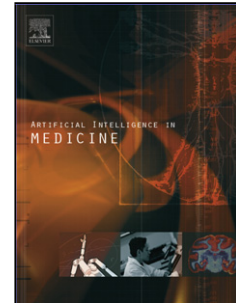


Journal Pre-proof

A modular cluster based collaborative recommender system for cardiac patients

Anam Mustaqeem, Syed Muhammad Anwar, Muhammad Majid



PII: S0933-3657(18)30540-2
DOI: <https://doi.org/10.1016/j.artmed.2019.101761>
Reference: ARTMED 101761

To appear in: *Artificial Intelligence In Medicine*

Received Date: 3 September 2018
Revised Date: 15 December 2018
Accepted Date: 7 November 2019

Please cite this article as: { doi: <https://doi.org/>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier.

A Modular Cluster Based Collaborative Recommender System for Cardiac Patients

Anam Mustaqeem^a, Syed Muhammad Anwar^{a,*}, Muhammad Majid^b

^aDepartment of Software Engineering, University of Engineering and Technology Taxila, Pakistan

^bDepartment of Computer Engineering, University of Engineering and Technology Taxila, Pakistan

Abstract

In the last few years, hospitals have been collecting a large amount of health related digital data for patients. This includes clinical test reports, treatment updates and disease diagnosis. The information extracted from this data is used for clinical decisions and treatment recommendations. Among health recommender systems, collaborative filtering technique has gained a significant success. However, traditional collaborative filtering algorithms are facing challenges such as data sparsity and scalability, which leads to a reduction in system accuracy and efficiency. In a clinical setting, the recommendations should be accurate and timely. In this paper, an improvised collaborative filtering technique is proposed, which is based on clustering and sub-clustering. The proposed methodology is applied on a supervised set of data for four different types of cardiovascular diseases including angina, non-cardiac chest pain, silent ischemia, and myocardial infarction. The patient data is partitioned with respect to their corresponding disease class, which is followed by k-mean clustering, applied separately on each disease partition. A query patient once directed to the correct disease partition requires to get similarity scores from a reduced sub-cluster, thereby improving the efficiency of the system. Each disease partition has a separate process for recommendation, which gives rise to modularization and helps in improving scalability of the system. The experimental results demonstrate that the proposed modular clustering based recommender system reduces the spatial search domain for a query patient and the time required for providing accurate recommendations. The proposed system improves upon the accuracy of recommendations as demonstrated by the precision and recall values. This is significant for health recommendation systems particularly for those related to cardiovascular diseases.

Keywords: recommender system, clustering, collaborative filtering, cardiovascular disease, decision support.

1. Introduction

E-health is one of the most significant application of communication technology. Remote health management on a personal level is evolving as a mean of early detection and prevention of diseases in distant areas. The advent of e-health promises to provide solutions for issues effecting current health-care systems such as cost, availability, and accessibility. E-health based system is an effective way to motivate people, specially those living far from clinical facilities, to actively participate in monitoring and managing their health with in the comfort of their living space. This is predominantly effective in monitoring and curing chronic diseases, especially cardiovascular diseases. Existing e-health systems incorporating wearable sensors are aimed at data collection and transmission such that both subjects and clinical experts can observe significant clinical markers [1].

Among chronic diseases, heart attack is a common reason for hospitalization, especially in aging population [2], where patients demand a high-quality in disease monitoring and personal care. Cardiovascular

*Corresponding author

Email address: s.anwar@uettaxila.edu.pk (Syed Muhammad Anwar)

diseases generally require an intensive follow up for proper treatment, and are a leading cause of approximately two-third deaths around the globe [3]. General clinical recommendations for curing these diseases are associated with lifestyle changes, such as to enforce a healthy diet, physical exercise, and a reduction in sugar intake and alcohol consumption [4, 5]. The automation of health monitoring systems cannot fully replace physical patient monitoring, where a patient gets a chance to directly interact with a health physician. The utilization of automated e-health systems can save time and reduce the amount of resources required in cases of emergency [6].

Recommender system is a recognized technology, which is used in various fields such as social media, marketing, academics and music industry [9]. Over the past few years, recommender systems have developed significantly, yielding refined and specialized approaches, which depend on the purpose, domain, and level of personalization [10]. Unlike traditional machine learning methods, recommender systems can efficiently deal with data acquisition and processing, and are able to authorize an insight in to the process of decision making and result analysis. However, current recommender systems have not yet found broad application in medicine to utilize its potential and benefits.

A large and multidimensional volume of data demands technical aids for an efficient utilization and integration so as to develop clinical decision support systems [11]. These support systems are designated to assist medical experts and professionals to take rapid medical decisions regarding disease diagnosis and treatment. The clinical support systems are implemented using data mining and machine learning methods for extracting useful information from complex medical datasets, which are used as an aid in disease diagnosis and treatment [12, 13]. Collaborative filtering is one of the most commonly used technique among recommender systems and is based on the concept that users having interest in a particular item may as well have similar interest in other items [14]. The process of collaborative filtering is performed in three major steps i.e., pre-processing for building the user-item matrix, computing similar users for the query user, and generating recommendations for the query user. The user-item matrix is composed of the user's ratings for the items of interest. These preferences are explicitly obtained by ratings of an item on a scale of (1 – 5), or implicitly by their purchase history [15].

In this study, clinical data records of cardiac patients are collected for recommending general medical treatments to four types of cardiovascular diseases including angina (AN), non-cardiac chest pain (NCCP), silent ischemia (SI), and myocardial infarction (MI). A team of expert cardiologists is involved for labelling the dataset corresponding to the disease diagnosis and treatment. Moreover, consequence of the suggested treatment is collected from patients during follow up visits to verify the effectiveness of the prescribed recommendation. The proposed system works on modularization by implementing a clustering based collaborative recommender system. This approach overcomes the problem of data sparsity and scalability in collaborative filtering. The suggested medical advice is evaluated by the system based on several evaluation metrics. The profile of the related patient is modified by concatenating the recommendations to their medical records. The modified patient record is stored in the repository of training dataset after taking feedback from the patient. The feedback is helpful in improving the system performance for treatment recommendations. The proposed model has following contributions,

1. The collection and organization of cardiac patients dataset at a renowned local hospital, including the clinical and demographic features of admitted patients, clinical diagnosis of disease, suggested treatments, and patient's feedback on these suggestions.
2. Improvisation in the steps of collaborative filtering to address the scalability issue.
3. The execution of sub-clustering at two levels to ensure fast and robust similarity computation.
4. The involvement of cardiologists in the process for clinical approval and disapproval of the resulting outcomes.

The rest of the paper has following sections. In Section II, related work for health recommender systems is briefly described. Section III, presents a detailed overview of the proposed recommender system. The details of experimental set up including dataset and results are presented in Section IV. Finally, conclusion and future work is discussed in Section V.

2. Related work

Recommender systems for health are being used for the past few years [17], but due to the diversity of disease symptoms, varying categories of health professionals and asymmetry of available information, these systems are facing challenges [18]. The evolving web based medical data banks and online groups are providing user reviews on doctor's information, making it possible for the health recommender systems to provide relevant services [19]. A hybrid model was implemented for the prediction of heart disease and medical recommendations using statistical analysis techniques[20]. The aim of the model was to target the risk associated with heart diseases and calculate the probability of disease occurrence using clinical features. In [21], an improved collaborative filtering based algorithm was proposed for doctor assessment and achieved the goal of personalized health recommendations in large data sets.

A doctor recommendation framework was developed by taking in to account the patient disease symptoms and preferences [22]. The methodology was based on finding similar users using their consultations with doctors and the relevant doctor's profile, which was followed by measuring the strength of a doctor's profile based on user's opinion. In [23], opportunities and challenges required for merging health education techniques with recommender systems were highlighted. The utilization of social networking and web semantic technologies was demonstrated in detail. In e-health social networks, the purpose of a health recommender system is to suggest similar patients based on their semantic similarity to the social centric patient [24, 25]. A novel patient centric technique [26] was implemented for classifying patient's health records according to ontologies using web semantics and provide treatment plans by obtaining anonymous crowd-sourced data for patients having similar diseases. Similarly, a system was proposed to use correlations among nursing diagnoses, outcomes and interventions to create a recommender system for constructing nursing care plans. The proposed system utilized a prefix tree structure, common in item based data mining techniques, to construct a ranked list of suggested care plans [27].

