Journal of Management Science and Business Intelligence, 2017, 2–2
Aug. 2017, pages 17-23
doi: 10.5281/zenodo.834918
http://www.ibii-us.org/Journals/JMSBI/
ISBN 2472-9264 (Online), 2472-9256 (Print)

Health Recommender System using Big data analytics

J.Archenaa[1] and Dr E.A.Mary Anita[2]

[1]Research Scholar,Department of Computer Science & IT,AMET University,Chennai-India, [2]S.A.Engineering College,Chennai

E mail: archulect@gmail.com

Received on June 07, 2017; revised on July 18, 2017; published on July 22, 2017

Abstract

This paper gives an insight on how to use big data analytics for developing effective health recommendation engine by analyzing multi structured healthcare data. Evidence-based medicine is a powerful tool to help minimize treatment variation and unexpected costs. Large amount of healthcare data such as Physician notes, medical history, medical prescription, lab and scan reports generated is useless until there is a proper method to process this data interactively in real-time. In this world filled with the latest technology, healthcare professionals feel more comfortable to utilize the social network to treat their patients effectively. To achieve this we need an effective framework which is capable of handling large amount of structured, unstructured patient data and live streaming data about the patients from their social network activities.

Apache Spark plays an effective role in making meaningful analysis on the large amount of healthcare data generated with the help of machine learning components and in-memory computations supported by spark. Healthcare recommendation engine can be developed to predict about the health condition by analyzing patient's life style, physical health factors, mental health factors and their social network activities.

Machine learning algorithms plays an essential role in providing patient centric treatments. Bayesian methods is becoming popular in medical research due its effectiveness in making better predictions.For example on training the model with the age of women and diabetes condition helps to predict the chances of getting diabetes for new women patients without detailed diagnosis.

Keywords: Predictive analytics, Recommendation Systems, Bayesian rules, Big Data, Machine learning algorithms

1. Introduction

In today's digital world people are prone to many health issues due to the sedentary life-style. The cost of medical treatments also keeps on increasing. It's the responsibility of the government to provide an effective health care system with minimized cost. This can be achieved by providing patient centric treatments. More cost spent on healthcare systems can be avoided by adopting big data analytics into practice [(1)]. It helps to prevent lot of money spent on ineffective drugs and medical procedures by making useful analysis on the large amount of complex data generated by the healthcare systems. There are also challenges imposed on the growing healthcare data. It's important to figure out how the big data analytics can be used in handling the large amount of multi structured healthcare data.

What is the need for predictive analytics in healthcare?

To improve the quality of healthcare, it's essential to use big data analytics in healthcare.

Data generated by the healthcare industry increases day by day. Big data analytics system with spark helps to perform predictive analytics on the patient data [(3)]. This helps to alarm the patient about the health risks earlier. It also supports physicians to provide effective treatments to their patients by monitoring the patient's health condition in real-time. Diagnosis can be improved by utilizing the expert recommendations from medical forums. Customized treatment can be achieved with the help of big data analytics, which helps in improving the quality of healthcare services. It also helps to alarm government about the seasonal disease that may occur in particular locality due to the change in weather condition.

2. Big Data Use Cases for Healthcare

Many organizations are figuring out how to harness big data and develop actionable insights for predicting health risks before it can occur. Spark is extremely fast in processing large amount of multi-structured healthcare data sets, as it offers ability to perform in-memory computations. This helps to process data 100 times faster than traditional map-reduce. Spark's support for lambda architecture allows to perform both batch and real time processing.

2.1 Data Integration from Multiple sources

Spark supports fog computing which deals with Internet of Things (IOT). It helps to collect data from different healthcare data sources such as Electronic Health Record(EHR), Wearable health devices such as Fitbit, user's medical data search pattern in social networks and health data which is already stored in HDFS(5). Data is collected from different sources and inadequate data can be removed by the filter transformation supported by spark.

2.2 High performance batch processing computation and Iterative processing

Spark is really fast in performing computations on large amount of healthcare data set. It is possible by the distributed in-memory computations performed as different clusters. Genomics researchers are now able to align chemical compounds to 300 million [2] DNA pairs within few hours using the Spark's Resilient Distributed Dataset (RDD) transformations (6). It can be processed iteratively then.

