, doctor and patients have same health information, which saves time for the doctor. Also early medication helps patients to improve their lifestyle in early stages. ROCK algorithm helps to form robust cluster for health data. It can also be extended with additional lifestyle attributes.

VIII. ACKNOWLEDGMENT

I would like to thank Prof. Chris Bondy and Dr. Pengchen shi for guiding me on this project. I would also like to thank Omron and withings support team for their prompt assistant during utilizing their health APIs.

REFERENCES