Machine learning techniques for implementing recommender systems are gaining importance for the past few years. Among these collaborative filtering is a widely used and successful technique, which recommend items to users based on the similarities in their behaviour and preferences [30]. For similarity computation, many calculations are required, which becomes time consuming and difficult for a large amount of data. To overcome this problem, data grouping or clustering is introduced in collaborative filtering [31, 32, 33]. However, the issue of scalability will exist, if the data is not properly grouped. This requires a system, which allow grouping of supervised data. This grouping task must be performed at various levels to ensure fast and robust computations. A therapy decision system was implemented [12] for providing therapy recommendations to patients based on the outcome of previous cases. The system introduced a combination of collaborative and demographic filtering approaches, but it was proposed for skin diseases. The system also lacked the capability of handling data sparsity, which resulted in lower recommendation accuracy. The problem of data sparsity and scalability exist in most of the traditional collaborative filtering approaches. Therefore, to overcome the problem, data clustering was implemented in collaborative filtering [31, 32, 33].

However, these systems still lacked in providing a robust method, as data was not grouped properly and physicians were not involved in the recommendation process, which reduced the reliability of system. Moreover, these systems were not tested on a specific cardiac disease dataset, which raises questions on the accuracy of the proposed methods.

3. Proposed Recommender System

The proposed modular clustered based collaborative recommender (MC_bCR) system is shown in Figure 1, which is built to address the challenges of collaborative filtering including data sparsity and scalability. The model utilizes the benefits of clustering in providing a robust recommender system by addressing these issues, so that the effectiveness and accuracy is not compromised. This is of prime importance, since a health recommendation system is unique and critical as compared to other context based recommender systems. A health recommendation generated by an error or a system flaw can be life threatening for the patient. Moreover, the proposed methodology aims at delivering a modular clustered based health

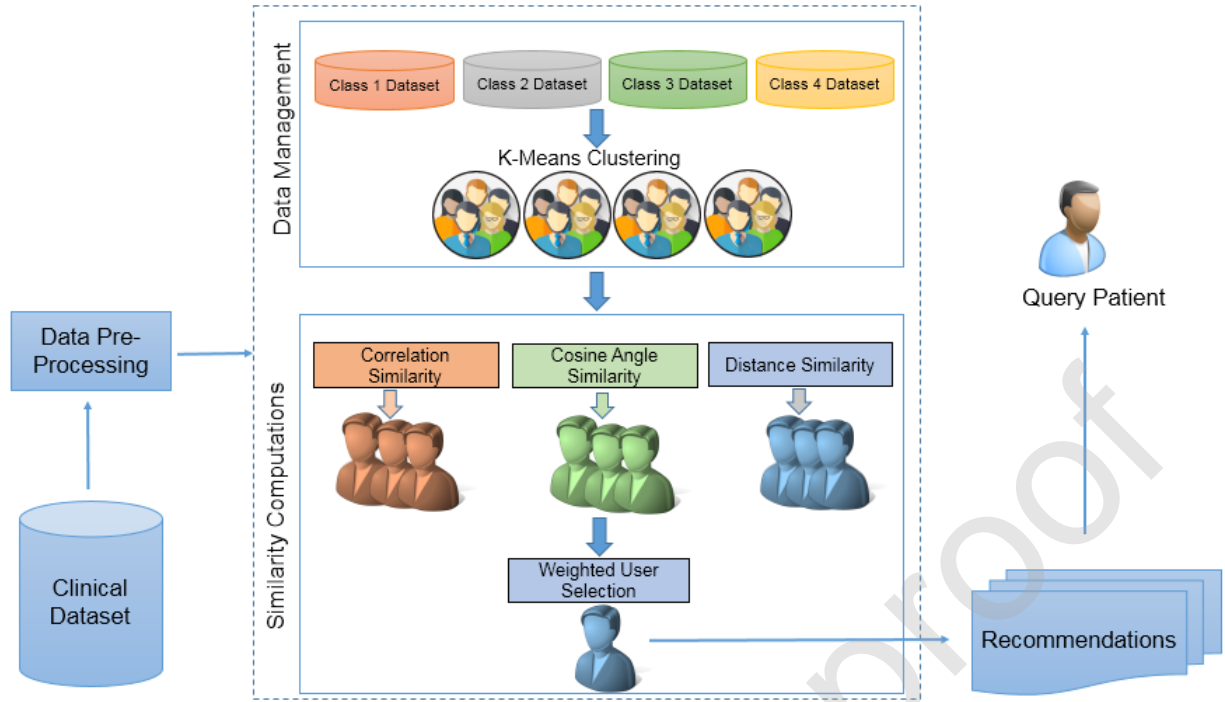


Figure 1: The proposed modular cluster based recommender system showing the process of generating recommendations for a query patient.

recommender system that can aid physicians in taking rapid clinical decisions for disease diagnosis and treatment depending on past experiences.

The clustered model is built on a supervised dataset collected from cardiology department of a renowned local hospital. Clustering is applied as a step during collaborative filtering for fast similarity computation. After clustering, similar patients within the cluster are identified for a query patient, and necessary medical advices are recommended, which depend on the treatment and feedback of similar patients. Recommendations are generated as medical advices selected from a list of labelled categories including, various kind of exercises, reference to another physician, type of food and medicine intake, and specific clinical treatment. Each disease cluster is tuned individually to introduce modularization at recommendation level to perform recommendation step for all clusters in parallel. In this way, more recommendations are generated simultaneously, which decreases the overall computational cost of the proposed recommender system. The modularization used is directly proportional to the number of clusters involved. An overview of the modularization model used in the proposed study is shown in Figure 2.

The proposed MC_bCR system improvises in each step of a collaborative recommender system to ensure rapid, accurate, and robust recommendations. An additional step of data management is added to organize the data for fast computation. The MC_bCR system for cardiac patients comprises of four main phases i.e., data pre-processing, data management, similarity computation and decision making. A detailed description of these steps is presented in the following sub-sections.

3.1. Data Preprocessing

In order to perform a meaningful analysis on the available data, preprocessing is performed on raw data values to introduce completeness by properly handling missing values and outliers. Missing values, unrealistic data and repeated records are handled using numerical cleaner filter [34], which cleanses the numeric data by setting a default value. The missing values in data are assigned the mean value of the data distribution.

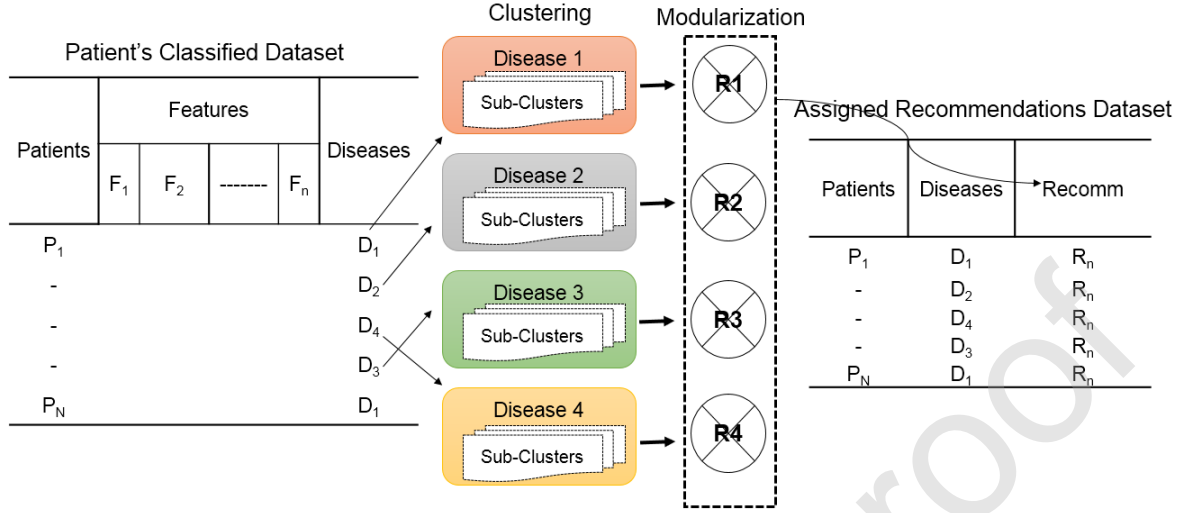


Figure 2: An overview of the modularization used to generate sub-clusters.

3.2. Data Management

The growing volumes of data and economical computational measures have made it possible to produce models that can analyze large and complex data, thereby providing fast and accurate results [35]. In the data management step, the labelled dataset is partitioned into four groups of cardiovascular diseases i.e., AN, NCCP, MI and SI, depending on their clinical features and labels. These diseases are represented by four different colors in Figure 1 and 2, and are selected for this study. This is a hard-clustering approach, as each patient is already classified with a specific disease. Further sub-clusters are introduced within each cluster using k-mean clustering as shown in Figure 1 to generate fast recommendations.