2.3 Predictive Analytics Using Spark Streaming

Spark streaming components such as MLib helps to perform predictive analytics on healthcare data using machine learning algorithm [10]. It helps to perform real-time analytics on data generated by wearable health devices. It generates data such as weight, BP, respiratory rate ECG and blood glucose levels. Analysis can be performed on these data using k-clustering algorithms. It will intimate any critical health condition before it could happen.

The below figure represent proposed big data healthcare ecosystem using Apache Spark and Hadoop.

Apache Spark's RDD based computations is extremely fast in processing large amount of data. Real time streaming data from social networking sites can be processed effectively. Mlib – Spark's in-built library supports machine learning which is essential for designing health recommender systems. Prediction and Recommendation component are built using machine learning algorithm.

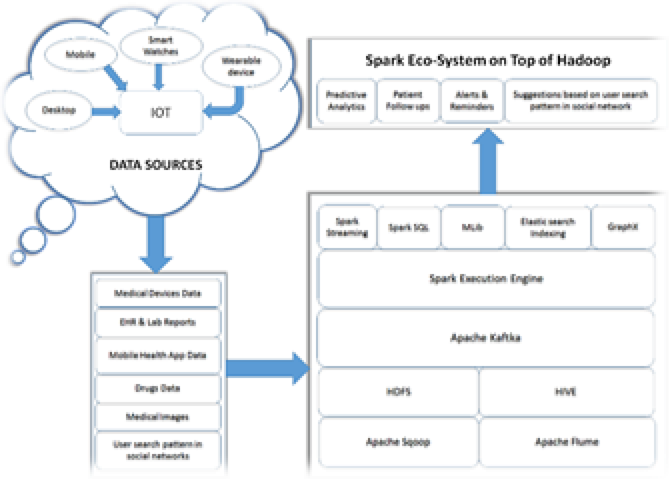

Fig. 1. Apache Spark Healthcare Ecosystem with Hadoop

3. Designing Healthcare Recommendation System

A health recommender system (HRS) suggests medical information which is meant to be highly relevant to the advancement in

Medical treatment associated with the patient history.

HRS provide physicians, staff, patients and other individuals with knowledge and patient-centric information, intelligently filtered and presented at appropriate times, to enhance quality of healthcare services. Common features of HRS systems includes providing patient oriented guidance such as clinical information, integrating data from various sources (such as lab reports, medication history, imaging, wearable sensors, social media sites, health forums) into CDS (Clinical diagnosis system) application and provide relevant recommendations such as list of diagnosis, drug interaction alerts, preventative care alerts, suggesting patient centric health insurance plans, sending alerts about the hospital transportation required, sending alerts to patients about follow-ups, diet recommendations, Refilling medicines etc.

Based on user's medical expertise an HRS should suggest medical information, which is relevant to that user. Depending on the expertise of a HRS user, at least two separate use cases can be defined as follows:

1. Use case A = Health professional as end-user

In this scenario an HRS is used by a health professional to retrieve relevant information for a certain case. For example, existing clinical diagnosis, Clinical pathway or research articles from health forums can be computed automatically. This form of case-related information enrichment might support a physician with the process of clinical diagnostics as latest research results can be used for treatment decision support. In addition, naive-friendly documents can also be retrieved for the purpose supplying high quality information to patients in order to cope with a certain disease or adapt his or her lifestyle habits.

2. Use case B = Patient as end-user

In this scenario a patient interacts with a HRS-enabled PHR without direct support by a physician. HRS computes user-friendly content according to the person's case history. The relevant items are recommended to the user. By selecting the highest ranking content a patient is empowered in terms of health information.

3.1 PHR Enhanced with HRS

A PHR (Patient health record) is an electronic application through which patients can access and share their health information in a private, secure and confidential environment [6]. System is useful only if it gives valuable insights from the user health history. PHR enhanced with health recommendation component provides relevant information to the users based on their needs. It will be valuable add on to the existing healthcare system which suggests personalized or case based health recommendations.

