ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Intelligent Theme Based Integration of Medical Cases

Mangesh Mali[1], Dr. Parag Kulkarni[2], Prof. Virendra Bagade[3]

M.E. Student, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India[1]

Chief Scientist, Research Department, iknowlation Research Labs, Pune, India[2]

Asst. Professor, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India[3]

ABSTRACT: The medical field contains the numerous amount of dataset about patient's history data. But this dataset has no use without any analysis which predict health care information of patients. In the analysis the similarity measures and clustering algorithms are used to group the medical cases. While to find similarity between cases and cluster them, several methods have been proposed, but measuring the similarity between medical cases and clustering is a challenging problem. Previous works on the clustering didn't give any pure clusters. Proposed work establish a clustering algorithm which integrate medical cases based on threshold value. Proposed algorithm gives high purity clusters on medical cases.

KEYWORDS: Clustering Technique, Similarity Measures, Medical Cases.

I. INTRODUCTION

With the advent of electronic health records, more data is continuously collected from individual patients, and more data are available for review from past patients. Despite this, it has not yet been possible to successfully use this data to systematically build computer-based decision support systems that can produce clinical recommendations to assist clinicians in providing individualized healthcare. Medical Decision making should use relevant data from many distributed systems instead of a single data source to maximize its applicability, but real-world medical data are often based on missing information. This is referred as the medical information challenge.

In the past, the doctor applies their knowledge in a medical decision and diagnosis system. Then make a careful treatment on the basis of patients clinical exam result in the combination of their history. There is need to provide accurate diagnosis and treatment to help in patient recovery. There are a number of factors which influence the traditional medical diagnosis process. So, the data mining is widely used in computer-based medical diagnosis, which uses the medical cases to obtain diagnosis rule.

Medical records preprocessing is an important step which cannot be avoided in most of the situations and when handling medical data set. The attributes present in medical records may be of different data types. Also, the values of attributes have a certain domain which requires proper knowledge from the medical domain to handle them. An efficient preprocessing of medical records may expand the informational nature of medical records. In this unique circumstance, information preprocessing procedures have accomplished critical significance from medical data analysts and data miners. Wrong and inappropriate information qualities may mislead the prediction and classification results, thereby leading to improper medical treatment which is an exceptionally hazardous potential risk. This dissertation basically goes for taking care of missing attribute values present in the medical records of a dataset.

A large number of clustering definitions can be found in the in lots of papers, from simple to present. The simplest definition is the grouping together of similar data objects into clusters. It is important to understand the difference between clustering (unsupervised classification) and a discriminate analysis (supervised classification). In supervised classification, we are provided with the collection of labeled pattern, the problem is to label a newly encountered, yet unlabeled pattern. In the case of clustering, a problem is to group a given collection of unlabeled pattern into meaningful clusters. In a sense, the label is associated with clusters also, but these category labels are data driven, that is they are obtained solely from data. The specialized applies their insight into the therapeutic choice and in

DOI: 10.15680/IJIRCCE.2017. 0509039

finding framework. After applying their insight they make a watchful treatment on the premise of patients clinical exam result in a blend of their history. There is the need to give precise determination and treatment to offer assistance in patient recuperation. Various variables which can impact customary restorative determination process are introduced. The data mining is broadly utilized as a part of PC based therapeutic analysis, which utilizes the medicinal cases to get the conclusive run the show.

We focus on the creating system with respect to the user. Here we say the user could be the patient or doctor. The patient is concerned about symptoms, type of treatment and more, where the doctor is concerned about symptom study, possible causes related to new patient symptoms. In the proposed system, we analyze user (patient/doctor) search query and retrieve similar cases based on user theme. Case-based reasoning is the model which solves the problem by analyzing previously available cases and by reusing information and knowledge of the available cases. The System calculates the distance between search case and case in the cases repository using similarity measurement methods. Both case search and matching process need to be successful and time efficient.

The objective of this research is, to develop a user interactive system to provide services to a user via an interactive search for personalized patient needs. To design an appropriate structure for case content and indexing of case repository. Extract patient-specific features from medical cases. These are the most describing cases and useful for finding similarity metric. Apply similarity measures for retrieving cases relevant to search case and user theme from case repository. Recommend most similar cases to a user.