3.2.1. Cluster Centroid based Sub-clusters (Tuning)

Once the clusters are formed, a step is needed for tuning clusters to introduce more meaning to data. Tuning can be different for each cluster depending on the sparsity and density of data. This step is performed when more training data is added in the respective clusters. Each cluster is further divided into sub clusters by applying k-mean clustering. The purpose of sub-clustering is to minimize the similarity computations for the query patient, hence reducing time, while not compromising on accuracy. This helps in reduction of the training data to purposeful data groups, which are evaluated to generate predictions for a query patient. K-mean clustering is applied on the partitioned data for sub-clustering and is described below.

3.2.2. K-mean Clustering

K-mean clustering involves the partitioning of a given number of observations N into k clusters, where k is defined in advance, such that $k < N$. The first step in the algorithm requires the selection of k cluster centroids, such that the centroids have minimum similarity among themselves [36]. K-mean algorithm is highly unstable in choosing initial centroids for clusters, which heavily effects the performance in terms of efficiency [37]. The algorithm randomly chooses initial centroids by default, and it does not provide any assurance of producing unique results after clustering. To produce effective results, initial cluster centroids are chosen using a criterion based on standard deviation. The centre points are chosen such that the variance

between clusters is maximized. The initial point is selected randomly, and the other points are selected by calculating the standard deviation. Those points having the minimum deviation from the mean value are chosen as seeds.

After selecting the initial seed points, the association of data points with the specified centroids are identified. The points are allocated to the cluster based on the distance measure, where smaller the distance to the mean point of a cluster centroid, the more are the chances of allocating the point to that cluster. Once the clusters are populated with data points, a new mean of cluster is calculated, which acts as the new centroid and the allocation process is repeated. This process keeps on repeating until the iterations end or the centroid points stop changing their location. The new mean point for the cluster is calculated using,

$$\mu(cluster_k) = \frac{\sum_{i=1}^{m_k} x_i}{m_k}, \quad (1)$$

where m_k shows the total number of data points in the cluster k , and x_i represents the individual data points in cluster k . The aim of the process is to minimize the objective error within the cluster, which is calculated using,

$$\arg \min_x \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - m_i\|^2, \quad (2)$$

where k represents the number of clusters and m_i shows the mean of data points in the cluster S_i . In this work, k is initialized with a value of 9. The number of iterations are varied to obtain the best possible results.

3.3. Similarity Computation

The similarity is computed by using statistical techniques to find users similar to a query user on the basis of their past interests [38, 39]. The measure of similarity defines the relationship between a pair of users and map results in an interval of (0,1), where 0 indicates no similarity, and 1 indicates maximum similarity. The creation of sub-clusters ensures further division of patients having similar disease class, with the purpose of providing simpler and cost-effective computations for a query patient. A patient from the test data, who needs to be diagnosed with a specific type of cardiovascular disease may require a treatment based on the experience of similar patients. For this purpose, the query patient is first placed in the desired disease cluster, and then the patient is further placed in the sub-cluster in accordance with the process defined in k-mean clustering. In this scenario, the query patient will only be compared with sub-cluster patients, thus reducing the effort and cost for similarity computation. The steps involved in finding similar patients to an active patient and then providing useful recommendations are shown in Figure 3. The similar patients are selected based on the distance measure. If all selected patients are distinct, as represented by the distinct patient box, those patients having the minimum distance are selected. Otherwise, a majority voting scheme approach is used, where patients selected by multiple distance measures are selected. The flow diagram also highlights the fact that the multiple recommendations could be selected based on the final score. In the proposed model, three commonly used similarity measures are used for finding nearest neighbors of a query patient. K-nearest neighbor (KNN) approach is used for finding the nearest neighbors with value of K set to three. The value of K is selected after a series of experiments, where the selected value gave the highest performance. The value of k is selected to be 9 and with three distance measure, the maximum allowed size for K is 3.

The idea is to find K similar patients to a query patient using three different similarity measures i.e., Euclidean distance, cosine similarity and Pearson correlation. These three similarity measures are used to reduce the ambiguities, which might occur due to the limitations of an individual algorithm. Each similarity measure is used to find three similar patients, therefore a set of nine similar patients are obtained. To select the three most similar patients, a majority voting criteria is used. If all nine patients obtained through the process are distinct, then the top K most similar results are used for further processing. In case of non-distinct results, a majority voting criteria is employed for selecting the k-nearest neighbors as shown in Figure 3. This process is executed in all clusters, if active patients fall in each cluster as shown in Figure 2.

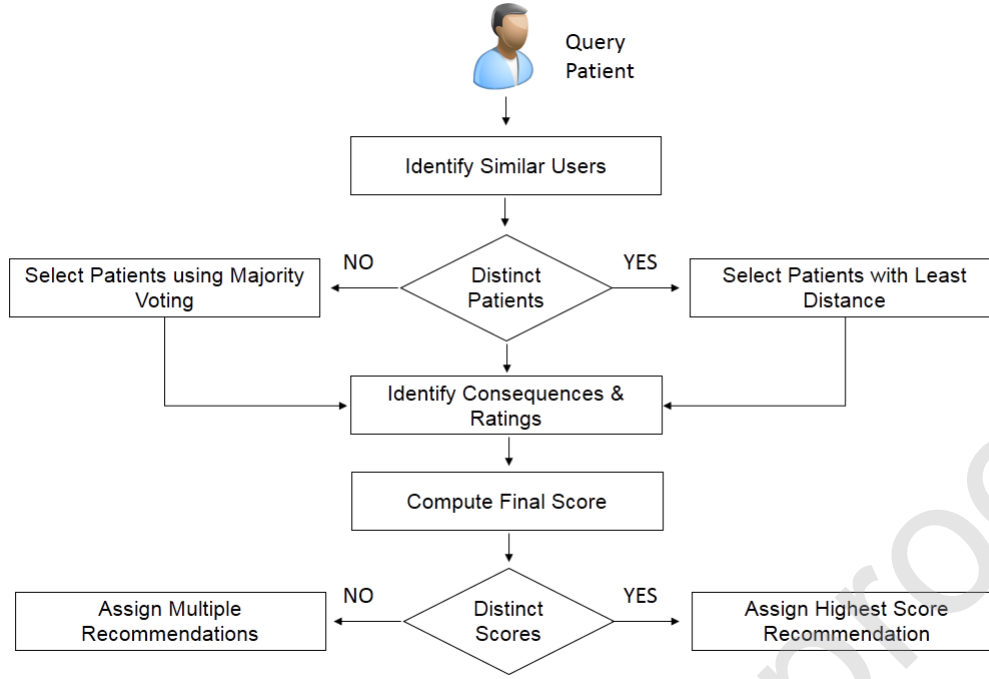


Figure 3: A flow chart for the proposed method used for similarity and final score computation.

In this case, effectively four different active patients can be evaluated simultaneously, hence introducing the concept of modularization. A brief description of the similarity measures used in this work are presented in the following subsections.

3.3.1. Pearson Correlation based Similarity

This method is based on a similarity score among two variables represented as Pearson coefficient, whose value is between -1 and $+1$, and represent the magnitude and level of association or correlation between two variables [39]. By getting the Pearson correlation value, one can also project the direction of relationship for next iterations. Generally, three types of results are derived from this coefficient i.e., -1 (negative relationship between two variables), $+1$ (positive relationship between two variables), and 0 (no relationship between two variables). The Pearson correlation value is calculated as,

$$r(a, b) = \frac{na \sum b_i - a(\sum_{i=1}^n b_i)}{\sqrt{[n(a^2 - (a^2))] [n \sum_{i=1}^n (b_i)^2 - (\sum_{i=1}^n b_i)^2]}}, \quad (3)$$

where r is the correlation coefficient, n shows the total number of patient records, a shows the data value of a query patient and b_i shows the variable data of the patient with whom similarity has to be measured.

3.3.2. Vector Cosine based Similarity

Cosine similarity is also used to find correlation between two variables. This similarity measure take items and their respective ratings as a vector and computes correlation by measuring the angle between them [39]. Cosine works on the orientation of two vectors and does not consider the magnitude value. Similarity will be maximum, when two vectors are at 0 degrees, whereas two vectors at 90 degrees will have no similarity. The cosine similarity is calculated as,

$$sim(a, b) = \frac{(r_a)(r_b)}{\|r_a\| \|r_b\|} = \frac{\sum_{k \in item} R_{ak} R_{bk}}{\sqrt{\sum_{k \in item} R_{ak}^2} \sqrt{\sum_{k \in item} R_{bk}^2}}, \quad (4)$$

Table 1: Treatment consequences and weights

Treatment consequence	Weight
Improvement	1
Therapy	0.6
No-Improvement	0.2
Death	0

where r_a and r_b are the ratings given by user a and user b , R_{ak} shows the rating by user a for item k and R_{bk} shows the rating by user b for item k , and $item$ represents cardiovascular diseases.