Health Recommender Systems can assist its users in various stages in the care process, from preventive care through diagnosis and treatment to monitoring and follow-up. The most common use of HRS is for addressing clinical needs, such as ensuring accurate diagnoses, screening in a timely manner for preventable diseases, suggesting appropriate health insurance plans, alternative medicines, drug dosage recommendations or alerting adverse drug events.

4. Healthcare Recommender System Framework

Healthcare recommender system is represented by prediction and recommendation. It depends on a set of patients' case history, expert rules and social media data to train and build a model that is able to predict and recommend disease risk, diagnosis and alternative medicines. Predictions and recommendations are approved by physicians. HRS system requires input information to generate predictions and recommendations. In this work diabetes data is used as case study.

Training Data

Pile of historical medical records of diabetic patients (935 records) has been collected from hospitals. The collected data records are represented by a number of attributes, values and doctors diagnosis for each case. Diagnosis scale ranges from 1 to 10 based on the severity of the disease, 5-represents critical condition, 4-represents severe requires immediate treatment, and 3-represents moderate requires further investigation, 2-represents normal, 1-represents within control.

Demographic data of active patient

It refers to the user's data such as: name, age, location, education level, wearable device, lifestyle, food habits and type of connectivity.

Medical Case History of Diabetes Patient:

It comprises home test details such as blood sugar, blood pressure, weight. Diagnosis data comprises of physician notes, lab results, and medications. Diabetes data set is collected from KN specialty clinic and also downloaded from UCI repository.

The output of the system is

Prediction and recommendation: prediction is expressed as a numerical value that represents the disease risk diagnosis for future cases based on active patients. Recommendation is expressed as the suggestion required by the users. For example non healthcare professional might be requiring alternative remedies for treating diabetes. Healthcare professionals may be looking for disease diagnosis methods based on patient similarity.

4.1 Building the Predictive Model

Data Preprocessing

Feature selection methods can be used to identify and remove irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model.

Data filtering is essential to avoid the creation of ambiguous or inappropriate models and improve the learning model performance. In our system, the diabetes dataset is filtered by determining the relevant features through InfoGainAttributeEval Attribute Selection method, furthermore, the data is also transformed to a form appropriate the classification.

4.2 Classification using Bayesian Network

Bayesian methods have become increasingly popular in medical research due its effectiveness in making better predictions. Diabetes is a chronic condition that occurs when the body cannot produce enough or cannot effectively use insulin [1].

Diabetes can mainly be of 3 types: Type-1 diabetes, Type-2 diabetes and Gestational diabetes. Type-1 diabetes results from non-production of insulin & Type-2 diabetes results from development of resistance of insulin, as a result of which the insulin produced is not able to metabolize the sugar levels properly. Bayesian classifier is used to predict diabetes accurately even with less amounts of training data.

Naïve Bayes is considered to be one of the most efficient and effective inductive learning algorithms for machine learning and data mining[5]. Bayesian statistics allow one to make an estimate about the likelihood of a claim and then update these estimates as new evidence becomes available.

In Bayes' probability of a hypothesis is obtained by multiplying the prior probability with the strength of the new data. The new, updated probability is called the posterior probability, or just 'the posterior'. This is the sum total of probabilities of all possible relevant hypotheses.

The posterior then becomes the new prior and the process may repeat. Let's consider how we can put Bayes' Theorem to practical use in everyday medical decision making.

1. For example 1 out of 1000 people die in diabetes it is known as the prior data.
   Prior Probability = 1/1000 = 0.001
2. Another test result indicates that there are 10% false positive result-indicates people who does not die due to diabetes 10%*1000 = 100 false positives
3. On an average, people 101 test positive for death due to diabetes out of 1000 people(1 true positive- 1 die in accident and 100 false positive- who have diabetes but does not die)
4. Therefore 1 dies due to diabetes out of 101 people.

Without the Bayesian perspective, these 101 people will likely all become convinced that they will die due to diabetes. Bayesian statistics allows to get clearer perspective about test results by combining prior knowledge with new data and updating our position.