Some well know similarity measures are such as cosine similarity, Euclidean distance, Manhattan distance, etc. These measures are used with different clustering algorithms such as DBSCAN, K-means, hierarchical clustering etc. We discuss all clustering algorithms which are applicable to our datasets. After analyzing the results we take better algorithm for our research work.

II. RELATED WORK

A clustering technique defines classes and put objects which are related to them in one class on the other hand, in classification objects are placed in predefined classes. Clustering means putting the objects which have similar properties into one group and objects having dissimilar properties into another group. Clustering has alienated the large data set into groups or clusters according to similarity of properties.

Partitioning Clustering

Partitioning clustering is based on the general criterion of combining high similarity of the samples inside of clusters with a high degree of dissimilarity among distinct clusters. Most partitioning methods are distance-based. These clustering methods are working well for finding spherical shaped clusters in small to medium size databases [17].

Density Based Clustering

Most partitioning methods cluster objects based on distance between objects. In these methods the cluster continues to grow as long as the density in the neighbourhood exceeds some threshold [19].

Grid Based Clustering

In Grid based methods, the object space is quantized into a finite number of cells that form a grid structure. It is a fast method and is independent of the number of data objects and is dependent on only the number of cells present in each dimension in the quantized space.

Hierarchical Methods

In this method, hierarchical decomposition of the given set of data objects is created. It can be classified into two categories named as agglomerative or divisive, on the basis that how hierarchical decomposition is formed. Agglomerative approach is the bottom up approach, starting with each object forming a separate group. Hierarchical algorithms create a hierarchical decomposition of the data set containing data objects. It is represented by a tree structure, called dendrogram. It does not need clusters, as inputs.

Divisive approach is top down approach which starts with all the clusters in the same cluster and then at each iteration step a cluster is split into smaller clusters until each object are in one cluster.

K-Means Clustering

The k-means clustering algorithm is the basic algorithm which is based on partitioning method which is used for many clustering tasks especially with low dimension datasets. It uses k as a parameter, divide n objects into k clusters so that the objects within the cluster are identical to each other but dissimilar to other objects in other clusters. The algorithm attempts to find the cluster centers, C_1, C_2..... C_k, such that the sum of the squared distances of each data point, x_i, 1 <= i <= n, to its nearest cluster center C_j, 1 <= j <= k, is minimized. First, the algorithm randomly selects the k objects, each of which initially represents a cluster mean or center. Then, each object xi in the data set is assigned to the nearest cluster center, i.e. to the most similar center [11]. Then new mean is computed for each cluster and each object is reassigned to the nearest new center. This process iterates until no changes occur to the assignment of objects.

Gap Identification through Literature Survey

H. Cao [1] presented an approach of combing abstracted patient-specific features medical cases. The information-theoretical measure to compute similarity between cases. It is efficient method to represent cases. They implemented two information-theoretical measure in this study are corpus-dependent weighing models (the Nats model and the Bin model). Both methods, then tested by expert evaluation of case similarity. Some limitations of these paper first, study relies upon an abstracting system which abstracts the feature from medical text. Second, abstracted features are only applied into an informational-theoretic measure using two corpus-based weighing models.

D. Girardi [2] proposed a new distance measure that is better suited than traditional methods at detecting similarity in the patient record by referring to a concept hierarchy. They measure the distance to new distance measure for categorical values by considering the path distance between a concept in a hierarchy in an account. The new distance measure is an improvement over the current standard hierarchical arrangement of categorical values is available.

The author [3] presented a probabilistic approach to measuring the similarities between patient traces for client pathway analysis. They introduce three possible applications i.e., patient trace retrieval, clustering, and anomaly detection. To evaluate applications via real-world dataset of specific clinical data collected from a Chinese hospital. The patient traces could be measured based on their behavioral similarities.