3.3.3. Euclidean Distance

Euclidean distance is used to find the distance between two points [38] represented as the length of a path that joins these points. This measure is only used, when data is calculated using the same scale. Similarity in this case is inversely proportional to the distance measure. The Euclidean distance is calculated as,

$$d(a, b) = \sqrt{\sum_{k=1}^n (a_k - b_k)^2}, \quad (5)$$

where n is the total number of dimension attributes, a_k and b_k are k^{th} value of data instances a and b . If scale is different, then the data needs to be standardized to get accurate results.

3.4. Decision Making

In this study, a data repository of patients suffering from cardiovascular diseases is developed. Medical recommendations are provided by the physician, which are used for training the model. Once the new active patient is diagnosed with a similar disease, its attributes are compared with other patient records stored in the knowledge base, to find the most similar patients. A set of similar patients selected through similarity algorithms are then used for the purpose of generating recommendations. The labelled dataset, where recommendations for different cardiac patients are collected with the assistance of medical experts, is used during the training phase. The patients and consequences after treatment are continuously observed. The consequences include classes such as therapy, death, medication with improvement, and medication without improvement. Each consequence is assigned weights depending on its severity in consultation with the medical experts. The weight of these consequences is shown in Table 1. The four consequences and their weight values are assigned after consultation with multiple clinical experts involved in this study.

Rating and feedback is an important aspect of any recommender system as they indicate their level of success. Positive feedback helps in utilization of a system on a professional level, while a negative feedback can help in identifying the flaws and limitations of the recommender system [15]. In health care systems, this aspect gets the highest priority, as it helps in identifying the correct diagnosis and treatment of various diseases. A recommended medicine or therapy, which does not help in improving the health of a patient, undermines the quality of the recommender system. Explicit feedback need user ratings collected through surveys, or by asking them questions on social media. Implicit feedback is gathered by using likes and dislikes of a user by monitoring their actions. In this model, we have collected the patient's ratings for the recommendations given to them. The patients are asked to rate the treatment given to them on a scale of 1 – 5, where 1 being the most effective and 5 being the least effective treatment. The scale for ratings used in this study is shown in Table 2. Finally, recommendations to a query patient are assigned based on the selected similar patient ratings and their consequence weights. A final score is computed based on the identified consequences and ratings using,

$$F_S = R_i \cdot C_i; i \in k, \quad (6)$$

where F_S is the final computed score, k represents the set of nearest neighbors, R_i shows the rating provided by the i^{th} similar patient and C_i shows the consequential score of i^{th} patient. Final scores are computed

Table 2: Rating scale for treatments.

Ratings Scale	Definition (in terms of effectiveness)
1	Extremely effective
2	Very effective
3	Moderately effective
4	Slightly effective
5	Not effective

for all similar patients, and when these scores are distinct, the recommendation for patient with the highest score is assigned to the query patient. Whereas, in case of similar scores, multiple recommendations or treatment options are suggested to the query patient. The overall process used in the proposed methodology is presented in Algorithm 1.

Algorithm 1: Cluster based Modular Recommendation System.

Input : Set of supervised training data points X ; the number of clusters k
Output: N Recommendations

- 1 Partition data points X w.r.t disease classes;
- 2 Set iterations count C ;
- 3 **while** ($C > 0$) **do**
- 4 Apply K-Means clustering on each partition.
- 5 Find new mean μ_k for each cluster.
- 6 $C = C - 1$;
- 7 **end**
- 8 **if** $X_k < \mu_k$ **then**
- 9 Place X_k in lower sub-cluster $Cluster_{lower}$;
- 10 **else**
- 11 Place X_k in higher sub-cluster $Cluster_{higher}$;
- 12 **end**
- 13 **foreach** $Query_{patient}$ in D_{test} **do**
- 14 Find desired sub-cluster C_{sub}
- 15 Compute X_m of m similar users from C_{sub} ,
- 16 where $m = 9$;
- 17 **if** X_m are distinct **then**
- 18 Select top 3 similar users.
- 19 **else**
- 20 Select 3 users through majority voting.
- 21 **end**
- 22 Find $Recomm$ and $Consequence_{Weightage}$ of similar users;
- 23 Identify final score;
- 24 $F_S = R_i C_i$; where $i \in k$
- 25 **end**
- 26 Assign $Recomm$ with highest F_S to $Query_{patient}$;

4. Experimental Results and Discussion

The details of clinical dataset and experimental set up employed for evaluation of the proposed recommendation model, and the results achieved are presented in the following subsections.

Table 3: Characteristics of the dataset parameters.

Sr.No	Feature name	Feature Type	Feature units	Feature values
1	Patient-ID	Numeric	Numbers	1, 2, 3, ...
2	Age	Numeric	Years	20-100
3	Gender	Numeric	{0=Female; 1=Male}	0,1
4	Systolic Blood Pressure	Numeric	mmHg	80-250
5	Diastolic Blood Pressure	Numeric	mmHG	40-160
6	Fasting Blood Sugar	Numeric	mM	30-600
7	Cholesterol	Numeric	mg/dL	50-350
8	Oxygen Saturation	Numeric	%	70-100
9	Heart Rate	Numeric	Beats/Minute	50-200
10	Heart Beat Rhythm	Categorical	{0=REGULAR; 1=IRREGULAR}	0,1
11	ECG Readings	Categorical	{1=Normal; 2=left-ventricular hypertrophy; 3=ST-T wave abnormality; 4=MI}	1,2,3,4
12	Chest Pain Type	Categorical	{1=Typical; 2=Atypical; 3=nonanginal; 4=Asymptotic}	1,2,3,4
13	Chest Pain Location	Categorical	{0=Substernal; 1=otherwise}	0,1
14	Hypertension	Categorical	{0=No; 1=YES}	0,1
15	Smoker	Categorical	{0=No; 1=YES}	0,1
16	Family History of Diabetes	Categorical	{0=No; 1=YES}	0,1
17	Family History of Coronary Artery Disease	Categorical	{0=No; 1=YES}	0,1
18	BMI	Numeric	Kg/m2	10-50
19	Cardiac Disease Label	Categorical	{NCCP, Silent Ischemia, Angina, Myocardial Infarction}	1,2,3,4

4.1. Dataset

The clinical dataset of cardiac patients for this study was collected from the cardiology department of a renowned local hospital, with the consent of the participating patients. The collected data is anonymized for upholding patient privacy. The purpose of data collection was to propose a recommender model for treatment of four well known cardiovascular diseases i.e., angina, non-cardiac chest pain, ischemia, and myocardial infarction. The dataset comprises of 1070 records of a diversified range of cardiovascular disease severities, including critical, moderate, and mild. The data of critical disease severity was collected for patients admitted in the critical cardiac unit of the hospital, while other admitted patients and visitors for routine check up were targeted for moderate and mild severity cardiovascular diseases. A set up was created in the hospital with specialized equipment for measuring clinical tests required for this study. A medical team comprising of three medical experts supervised the entire process of data collection. The clinical features selected for this study were finalized under the supervision of these cardiologists and the collected data was verified by these experts. The disease diagnosis and suggested treatment was labelled by experts to remove ambiguities, which may lead to a failure of the clinical decision support system. To provide trust worthy recommendations to the patient, the rating of patients on their treatment were collected. The same patients were also observed during the follow up visits to measure the consequences of their treatment. This information is helpful in providing accurate medical recommendations to the desired patients. The treatments were rated on a discrete scale of 1 (extremely effective) to 5 (not effective) as shown in Table 2. The weighted consequences include four categories: therapy, death, medication with improvement, and medication without improvement.

The samples of dataset were collected using the stratified sampling technique [40] to target different sub-groups of patients including age, gender, smoking habits, and family history of diseases. The varying nature of data is helpful in providing a general and adaptive system to medical experts. Among the collected data, approximately 65% were male patients and 35% were female patients with age in the range of 20 to 100 years. A diversified range of age and gender is included, since these have significant importance in diagnosis of cardiovascular disease type and its treatment. A detailed description of the 18 features included in the dataset are presented in Table 3. The feature set contains a mixture of demographic, clinical and historical features. The proposed recommender system categorizes the patients into four disease clusters representing the disease diagnosis of SI, AN, MI and NCCP, subjected to the symptoms detected from the recorded clinical features. The output of the model is the treatment of disease suggested to the query patient as recommended by the medical expert to other similar patients suffering from the same disease type.