Baye’s theorem is represented by the below formula:

$$P(H/D) = \frac{P(D/H)*P(H)}{[P(D/H)*P(H)] + [P(D/H_0)*P(H_0)]} \quad (1)$$

P(H/D) is the probability of the hypothesis (H) given the data (D), P(D/H) is the probability of the data (D) given the hypothesis (H),

P(H) is the probability of the Hypothesis prior to the new data (also called the "prior probability" or just the "prior"), and $P(H_0)$ is the null hypothesis.
Combining background knowledge and evidence derived from data and missing data can be handled both in the construction process and in using a Bayesian network model. Expert systems based on Bayesian networks have the advantages of a formal mathematical foundation, relative computational tractability, and a graphical representation for presentation to an expert.

4.3 Parameters used in Estimation

Dataset of 1000 cases was prepared by collecting the data randomly from different groups of the society with an aim to have a variety in the dataset. To maintain accuracy and to avoid errors, data was preprocessed carefully.

Attributes	Description	Values used
Age	Age of the user	Discrete Integer Values
Sex	Male or Female	Male or Female
BMI	Body Mass Index (Height to weight ratio)	Discrete Integer Values
Family History	Any family member of the subject is suffering/ was suffering from diabetes.	Yes or No
Smoking	Smoking habits of the user	Yes or No
Drinking	Drinking habits of the user	Yes or No
Lifestyle	Lifestyle of the user	Active, Moderate, Sedentary
Eating Habits	Food habits of the user	Healthy Foods, Junk foods
Frequent Urination	Urination habits of the user	Frequent or Normal
Increased Thirst	Urge to drink more than usual	Yes or No
Fatigue	Does the user feel fatigue often?	Yes or No
Blurred Vision	Do you have blurred vision?	Yes or No
Waist Size	Waist size of the user in inches	Discrete Integer Values
Gestational Diabetes	Do you have gestational diabetes?	Yes or No
Polycystic ovaries	Do you have polycystic ovaries?	Yes or No
Fasting Plasma Glucose	Values of Fasting Plasma Glucose	Discrete Integer Values
Casual Glucose Tolerance	Values of Random Glucose tolerance test	Discrete Integer Values

Expert Rules

Fasting Plasma Glucose and Casual Glucose Tolerance Test

No Diabetes Range

If you had a fasting plasma glucose test, a level between 70 and 100 mg/dL (3.9 and 5.6 mmol/L) is considered normal.
If you had a casual glucose tolerance test, a normal result depends on when you last ate. Most of the time, the blood glucose level will be below 125 mg/dL
If your HBA1C test values is below 97 mg/dL then its normal.

Pre-diabetes Range

If your Fasting Plasma Glucose test ranges between 100 mg/dl to 125 mg/dl
If your Casual Glucose test ranges between 140 mg/dl to 199 mg/dl
If your HBA1C test values ranges between 97-154 mg/dL

Diabetes Range
If your Fasting Plasma Glucose test is 126 mg/dl or higher
If your Casual Glucose test ranges is 200 mg/dl or higher
If your HBA1C test values is greater than 180 mg/dL

Table 1. Life Style Based Analytics-Diabetes Profiling

Features	Employee A	Employee B	Diabetes Ratio A to B

Age	40	40	1 to 1
Vehicle Type	Cycle	Mini Van	1 to 10
Fast Food	Rarely	Frequent	1 to 40
Hobbies	Active Outdoor	Reading	1 to 80

This table describes about lifestyle based diabetes risk.

5. Implementation

Data from various sources combined with powerful learning algorithms and domain knowledge led to meaningful insights. Supervised pattern classification is the task of training a model based on labeled training data which then can be used to assign a pre-defined class label to new objects. Naive Bayes classifiers *are linear classifiers based on Bayes theorem*. Based on the conditional independence the presence of features are independent of each other [12]. Individual probability for all the features are calculated and classified into 3 classes: Diabetic, Pre-diabetic and No diabetic.
Individual probability is computed for all the features to classify the case as diabetic, pre-diabetic and no diabetes. **P(Diabetic='Yes') given** "Casual Glucose Tolerance Test" = 'Value from Test Data' and "Increased Blurred Vision"='Value from Test Data' .**P(Diabetic='No') given** "Casual Glucose Tolerance Test"= 'Value from Test Data' and "Increased Blurred Vision"='Value from Test Data'. Similarly the probabilities of all the features are calculated. To deal with the condition of zero probability values for unknown features Laplace smoothing is used. By calculating individual probability values, the test data gets classified into one of the three categories-Pre-Diabetic, Diabetic or Not Diabetic. The development of the system is done using Apache Spark and Python.