M. K. Kiragu [4] developed a Case-Based Reasoning Application for treatment and management of diabetes using the jCOLBIRI CBR framework. That application uses available past patient cases to present reasoning. The system employed case based methodology of reasoning which involves the process. The success of the system depends on the use of a similarity matching between the available cases and the new search case. The system deployed and tested with real life cases and then updated by a medical expert. The accuracy of the system can further be improved by combining different pattern matching algorithms such as (Euclidean, Hamming distance, neural networks etc.).

The author [9] designed a patient similarity framework which combines both unsupervised information and supervised information. They propose a novel patient similarity algorithm that uses spline regression to capture the unsupervised information. They also propose an algorithmic framework that could incrementally update the existing patient similarity measure from Patient similarity framework using matrix theory. They should speeds up the physician feedback and newly available clinical information by introducing a general on-line update algorithm for PSF matrix.

The author [6] proposed a framework for the recommendation of the doctor and build doctor profile. They firstly suggested for finding the similarities between user's consultation and doctor's profiles. Then, to measure doctor's quality, experiences, and different users' opinions are considered. Finally, to combine the results of the relevance model and the quality model, and then recommended a doctor. A mobile recommender APP is proposed.

III. ARCHITECTURAL DESIGN

Creation of a medical case repository is the most important model of our system. We discuss below a medical case repository model, system architecture.

A. Medical Cases Creation

1. Extracting Attributes

A medical data of a patient, which normally consists admission entry, progress entry and discharge digest, usually describes patient diagnoses, signs and symptoms, diagnostic and treatment procedures, and medications. Therefore, we considered only gender, blood pressure, age, symptoms, and clinical findings. We further examine that the importance of an attribute to compute similarity. Importance might depend on its type and that features in each category might be informed. Symptom attributes and finding attribute are considered to be highly informative because these attributes give direct information about the patient’s situation. We building medical cases by extracting attribute from each patient medical data. Cases that are passed to preprocessing stage refer figure 1.

2. Preprocessing

- The collected medical cases had noise and missing values.
- These medical cases are preprocessed using MySQL workbench tool.
- Age is normalized into the child, young and senior.
- BP is normalized into high, low and normal on the basis of their reading.
- We have processed medical cases that are stored in the csv file.

3. Clustering

- Initially, we use hierarchical, k-Means, dbscan clustering algorithm, but the grouping of records was not done as the requirement on the medical dataset.
- The threshold clustering algorithm is proposed.
- Each record attribute is represented by one vector where each vector element represents an id-attribute pair, the mapping between the attributes and gives dictionary.
- Tokenized documents are converted into a sparse vectors matrix. The function doc2bow simply counts the number of occurrences of each distinct word, converts the word to its integer word id and returns the result as a sparse vector.
- In preparation for similarity matrix, the transformation is used to convert documents from one vector representation into another. We used Tf-idf, a simple percentile, which takes documents represented as bag-of-words counts and applies a weighting which discounts common term. It also scales the resulting vector to unit length.
- The threshold value is decided by percentile function. A percentile is where a sample is divided into equal-sized, adjacent. It can also refer to dividing a probability distribution into areas of equal probability.
- Comparing threshold value with similarity value, similarity matrix is converted into the logical matrix.

- The attribute of a row having maximum sum is considered in one group and the same procedure is done until no row remain.
- The cluster is nothing but the result of groups.
- The medical cases are stored according to cluster in the database.