4.2. Evaluation Parameters

The clinical records collected for this study are randomly divided into training and testing sets for implementation of the proposed model. The train set contains 66% of the data instances (706 subjects),

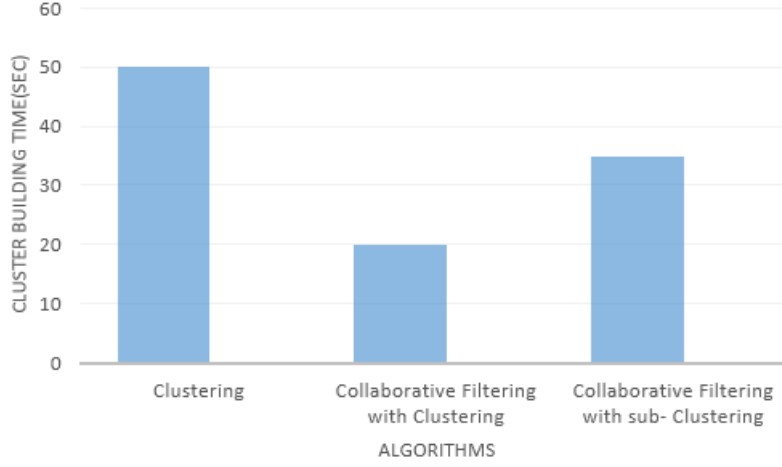


Figure 4: Cluster building time of different algorithms used.

whereas the remaining 34% of instances (304 subjects) are part of the test set. The training set is further divided into train and validation set with a 90 – 10 ratio, to accurately train the classifier parameters.

The disease classification using clustering and disease recommendation through collaborative filtering are compared with the actual classification and recommendations as labelled by medical experts to check the accuracy of the training model. The model is evaluated on the test set using four performance metrics i.e., mean absolute error (MAE), precision, recall and computational time. MAE [37] computes the absolute deviation between the predicted and the actual results and is calculated as,

$$MAE = \frac{1}{|D_{test}|} \sum_{i=1}^{|D_{test}|} |p_i - o_i|, \quad (7)$$

where $|D_{test}|$ is the total numbers of instances in the test set, p_i represents the predicted outcome for the i^{th} item in the test set and o_i represents the actual outcome for the i^{th} item. The proposed recommender system aims to reduce MAE score by reducing the difference between actual and predicted outcomes for disease recommendation. In machine learning based recommender systems, metrics such as precision and recall are used to measure the quality of the recommended items [41]. Precision is defined as the number of relevant recommendations produced in the set of total recommendations and is calculated as,

$$Precision = \frac{\text{correct recommendations}}{\text{total recommendations}}. \quad (8)$$

Recall is defined as the number of relevant recommendations retrieved successfully from the algorithm and is calculated as,

$$Recall = \frac{\text{correct recommendations}}{\text{relevant recommendations}}. \quad (9)$$

4.3. Results

In this section, performance of the proposed MC_bCR system is presented in terms of computational cost, cluster quality, cluster convergence, cluster building time, mean absolute error, and recommendation precision and recall. The results of clustering are compared with conventional k-mean approach and recommendation results are compared with traditional collaborative filtering approach.

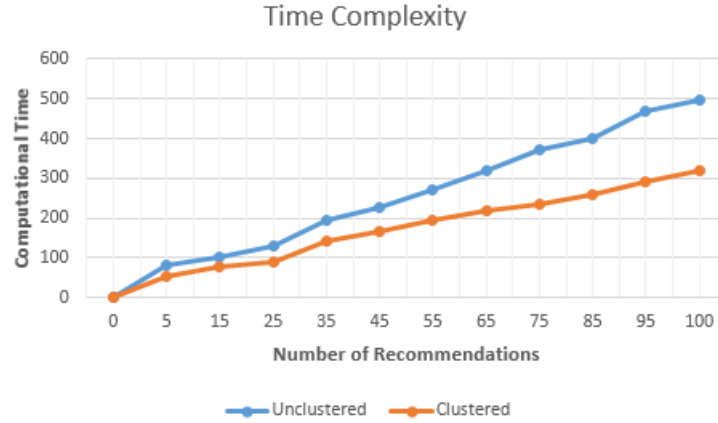


Figure 5: Comparison of proposed (MC_bCR) system with traditional collaborative filtering.

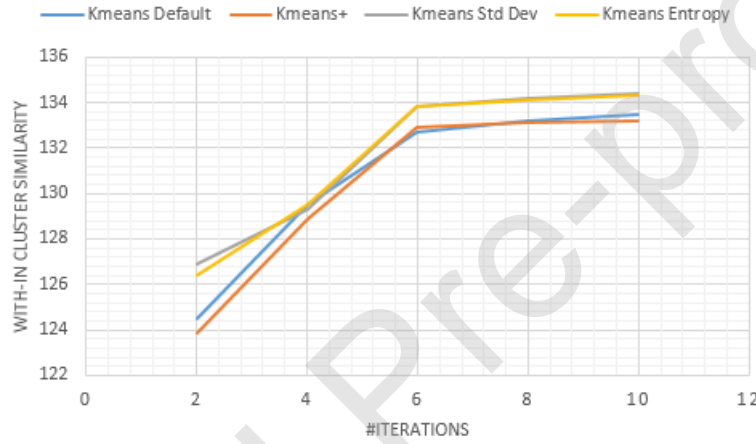


Figure 6: Quality of clusters with respect to iteration count.

4.3.1. Computational cost

In a clinical recommendation system, the accuracy of generating relevant recommendation on time is a critical task. Experiments are performed on clustered, sub-clustered, and non-clustered data to analyze the performance of the proposed system in terms of time required for building a cluster. The results are demonstrated in Figure 4, which shows that clustering with collaborative filtering takes less time as compared to the sub-clustered approach. However, as sub-clustering reduces the overall computational time for providing recommendations, there is a trade-off between the two approaches in terms of performance.

The model is analyzed and evaluated for measuring the total computational time in providing recommendations. The results of modular clustering approach are compared with non-clustered data to evaluate the efficiency of modular recommender system as shown in Figure 5. Results show that modularization has significantly improved recommendation efficiency. The clustered approach provides more recommendations in less time as compared to the traditional collaborative filtering approach. The results also show that the performance of the proposed system continuously improves, with an increasing number of recommendations.

4.3.2. Cluster Quality

A centroid selection approach based on standard deviation is proposed in this study, which differs from the default k-mean and k-mean+ approaches. Another approach using entropy measure is evaluated for

Table 4: Performance comparison of centroid selection approaches.

K-centroid selection method	Convergence rate	MAE
K-mean	9	0.278
K-mean ++	9	0.323
K-mean StdDev	7	0.173
K-mean Entropy	8	0.243

centroid selection, which works on the concept of maximizing the entropy between clusters. The entropy approach performs better than the traditional k-mean approach, but lags the standard deviation approach. The performance of the centroid selection approach is measured in terms of the cluster quality i.e., within cluster similarity between the data points and cluster centroids. The total within cluster similarity for each training user is computed using,

$$sim(U, L) = \sum_{l_j \in L} \sum_{u \in U^{l_j}} similarity(u, c_j), \quad (10)$$

where c_j is the centroid of the cluster l_j , U^{l_j} represents the users of the corresponding cluster l_j , and L shows the total number of clusters. Figure 6 shows the within cluster similarity of users for varying k-centroid selection approaches, and iterations. The graph shows that the standard deviation based centroid selection approach achieves the highest within cluster similarity. The number of iterations in Figure 6 represent the optimal number of iterations after which the algorithm starts converging. The convergence rate for all four centroid selection measures is shown in Table 4 and is calculated using,

$$con_{rate} = \sum_{itr=1}^n (sim(U, L)_{itr+1} - sim(U, L)_{itr}), \quad (11)$$

where con_{rate} is the rate of convergence and $sim(U, L)$ shows the similarity measures at different iterations. The selected approach results in a speedy convergence as seen in Figure 6. This shows that the initial centroid selection heavily effects the convergence rate and cluster quality, which ultimately improves the recommendation process.