Table 2. Diabetes Classification using Naïve Bayes Algorithm

Class	Entries	Percentage of Persons
Pre-Diabetic	288	32
Diabetes	225	25
No Diabetes	387	43

This table describes about the results obtained after applying Bayesian network

5.1 Confusion Matrix

Confusion matrix gives a complete picture of how your classifier is performing [10]. It helps to get a picture of what your classification model is getting right and what types of errors it is making. Classification accuracy is the ratio of correct predictions to total predictions made.

Accuracy is calculated as = $(TP+TN) / (P+N)$ (2)
$P = TP + FN$ $N = TN + FP$

True positive (TP) - are the positive data set that were correctly labeled by the classifier. If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP)[5].
True negative (TN) – are the data set predicted correctly for no diabetes. It is represented by TN.
False positive – are the data set predicted incorrectly for having diabetes. It is represented by FP.
False negative – are the data set predicted incorrectly for no diabetes. It is represented by FN.

Table 3. Number of Health Records

Number of Training Records	650
Number of Test Records	250

TP=100 FN=34 TN=70 FP=46
Accuracy= 170/134+116 = 0.68

5.2 Building the Recommendation Model

Hybrid Recommender System
Medical expert systems are a branch of artificial intelligence that applies reasoning methods and domain specific knowledge to suggest recommendations like human experts [6]. To enable reliable and fast decision making process, medical expert knowledge needs to be converted to a knowledge based system. Knowledge based system is not sufficient to suggest reliable recommendations due to the limitations in updating expert rules based on the population studies and limited personalization. Data driven approaches apply data mining and machine learning methods to extract insights from the heterogeneous data [10]. It provides individual recommendations based on the past learning experience and the patterns extracted from clinical data. Combination of information retrieval and machine learning can be used for medical database classification.

The below figure represents the source of Hybrid healthcare

recommender system.

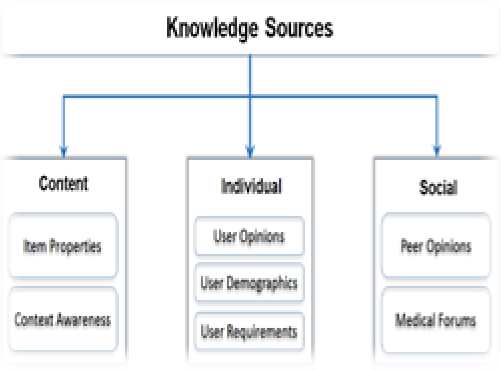

Fig. 2. Sources of Hybrid Recommender System

5.3 Types of Filtering

Collaborative filtering is the most common technique used by the recommender systems, in which the products are suggested to the user on the basis of users or items similarity. Correlations or similarities between users or items are calculated using K-Nearest Neighbor algorithm. Neighbor item ratings are combined to generate recommendations for the active user on unvisited or unrated items [7].

Content based filtering suggests the recommendations based on the user profile. For example type 2 diabetes diagnosis recommendations are suggested by keywords from patient case history [8]. The importance of words in the patient health profile can be evaluated using different weighted measure techniques, such as (a) Term Frequency/Inverse Document Frequency (TFIDF), (b) Bayesian classifiers, (c) clustering, and (d) Decision Trees (DT)[7]. For new users with few preferences, elicitation based recommendation method is used.

Domain ontologies are used to extract semantic information about items used in collaborative filtering algorithms and structured objects in medical websites as semantic entities. Ontologies are essential in healthcare domain as it helps to categorize disease, symptoms, medications, procedures, health insurance and so on.