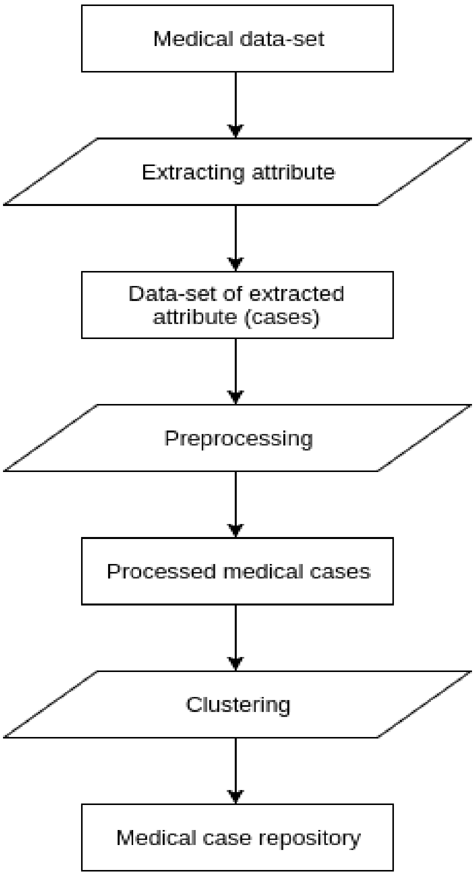

Fig. 1. Flow of Medical Cases Repository

B. System Overview

1. The system is to retrieve the health care information to the user based on user theme.
2. Our system divides into two subsystem one called retrieval system and other similarity system.
3. The retrieval system takes the input from the user and gives the health care information to the user.
4. The similarity system has extracted feature model which extract features of search theme.
5. The case similarity model finds appropriate cluster from the medical case repository for the features refer figure 2.

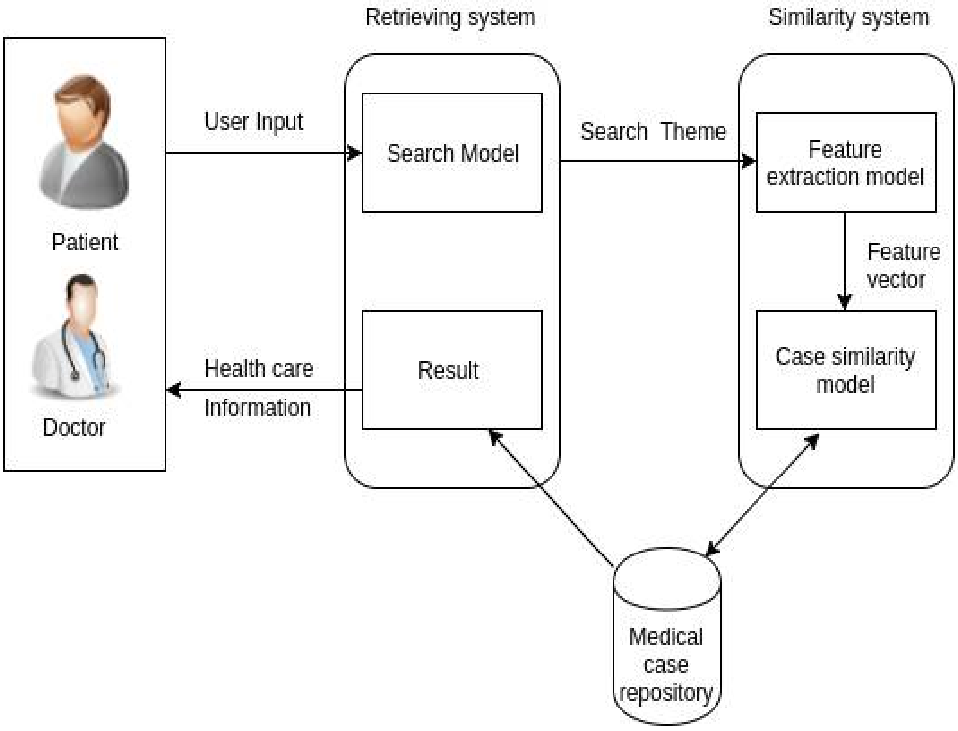

Fig. 2. System Architecture

Medical patients records are represented in csv table. This is a tiny corpus of ten records, each consisting of five attribute value. Each record is represented as a string. Records are transformed into vectors, we'll use a document representation called bag of words. In this representation, each record is represented by one vector. The dictionary is generated from the records which contains the unique words from our corpus along with ids of it. Mapping between words and their ids:

IV. PROPOSED ALGORITHM

A. Data description and dictionary creation:

{'normal/normal': 13, 'zxc': 20, 'FEVER': 1, 'PAIN IN LT KNEE JT': 6, 'BURNING IN FEET ': 18, 'high/normal': 5, 'PAIN IN RT SIDE NECK': 11, 'young': 3, 'OEDEMA ON FACE': 4, 'FUNGAL INFECTION': 14, 'COUGH STILL': 9, 'URTI': 10, 'FEMALE': 15, 'PROLAPSED HAEMORRHOIDS': 12, 'COUGH': 17, 'LT RENAL COLIC': 7, 'high/high': 0, 'MALE': 2, 'senior': 8, 'EXT HAEMORROIDS': 19, 'CLOTHES': 16}