4.3.3. Cluster Tuning

The study is conducted on supervised data, which is initially partitioned into four separate clusters representing each disease class i.e., AN, NCCP, MI, and SI. After partitioning the data, k-mean clustering is applied on each division to produce several sub-clusters. The process of selection of an optimal number of sub-clusters, iterations, and neighbors is presented in Figure 7. The optimal number of sub-clusters selected for this study are measured by the corresponding MAE as shown in Figure 7(a). The graph shows that MAE decreases by increasing the number of clusters up to nine clusters, after which it starts increasing. The highest cluster quality is achieved with nine sub-clusters. Similarly, the optimal number of iterations and neighbors are selected based on their corresponding MAE value. The results are shown in Figure 7(b), where the number of iterations are varied between 1 and 11, while other parameters are kept fixed. It is observed that 5 iterations give the minimum MAE value. The number of neighbors are changed from 3 to 15, and results in Figure 7(c) show that keeping 9 neighbors give the minimum MAE value. An increase in number of neighbors afterwards tend to increase the corresponding MAE.

4.3.4. Accuracy

The accuracy of a clinical recommender system is the core characteristic that must not be compromised at any cost. The cluster based modular system is analyzed in terms of precision and recall. The quality of recommendation depends on the size of the neighborhood selected for similarity. Experiments are conducted by implementing traditional collaborative and clustering methods, collaborative filtering with clustering method and the proposed MC_bCR system and evaluated using precision and recall. The results are presented

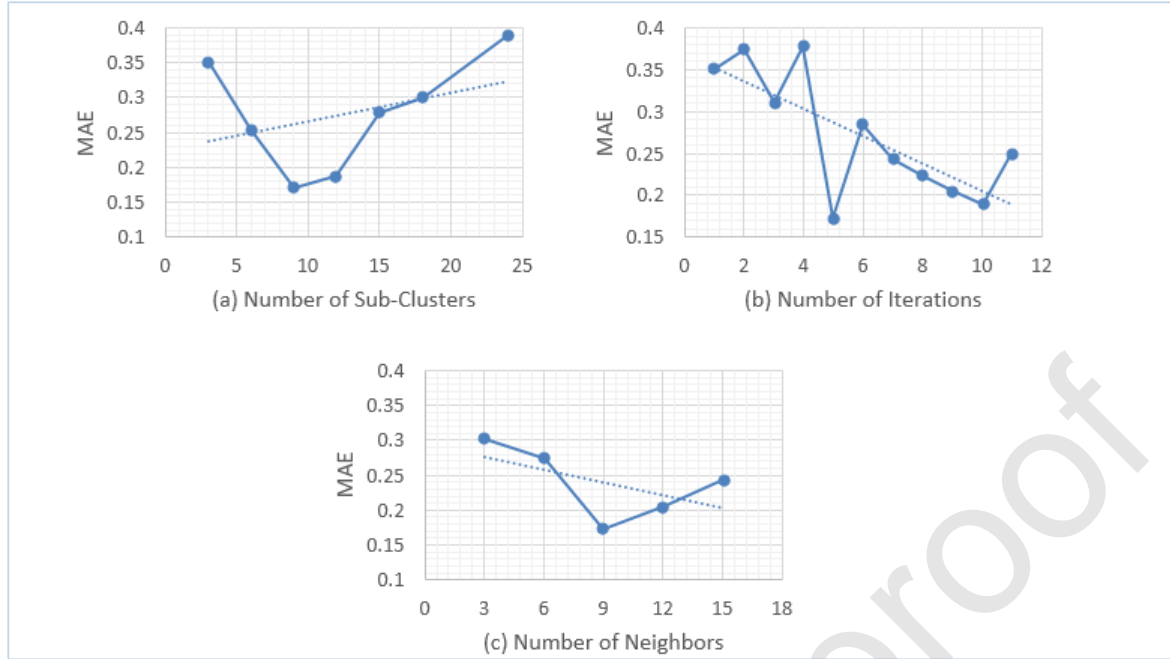


Figure 7: Selecting the optimal number of (a) sub-clusters, (b) iterations and (c) neighbors with respect to MAE.

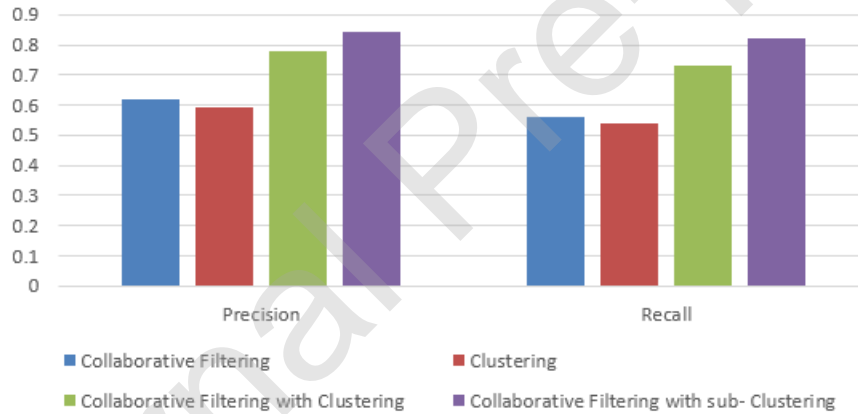


Figure 8: A performance comparison of the proposed approach with traditional collaborative filtering approaches in terms of precision and recall.

in Figure 8 which shows that the proposed technique achieves comparatively higher precision. This results in better quality of top N recommendations, as compared to other techniques. Similarly, it is observed that the proposed MC_bCR system outperforms all other traditional approaches in recall measure.

4.3.5. Majority Voting

A hybrid majority voting technique is used for prediction, which is a combination of multiple similarity computational algorithms. The results in terms of precision and recall for individual similarity measure, and the hybrid approach are shown in Figure 9. It is observed that the proposed hybrid approach for similarity measure gives higher performance in terms of precision and recall in providing fast recommendations.

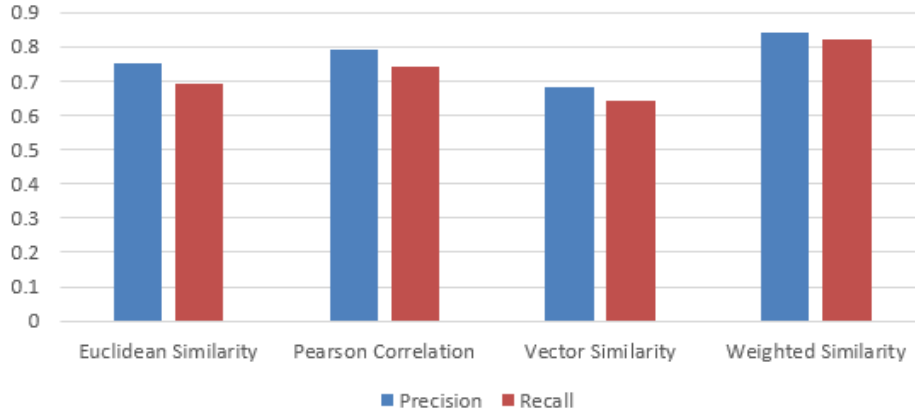


Figure 9: A performance comparison of the proposed weighted similarity approach with traditional approaches in terms of precision and recall.

4.3.6. Test-case Scenario

Consider a scenario, where a patient suffering from cardiovascular disease is admitted to the hospital for medical assistance. After collecting clinical tests, the type of disease is diagnosed and the query patient is added to a cluster belonging to the type of disease. The patient is then placed in the desired sub-cluster using k-mean clustering. In sub-cluster, the patient is further grouped into a smaller subset by comparing with mean of sub-clusters. The mean of the sub-cluster is computed by using Eq. 1, and then the test data points in the sub-cluster are compared with the mean point, such that the mean of sub-cluster is set as a threshold point. If $\mu(cluster)_k$ represents the sub-cluster mean point, x_q represents the query data point in the $\mu(cluster)_k$, and $D(x_q)$ represents the distance of query data point from the mean point, then the following conditions are checked,

1. If $D(x_q) > \mu(cluster)_k$, then place the query patient in a sub-group, which contain patients having a mean higher than the cluster mean such that,

$$\mu_{higher} = \sum_{i=j+1}^k x_i \quad (12)$$

2. Similarly, if $D(x_q) < \mu(cluster)_k$, then the query patient is placed in the lower mean sub-group such that,

$$\mu_{lower} = \sum_{i=1}^j x_i \quad (13)$$

The objective is to find a threshold value that maximizes the variance between clusters and minimizes the variance within clusters. If a patient has a small set of neighboring patients from which a set of k similar patients must be identified, the set of similar patients is selected using similarity measures and majority voting based technique. Three different scenarios for selection of weighted users from a set of nine users represented as $U_1, U_2, U_3, \dots, U_9$, are shown in Figure 10.

The first scenario is based on the concept of majority voting, where three users selected by three similarity measures have common similar users. As shown in case 1 in Figure 10, nine users are arranged in ascending order with respect to their votes or frequent occurrence in the similarity matrix. As U_1 is selected by all three similarity measures, therefore the user is selected for recommendation process. U_2 and U_3 have two votes each and is more than all other users on the list, hence these are also selected for the recommendation process.