To provide personalized healthcare recommendations, we proposed social profile enhanced recommendations based on the following criteria: (1) users with similar health concerns rate similar healthcare products, service, medication, home remedies and so on (2) users who liked similar healthcare-related items tend to like the same item in the future.

5.4 Social profile enhanced recommendation

In this approach profile similarity is computed based on the following:
1. Health Profile similarity
2. Patient behavior similarity.

Health profile information describes the patient age, location, gender and health-related concerns. Patient behavior similarity describes about the medical information accessed, healthcare social network actions, links accessed by user, user tagged by friends to health information, user's subscription to healthcare groups [5]. Combination of case based similarity and social health profile similarity is computed. If the value exceeds the threshold then the recommendation is given to the user.

5.5 Computing Case Based Similarities

Hybrid filtering approach is used to improve the accuracy of recommendations. Rule based filtering is used to filter profiles initially based on the user queries. Case based filtering is used to extract similar profiles based on patient health history. Case similarities are computed based on the KNN algorithm. Highest similarity score is suggested as recommendation.

New cases input to the HRS are compared with the existing case library. If there are no identical cases, HRS searches for the next similar cases. Case similarity is computed by KNN weighted average.

A weighted *K*-NN performs an evaluation on the attributes of the instances. Each attribute is evaluated to obtain a weight value based on how useful this attribute is for correctly identifying the classes of the dataset.

Similarity between cases is measured by the set of independent attributes. Attribute similarity is determined by the healthcare subject matter expert and stored as guiding rules. Similarity is measured by the numbers 0 or 1. Zero represents attributes are highly dissimilar and One represents attributes are likely similar. Importance of the attribute is measured as 1 –low importance or 5-high importance.

The below figure represents hybrid case based reasoning model:

Rule based methods are used for initial filtering which filters cases based on the expert rules.

Case based filtering is used to extract similar cases based on the similarities between their attributes.

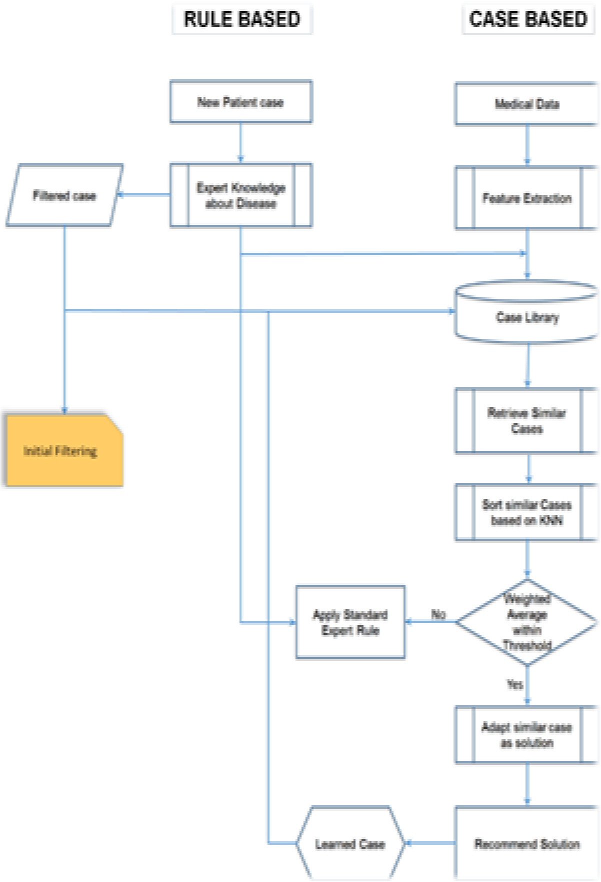

Fig 3. Hybrid Case Based Reasoning Model

The equation for computing similarity using KNN weighted average algorithm is represented as follows:

$$[1/\sum_F^N(I_F * A_F)] * \sum_X^N(I_X * A_X) \qquad (3)$$

I_F and I_X represents the importance of attributes.