B. Generation of vectors:

The function doc2bow() simply counts the number of occurrences of each distinct word, converts the word to its integer word id and returns the result as a sparse vector. The sparse vector [(0, 1), (1, 2), (2, 1), (3, 1)] says that record contains -word id (0, 2, 3) appears once and word id-1 appears twice. And other dictionary words appear zero times [7].

[(0, 1), (1, 2), (2, 1), (3, 1)]
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1)]
[(2, 1), (3, 1), (5, 1), (6, 1), (7, 1)],
[(2, 1), (5, 1), (8, 1), (9, 1), (10, 1)]
[(0, 1), (2, 1), (3, 1), (11, 1), (12, 1)]
[(3, 1), (13, 1), (14, 1), (15, 1), (16, 1)]
[(0, 1), (1, 1), (3, 1), (15, 1), (17, 1)]
[(5, 1), (8, 1), (10, 1), (15, 1), (18, 1)]
[(1, 1), (5, 1), (8, 1), (15, 1), (19, 1)]
[(0, 1), (1, 2), (2, 1), (3, 1)]

Term Frequency and Inverse Document Frequency (Tf-idf), expects a bag-of-words (integer values) training corpus during initialization. During transformation, it will take a vector and return another vector of the same dimensions, except that features which were rare in the training corpus will have their value increased. It therefore converts integer-valued vectors into real-valued ones, while leaving the number of dimensions intact. It can also optionally normalize the resulting vectors to (Euclidean) unit length.
A common reason for such a Tf-idf is that we want to determine similarity between pairs of documents, or the similarity between a specific document and a set of other documents. The similarity between our documents, we need to enter all documents which we want to compare with each other.

C. Generating similarity matrix:

From the above example set we get next values:
00= 1.00, 01= 0.39, 02= 0.08, 03= 0.07, 04= 0.15, 05= 0.01 , 06= 0.29, 07= 0.00, 08= 0.15, 09= 1
11= 1.00, 12= 0.04, 13= 0.04, 14= 0.09, 15= 0.01, 16= 0.11, 17= 0.00, 18= 0.04, 19= 0.39
22= 1.00, 23= 0.11, 24= 0.03 25= 0.07 , 26= 0.01, 27= 0.08, 28= 0.09, 29= 0.08
33= 1.00, 34= 0.03, 35= 0.00, 36= 0.00, 37= 0.47, 38= 0.25, 39= 0.07
44= 1.00, 45= 0.06, 46= 0.04, 47= 0.00, 48= 0.15, 49= 0.04
55= 1.00, 56= 0.06, 57= 0.04, 58= 0.04, 59= 0.01
66= 1.00, 67= 0.06, 68= 0.11, 69= 0.29
77= 1.00, 78= 0.31, 79= 0.00
88= 1.00 , 89= 0.29
99=1.00

We need to decide the threshold value. The percentile function is used to calculate threshold value of the similarity matrix. The value is based on how we set the percentile function on the similarity matrix, i.e. 25%, 50%, 75% etc. As the threshold value changes differ in the clustering results. That will need to be grouped the documents. If similarity is greater than a given threshold value. We decide here threshold value = 0.2. We create a zero-matrix having a length similar to the number of documents in the corpus. Rows and column representing exact document. For example, if row index 1 and column index 8, that is, we suppose the similarity between document 1 and document 9. Here Document (1, 9) = 0.39, so 0.39 greater than the threshold value. That is, we are placing `1' at that position (1, 9) in the matrix. Likewise comparing similarity value between all documents and updating corresponding positions of the matrix. We get the chain of similarity between documents. Here we get a logical matrix.