In case 2, all distinct users appears in the similarity matrix, where the three similarity measures give different set of users hence making it difficult to select the most common users. The users are arranged in

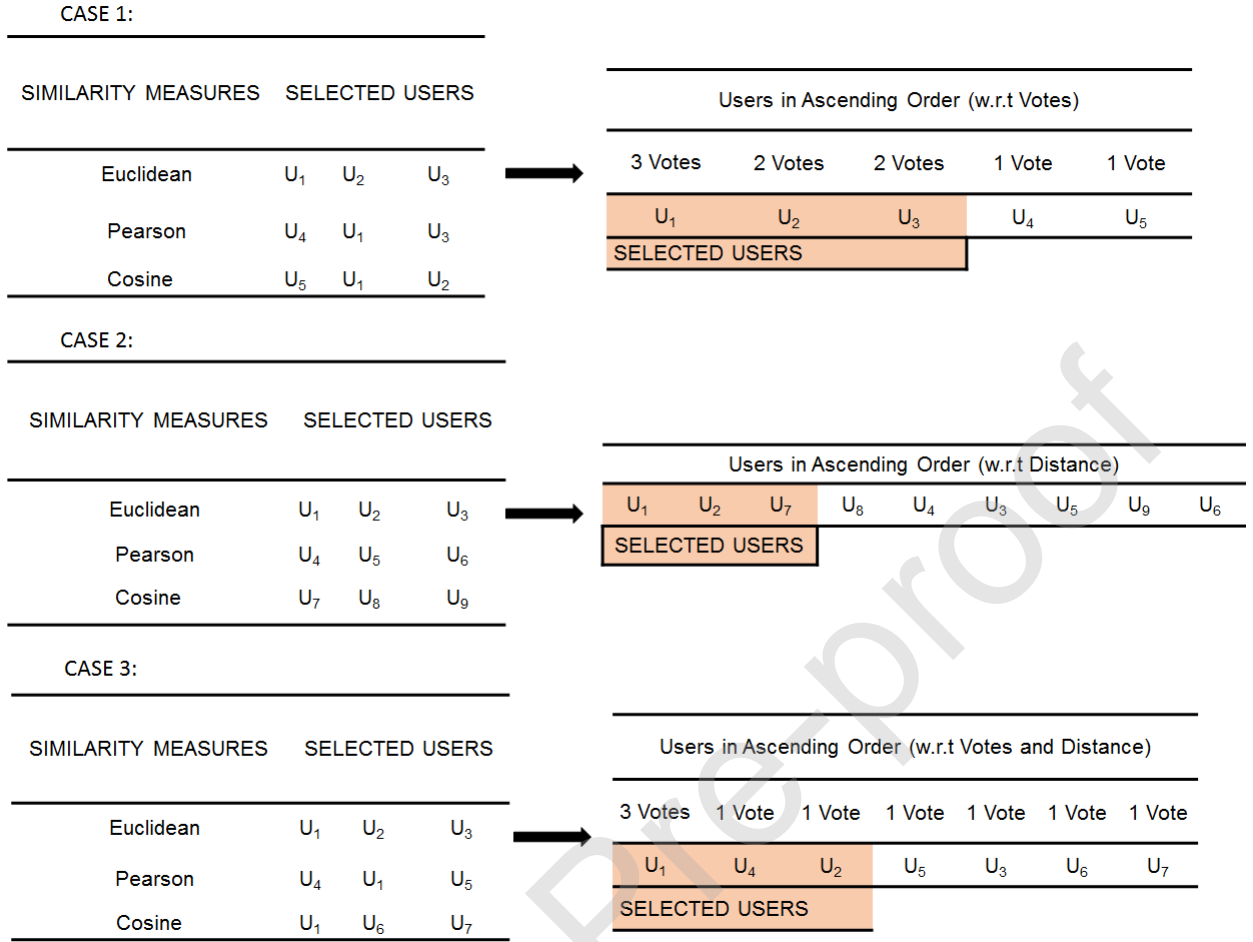


Figure 10: Test-case scenarios for selecting similar users.

ascending order according to their distances from the query user and the first three users on the list i.e., U_1 , U_2 , and U_7 showing least distances are selected for the recommendation process.

In case 3, there is at least one common user from all three similarity measures, whereas the remaining users are distinct. The most frequently occurring user is selected using majority voting, while the other two users are selected in correspondence to their distances from the query user. The selected users are analyzed based on their disease, assigned treatment, consequence of their treatment, corresponding weight assigned to their treatment and their ratings for the recommendation. All these factors are assembled to get a final score for each selected user. The user with highest final score is considered as the most similar user to the query patient and the corresponding treatment recommendation is recommended to the query users.

The process of calculating a final score is shown in Table 5. For example, three similar patients A , B and C are selected through the weighted user selection process. Patient A was recommended with medication and that patient went through a surgery. Patient B was recommended exercise with dietary modifications and that patient showed improvement. Patient C was recommended dietary modifications and that patient showed no improvement. The rating of patient A , B and C in response to the clinical recommendation is used in computation of the final score. Patient A rated the recommendation as 2, patient B rated the recommendation as 4, and patient C rated the recommendation as 1. Based on the ratings and consequences of similar patients a final score is computed using Eq. 6, and the highest weighted recommendation is

Table 5: Final score computation for assigning relevant recommendation.

Patient	Consequence	Weight	Recommendation	Patient's Rating	Final Score
A	Therapy	0.6	Exercise	2	1.2
B	Improvement	1	Exercise, Dietary Modifications	4	4
C	No-Improvement	0.2	Dietary Modifications	1	0.2

recommended to the query patient. The recommendation class of patient *B* has the highest score and is assigned to the query patient.

4.4. Discussion

The proposed MC_bCR system is designed to provide fast and robust treatment recommendations to similar query patients using a clustering based collaborative filtering technique. The results have shown that the proposed system achieves high scalability, while not compromising on accuracy. The system is reliable and accurate as it involves physicians in formulation and labelling of dataset. Moreover, most of the recommender systems in health domain are general systems and are not aimed for a specific disease, whereas the proposed system provide medical recommendations to cardiac disease patients suffering from AN, NCCP, SI, and MI. A similar system has been implemented for heart patients [20] using statistical techniques to recommend specific medical treatment to patients. However, the system is dependent on clinical features and is intended to be used by an expert physician only as it requires expert validation.

It is important to highlight here that a query patient can have multiple recommendation, as is the case for the query patient who is assigned the recommendation class of patient *B* comprising of two recommendations for the example shown in Table 5. The underlying data for this study has included patients suffering from any one of the four cardiac diseases considered. In future, those patients that have multiple diseases will be included to further enhance the performance of the recommender system. The whole process of recommendations is performed and verified under the supervision of clinical experts.

The proposed model allows grouping of data gathered under the supervision of medical expert, where grouping is performed at various levels to ensure fast and robust computation. This partitioning of data with respect to disease classes helped in introducing the concept of modularization, which allows the system to adapt changes at a sub-cluster level. Similarly, at the recommendation layer we can define different criteria for each cluster, so that each disease can be treated in its own different way. However, the system requires heavy processing for parallel computations in modularization and deep sub-clustering can introduce a trade-off in performance measures such as time and accuracy.

5. Conclusion

This paper is focused on improving the traditional collaborative filtering algorithm and proposes an improvised algorithm for recommending medical advice to cardiac patients. The proposed MC_bCR system is built by introducing clustering and sub-clustering in collaborative filtering to reduce the search domain for finding similar users and to improve the performance of recommender systems in terms of time and accuracy. The recommendations to a query patient are assigned from a set of labelled recommendation classes. The supervised dataset of cardiac patients suffering from four well known cardiovascular diseases including SI, AN, NCCP, and MI is used in this study. A query patient is first placed in the desired disease class partition, where a suitable cluster for the patient is found by measuring the distance of the query patient with all cluster centres of the underlying partition class. The recommendation process is initiated by finding similar users from the sub-cluster and finally, a decision is taken for clinical recommendation under the supervision of clinical experts. The experimental results show that the improvised collaborative algorithm performs better in terms of time and accuracy. Moreover, modularization using clustering requires less computations to make a recommendation. As a result, multiple active patients can be treated simultaneously in different clusters.

In future, the dataset will be extended by adding more patients and partitioned into clusters based on patient demographics, as most diseases are treated differently in different regions. The idea will also be extended to incorporate multiple doctors, as a treatment or clinical recommendation varies among different clinical experts.