A_F and A_X represents the similarity scores between attributes

Table 4. Case Similarity between attributes

Features	**Case1**	**Case2**	**Case3**	**New Case**	**Similarity Score**
Age	36	25	35	27	Case1 vs new Case:0.89
Sex	Male	Female	Male	Male	
Case history	Pre-diabetes	Type 2 diabetes	Type 2 diabetes	Pre-diabetes	Case2 vs new case:0.65
Lifestyle	Moderate Exercising	Sedentary	No Exercise	Little Exercise	
Other Health concerns	Obese	Fatigue	Depression, Fatigue	Overweight	Case3 vs New case:0.37

This table shows the similarity between existing and new cases.

Similarity (New Case, Case 1)
= 1/17 * [(1*0.8) + (1*1.0) + (5*0.9) + (5*0.9) + (5*0.9)]
= 1/17 * (0.8 + 1.0 + 4.5 + 4.5 + 4.5)
= 0.89

Similarity (New Case, Case 2)
= 1/17 * [(1*1.0) + (1*0) + (5*0.7) + (5*0.6) + (5*0.7)]
= 1/17 * (1.0 + 0 + 3.5 + 3.0 + 3.5)
= 0.65

Similarity (New Case, Case 3)
= 1/17 * [(1*0.8) + (1*0) + (5*0.9) + (5*0.2) + (5*0)]
= 1/17 * (0.8 + 0 + 4.5 + 1.0 + 0)
= 0.37

Similarity computation determines which case can be suggested as recommendation, high similarity score is suggested as solution. Based on the above computation, the system will choose case 1 as the suggestion for the new case (0.89 > 0.65 > 0.37).

Can we trust recommendations?
Mass of unreliable, redundant information on the websites makes it hard to use the information for health decision making [6]. It is necessary to present a solution that user can "trust" to information and knowledge that retrieve from recommender systems.

Trust-aware recommender systems can be built by suggesting recommendations from trustable sources such as sites approved by Health on Net authority (HON)

6. Conclusion & Future Work:

In this work both the prediction and recommendation system in the context of diabetes was studied. Prediction is done using Bayesian classifier. Healthcare recommender systems are important as people use social network to know about their health condition. Accuracy of the prediction is measured using confusion matrix. Recommender systems outcomes are recommending diagnosis, health insurance, clinical pathway based treatment methods and alternative medicines to users based on their health profile similarity with others. Hybrid Recommender system filtering approach followed is –Content filtering, Rule based and Case based filtering algorithms was used. Further study on using collaborative filtering-social profile enhanced recommendation technique will be considered to improve the accuracy of recommendations. Reliability and security of social health information will be considered. Data from wearable device such as Fitbit will be considered for improving the prediction and recommendation performance.

References

1. Edwin Morely, Big Data Healthcare, IEEE explore discussion paper, 2013
2. Keith, Nitesh, Scaling Personalized Healthcare with Big Data, International big data analytics conference in Singapore, 2014
3. Cilcia Pinto, A Spark Based Workflow For Probabilistic Linkage Of Healthcare Data, Brazilian Research Council White Paper, 2013
4. Abid Sarwar, Vinod Sharma, Intelligent Naïve Bayes Approach to Diagnose Diabetes Type-2, ICNICT 2012
5. M. Kamran1, A. Javed, A Survey of Recommender Systems and Their Application in Healthcare, Technical Journal, University of Engineering and Technology (UET) Taxila, 2015
6. Punam Bedi, Trust based Recommender System for the Semantic Web, International Joint conferences on Artificial Intelligence,2014
7. Martin Wiesner and Daniel Pfeifer, Health Recommender Systems: Concepts, Requirements, Technical Basics and Challenges, International Journal of Environmental Research and Public Health, 2014
8. A. Felfernig, R.Bruke, Constraint-based Recommender Systems: Technologies and Research Issues, ICEC, 2008
9. Vladimir Hahanov, Volodymyr MizBig Data Driven Healthcare Services and Wearables, CADSM 2015
10. J.Archenaa,Dr E.A.Mary Anita, Interactive Big Data Management in Healthcare Using Spark,Springer Smart Innovation series,2016
11. https://medlineplus.gov/ency/article/003482.htm
12. http://sebastianraschka.com/Articles/2014_naive_bayes_1.html