For Ex : logical Matrix

```
[[ 1.  1.  0.  0.  0.  0.  1.  0.  0.  1.]
 [ 1.  1.  0.  0.  0.  0.  0.  0.  0.  1.]
 [ 0.  0.  1.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  1.  0.  0.  0.  1.  1.  0.]
 [ 0.  0.  0.  0.  1.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  1.  0.  0.  0.  0.]
 [ 1.  0.  0.  0.  0.  0.  1.  0.  0.  1.]
 [ 0.  0.  0.  1.  0.  0.  0.  1.  1.  0.]
 [ 0.  0.  0.  1.  0.  0.  0.  1.  1.  0.]
 [ 1.  1.  0.  0.  0.  0.  1.  0.  0.  1.] ]
```

D. Select row with maximum sum:

We have a logical matrix, we find the row with the maximum sum. In this example, there are 2 rows with the maximum one's so we select the first one which is 0 th. (Considering indexing starts from zero)

E. Create new groups:

We have selected first row in the matrix (i=0) and now all the columns which contain number 1 in the selected row should be added to a new group which will be named G1.

G1 = {0, 1, 6, 9}

Now we pick up one by one indices in the current group and repeat previous procedure. Here, row with index 1, having {0, 1, 6} indices so we don't have to add indices to current group. Then we take the next index from G1 and so on.

F. Delete grouped documents:

After we got a new group is to delete values in the matrix, only rows and columns with indices have to delete. Our grouped documents with indices {0, 1, 6, 9} then the matrix should look like this:

```
[[1.  0.  0.  0. 0.  0.]
 [0.  1.  0.  0. 1.  1.]
 [0.  0.  1.  0. 0.  0.]
 [0.  0.  0.  0. 0.  0.]
 [0.  1.  0.  0. 1.  1.]
 [0.  1.  0.  0. 1.  1.]]
```

Now from step no. 4 we repeat, choose a new row to be processed. Repeat the same procedure until our matrix only contains zeroes.

G. Result of above procedure:

[[0, 1, 6, 9], [3, 7, 8], [2], [4], [5]].

V. RESULTS

The expected output for the above-mentioned work should be at least provide medical cases from the raw data set which does not contain missing attribute value and noise. The proposed approach has been tested on real data set.
Imputation measure uses the threshold based clustering algorithm which runs in the O(m * n) time where, m is a number of records and n is the number of attributes. We can consider the constant number of attributes to reduce the complexity of the algorithm.
For creating medical cases missing data should be separated from complete records and missing data should be imputed effectively. So, that there should be high accuracy in generated case library. We are using unsupervised machine learning algorithm because the available dataset belongs to medical and training data is not available. We have used different clustering algorithm from literature survey on real time medical dataset and the detail result analysis for different trials have been presented in Table I.

Table I
Result for Different Clustering Algorithm

Algorithm	Result Analysis
K-means	K-Means is an iterative process of clustering; which keeps iterating until it reaches the best solution or clusters in our problem space. But the basic question which should arrive is that how to decide the number of clusters (K)? There is no mathematical formula which can directly give us answer to K but it is an iterative process where we need to run multiple iterations with various values of K.
Hcluster	In hierarchical clustering, once a decision is made to combine two clusters, it cannot be undone. No objective function is directly minimized in it. This algorithm is sensitivity to noise and outliers and has difficulty in handling different sized clusters and convex shapes
DBSCAN	The quality of DBSCAN depends on the distance measure used in the function region query(P, epsilon). DBSCAN cannot cluster data sets well with large differences in densities, since the minPts epsilon combination cannot then be chosen appropriately for all clusters

The purity of cluster indicates that there is a good structure to the clusters, with most observations seeming to belong to the cluster that they're in. The table II shows how to use the value of purity:

Table II
Range of the purity Interpretation

Range of Purity	Interpretation
0.71-1.0	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial
<0.25	No substantial structure has been found

The purity of first five clusters formed in proposed threshold based clustering algorithm are shown in table III

Table III
Cluster Purity

Cluster Number	Purity Measure
Cluster 0	0.88
Cluster 1	0.87
Cluster 2	0.99
Cluster 3	0.81
Cluster 4	1.0