6. Acknowledgement

The authors are highly obliged to the cardiology department of POF Hospital Wah, Pakistan for their support in data collection of cardiac patients. We are also thankful to Dr. Abdul Rashid khan, Dr. Ahsan Imran, Dr. Farhan-ul-Haq, and Dr. Muneeba Akbar for devoting their time and knowledge in the formulation of dataset used in this research work.

7. References

- [1] A. Mukherjee, A. Pal, P. Misra, Data analytics in ubiquitous sensor-based health information systems, in: Next generation mobile applications, services and technologies (NGMAST), 2012 6th international conference on, IEEE, 2012, pp. 193–198.
- [2] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, S. R. Das, S. de Ferranti, J.-P. Després, H. J. Fullerton, et al., Heart disease and stroke statistics 2016 update, *Circulation* 133 (4) (2016) e38–e360.
- [3] L. Anderson, N. Oldridge, D. R. Thompson, A.-D. Zwisler, K. Rees, N. Martin, R. S. Taylor, Exercise-based cardiac rehabilitation for coronary heart disease: Cochrane systematic review and meta-analysis, *Journal of the American College of Cardiology* 67 (1) (2016) 1–12.
- [4] R. Zenun Franco, Online recommender system for personalized nutrition advice, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, ACM, 2017, pp. 411–415.
- [5] A. Alwan, et al., Global status report on noncommunicable diseases 2010., World Health Organization, 2011.
- [6] E. P. Zeitler, J. P. Piccini, Remote monitoring of cardiac implantable electronic devices (cied), *Trends in cardiovascular medicine* 26 (6) (2016) 568–577.
- [7] D. Goldberg, D. Nichols, B. M. Oki, D. Terry, Using collaborative filtering to weave an information tapestry, *Communications of the ACM* 35 (12) (1992) 61–70.
- [8] D. Li, C. Chen, Q. Lv, L. Shang, Y. Zhao, T. Lu, N. Gu, An algorithm for efficient privacy-preserving item-based collaborative filtering, *Future Generation Computer Systems* 55 (2016) 311–320.
- [9] J. Bobadilla, F. Ortega, A. Hernandez, A. Gutiérrez, Recommender systems survey, *Knowledge-based systems* 46 (2013) 109–132.
- [10] B. Yang, Y. Lei, J. Liu, W. Li, Social collaborative filtering by trust, *IEEE transactions on pattern analysis and machine intelligence* 39 (8) (2017) 1633–1647.
- [11] M. A. Musen, B. Middleton, R. A. Greenes, Clinical decision-support systems, in: *Biomedical informatics*, Springer, 2014, pp. 643–674.
- [12] F. Gräßer, S. Beckert, D. Küster, J. Schmitt, S. Abraham, H. Malberg, S. Zaunseder, Therapy decision support based on recommender system methods, *Journal of Healthcare Engineering* 2017.
- [13] S. SA, Intelligent heart disease prediction system using data mining techniques, *International Journal of Healthcare & Biomedical Research* 1 (2013) 94–101.
- [14] E. Q. da Silva, C. G. Camilo-Junior, L. M. L. Pascoal, T. C. Rosa, An evolutionary approach for combining results of recommender systems techniques based on collaborative filtering, *Expert Systems with Applications* 53 (2016) 204–218.
- [15] F. Isinkaye, Y. Folajimi, B. Ojokoh, Recommendation systems: Principles, methods and evaluation, *Egyptian Informatics Journal* 16 (3) (2015) 261–273.
- [16] M. A. Ghazanfar, Robust, scalable, and practical algorithms for recommender systems, Ph.D. thesis, University of Southampton (2012).
- [17] X. Lu, Z. Huang, H. Duan, Supporting adaptive clinical treatment processes through recommendations, *Computer Methods and Programs in Biomedicine* 107 (3) (2012) 413–424.
- [18] M. Wiesner, D. Pfeifer, Health recommender systems: concepts, requirements, technical basics and challenges, *International journal of environmental research and public health* 11 (3) (2014) 2580–2607.
- [19] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, Z. Zhang, Expertrank: A topic-aware expert finding algorithm for online knowledge communities, *Decision Support Systems* 54 (3) (2013) 1442–1451.
- [20] A. Mustaqeem, S. M. Anwar, A. R. Khan, M. Majid, A statistical analysis based recommender model for heart disease patients, *International Journal of Medical Informatics*.
- [21] C. Wang, M. Xu, The research of doctors recommendation algorithm based on clustering and collaborative filtering, in: Proceedings of the 23rd International Conference on Industrial Engineering and Engineering Management 2016, Springer, 2017, pp. 233–237.
- [22] H. Jiang, W. Xu, How to find your appropriate doctor: An integrated recommendation framework in big data context, in: Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on, IEEE, 2014, pp. 154–158.
- [23] L. Fernandez-Luque, R. Karlsen, L. K. Vognild, Challenges and opportunities of using recommender systems for personalized health education., in: *MIE*, 2009, pp. 903–907.

- 480 [24] S. Bissoyi, B. K. Mishra, M. R. Patra, Recommender systems in a patient centric social network survey, in: Signal Processing, Communication, Power and Embedded System (SCOPEs), 2016 International Conference on, IEEE, 2016, pp. 386–389.
- [25] F. Narducci, C. Musto, M. Polignano, M. de Gemmis, P. Lops, G. Semeraro, A recommender system for connecting patients to the right doctors in the healthnet social network, in: Proceedings of the 24th International Conference on World Wide Web, ACM, 2015, pp. 81–82.
- 485 [26] H. Hu, A. Elkus, L. Kerschberg, A personal health recommender system incorporating personal health records, modular ontologies, and crowd-sourced data, in: Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, IEEE, 2016, pp. 1027–1033.
- [27] L. Duan, W. N. Street, E. Xu, Healthcare information systems: data mining methods in the creation of a clinical recommender system, *Enterprise Information Systems* 5 (2) (2011) 169–181.
- 490 [28] Y. Zhang, D. Zhang, M. M. Hassan, A. Alamri, L. Peng, Cadre: Cloud-assisted drug recommendation service for online pharmacies, *Mobile Networks and Applications* 20 (3) (2015) 348–355.
- [29] H. Schäfer, G. Groh, J. H. Schlichter, S. Kolossa, H. Daniel, R. Hecktor, T. Greupner, Personalized food recommendation., in: DMRS, 2015, pp. 21–24.
- 495 [30] S. Hassan, Z. Syed, From netflix to heart attacks: collaborative filtering in medical datasets, in: Proceedings of the 1st ACM international health informatics symposium, ACM, 2010, pp. 128–134.
- [31] B. M. Sarwar, G. Karypis, J. Konstan, J. Riedl, Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering, in: Proceedings of the fifth international conference on computer and information technology, Vol. 1, 2002.
- 500 [32] F. Darvishi-Mirshekarlou, S. Akbarpour, M. Feizi-Derakhshi, et al., Reviewing cluster based collaborative filtering approaches, *International Journal of Computer Applications Technology and Research* 2 (6) (2013) 650–meta.
- [33] W. Zhe, Research on personalized recommendation algorithm based on collaborative filtering and partition clustering, *DEStech Transactions on Engineering and Technology Research (tmcm)*.
- [34] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, *ACM SIGKDD explorations newsletter* 11 (1) (2009) 10–18.
- 505 [35] X. Wu, X. Zhu, G.-Q. Wu, W. Ding, Data mining with big data, *IEEE transactions on knowledge and data engineering* 26 (1) (2014) 97–107.
- [36] S. Guha, N. Mishra, Clustering data streams, in: *Data Stream Management*, Springer, 2016, pp. 169–187.
- [37] S. Zahra, M. A. Ghazanfar, A. Khalid, M. A. Azam, U. Naeem, A. Prugel-Bennett, Novel centroid selection approaches for kmeans-clustering based recommender systems, *Information sciences* 320 (2015) 156–189.
- 510 [38] Y.-S. Lin, J.-Y. Jiang, S.-J. Lee, A similarity measure for text classification and clustering, *IEEE transactions on knowledge and data engineering* 26 (7) (2014) 1575–1590.
- [39] H. Liu, Z. Hu, A. Mian, H. Tian, X. Zhu, A new user similarity model to improve the accuracy of collaborative filtering, *Knowledge-Based Systems* 56 (2014) 156–166.
- 515 [40] E. Liberty, K. Lang, K. Shmakov, Stratified sampling meets machine learning, in: *International Conference on Machine Learning*, 2016, pp. 2320–2329.
- [41] N. Polatidis, C. K. Georgiadis, A multi-level collaborative filtering method that improves recommendations, *Expert Systems with Applications* 48 (2016) 100–110.