VI. CONCLUSION AND FUTURE WORK

In the present research, the challenges of forming the pure cluster in the medical dataset are addressed. The works come up with an approach of forming the clusters in the dataset. The threshold based algorithm generates the cluster based on threshold value which is calculated by a similarity matrix using the percentile function. Two cases are similar, if Similarity between two cases is greater than the threshold value. Compared with other clustering algorithms proposed work generates better purity clusters. So our method predicts the accurate health care information for given user search theme. Present system is does't have more accurate and detailed dataset. So the there is a restriction on predicting the correct information. In the future our approach should be to collect, the more detailed dataset. The proposed work has more complexity because of more attributes in the datasets. So our future approach should be to lower the complexity.

REFERENCES

1. H. Cau and G. B. Melton and M. Markatou and Hripcsak, G., `Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases', Journal of Biomedical Informatics, Vol. 41, pp. 882-888, 2008.
2. D. Girardi and S. Wartner and G. Halmerbauer and M. Ehrenmüller, ` Using concept hierarchies to improve calculation of patient similarity', Journal of Biomedical Informatics, Vol. 63, pp. 66-73, 2016.
3. Z. Huang and W. Dong and H. Duan and H. Li, `Similarity Measure Between Patient Traces for Clinical Pathway Analysis: Problem, Method, and Applications', IEEE Journa of Biomedical and Health Informatics, Vol. 18, 2014.
4. Kiragu M. K. and Waiganj P. W., `Case based Reasoning for Treatment and Management of Diabetes', International Journal of Computer Applications Volume, Vol. 145, 2016.
5. Daltayanni M. and Wang C.and Akella R. and Patil P. and Subbaraya C. K., ` A Fast Interactive Search System for Healthcare Services', Proc. Service Research and Innovation Institute Global Conference, 2012.
6. Jiang H. and Xu W., ` How to find your appropriate doctor :An integrated recommendation framework in big data context', Proc. Springer Science and Business Media New York, 2014.
7. Choudhury N. and Begum S. A., `A Survey on Case-based Reasoning in Medicine', International Journal of Advanced Computer Science and Applications. Vol 7, pp. 136-144. 2016.
8. Elavarasan, N. and Dr. K. Mani, `A Survey on Feature Extraction Techniques', International Journal of Innovative Research in Computer and Communication Engineering, Vol. 66, pp. 43-46, 2015.
9. K. Hande `PSF: A Unified Patient Similarity Evaluation Framework Through Metric Learning With Weak Supervision', IEEE Journal of Biomedical and Health Informatics, Vol. 19, 2015.
10. Raghupathi, W. and Raghuoathi, V. ` Big data analysis in heakthcare: promise and potential', Proc. Heath Inform. Sci. Syst., 2014.
11. T. Bossomaier ` ModEx and Seed-Detective: Two novel techniques for high quality clustering by using good initial seeds in K-Means', Journal of King Saud University - Computer and Information Sciences, Vol. 27, pp. 113-127, 2015.
12. Batagelj, V. and Mrvar, A. and Zaversnik, M. ` Partitioning approaches to clustering in graphs', Applied Artificial Intelligence, pp. 375-381, 2008.
13. Ester, M. and Kriegel, H. and Sander, J. and Xu, X., `A density-based algorithm for discovering clusters databases with noise', Proc. of 2nd Int. Conf. on Knowledge Discovery and Data Mining(KDD-96), pp. 226-231, 1996.

BIOGRAPHY

Mangesh Mali is a student pursuing M.E. in the Computer Engineering Department, Pune Institute of Computer Technology, pune. His research interests are Data Mining, Data Analysis and Machine Learning.

Dr. Parag Kulkarni is Chief Scientist and CEO of the iKnowlation Research Labs Pvt Ltd, an innovation, strategy and business consulting and product development organization. He has been visiting professor/researcher at technical and B-schools of repute including IIM, Masaryk University – Brno, COEP Pune.

Prof. Virendra Bagade is an Assistant Professor in the Computer Engineering Department, Pune Institute of Computer Technology, pune. His research interests are Data Mining and Data Warehouse, Information retrieval.