Health Consumer Usage Patterns in Management of CVD using Data Mining Techniques

Devipriyaa Nagappan
Department of Computer Science
University of Auckland
Auckland, New Zealand
dnag692@aucklanduni.ac.nz

Prof. Jim Warren
Department of Computer Science
University of Auckland
Auckland, New Zealand
jim@cs.auckland.ac.nz

Dr. Patricia Riddle
Department of Computer Science
University of Auckland
Auckland, New Zealand
pat@cs.auckland.ac.nz

ABSTRACT

The Healthcare system is exposed to the increasing impact of chronic diseases including cardiovascular diseases; it is of much importance to analyze and understand the health trajectories for efficient planning and fair allotment of resources. This work proposes an approach based on mining clinical data to support the exploration of health trajectories related to cardiovascular diseases. As the health data are highly confidential, we aimed to conduct our experiments using a large, synthetic, longitudinal dataset, constituted to represent the CVD risk factors distribution and temporal sequence of events related to heart failure hospitalization and readmission.

This research work analyses and represents the temporal events or states of the patient's trajectory with the aim of understanding the patient's journey in the management of the chronic condition and its complications by using data mining techniques. This study focuses on developing an efficient algorithm to find cohesive clusters for handling the temporal events. Clustering health trajectories have been carried out by proposing an improved version of the Ant-based clustering algorithm. Insights from this study can potentially result in evidence that these approaches are useful in understanding and analyzing patient's health trajectories for better management of the chronic condition and its progression.

CCS CONCEPTS

• **Information systems → Data Mining → Clustering**

KEYWORDS

Chronic diseases, Cardio-vascular diseases(CVD), Health trajectories, Clustering

ACM Reference format:

Devipriyaa Nagappan, Jim Warren and Patricia Riddle. 2019. Health Consumer Usage Patterns in Management of CVD using Data Mining Techniques. In *Proceedings of Proceedings of the Australasian Computer Science Week Multiconference (ACSW '19), January 29--31, 2019, Sydney, NSW, Australia,* https://doi.org/10.1145/3290688.3290732

ACSW '19, January 29–31, 2019, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6603-8/19/01..$15.00
https://doi.org/10.1145/3290688.3290732

1 INTRODUCTION

Chronic diseases are diseases which persists for a long period with slow progressions such as diabetes, cancers, cardiovascular diseases, respiratory diseases, mental illness, chronic pain, chronic kidney disease and dementia[1]. Among these diseases, Cardiovascular Diseases(CVD) is one of the fast-growing diseases of the modern world. With the increasing impact of chronic diseases, analyzing and understanding the nature of health trajectories is much of importance for healthcare professionals to make decisions based on sufficient knowledge, and also analyze, understand the chronic care activities[2].

Data mining has a high potential for the healthcare industry to use healthcare data efficiently and also for system analytics to identify inefficiencies, best practices in order to improve care and reduce costs. Cluster analysis (CA) is one of the frequently used applied statistical technique to reveal the hidden structures and groups in large data sets[4]. However, this methodology is not widely used in large healthcare databases where the distribution of data is commonly skewed, which complicates analyses [4, 6]. Despite this challenge, CA may aid in identifying clusters of patients, who have experienced similar changes in health care before and after treatment and particular interest may lie in focusing attention on consistently high-risk groups or groups for whom health care has dramatically increased after a change in treatment.

The present study employs to generate a synthetic dataset related to cardiovascular diseases(CVD), from all the cardiovascular diseases, heart failure (HF) is one of the most severe clinical problems with a high rate of mortality, morbidity, and a high frequency of hospitalizations[45]. The goal is to create a synthetic cohort, which simulates the patient health care trajectories of HF hospitalization and readmission based on health statistics and census information of New Zealand, with realistic multivariate CVD risk characteristics[3,5,7].

The present study focuses on generating a synthetic heart failure hospitalization and readmission data set as a sequence of temporal events, based on the real-time data. The primary objective is to use the cluster analysis technique to the patient's trajectories, for analyzing and investigating the changes in health care patterns by developing an efficient clustering algorithm based on the Ant-based clustering technique with longest

common subsequence(LCS) distance to discover cohesive clusters and also to predict the future sequences of events accurately.

This paper is organized as follows; section 2 describes the related work, section 3 depicts the approaches and methods used to implement the research, section 4 discusses computational experiments and the analysis of the results. Finally, section 5 describes the conclusion and future direction of the research work.

2 BACKGROUND

2.1 CVD Research:

PREDICT is a web-based Cardiovascular disease risk assessment and management decision support system developed for primary care in New Zealand[36]. The main aim of developing this cohort was to validate the Framingham CVD score and generate new CVD Risk equations. The primary use of Framingham-Risk score[FRS] is to find the cardiovascular risk of an individual over a 10-year period. The FRS was initially developed using the data obtained from the Framingham Heart Study, to predict the development of coronary heart disease in ten years [37]. Reddy et al., [38] used the Framingham risk score in order to predict the risk of CVD in non-cardiac patients. A 5-year cardiovascular risk score was calculated using the Framingham risk equation with individuals having no previous cardiovascular history [39]. Lloyd-Jones et al., [40] investigated the Framingham risk score to estimate the 10-year risk of coronary heart disease (CHD) and to differentiate lifetime risk for CHD Patients. The work of Tsipouras et al., [41] represents a Treatment Tool, which is a decision support framework focusing on the management of heart failure (HF) patients. It is a web-based element with primary function including calculation of various risk scores along with the adverse events related to appearance prediction for treatment assessment (eg., hospital readmission prediction). The Treatment Tool provides two functionalities namely, risk scores calculation and treatment prediction based on adverse events appearance. Various researches in the past have been carried out related to CVD for predicting risk and adverse events, among all data mining techniques such as clustering technique, have a high potential for generating better results about predicting future events.

2.2 Cluster Analysis:

Many standard clustering methods have achieved positive results in solving clustering problems. However, each method has a few drawbacks, K-means is not effective in identifying cluster numbers and optimal initial partitions, is very sensitive to outliers and noise, and moreover is only applicable to digital datasets[8,14]. The hierarchical algorithm is impractical for working with the largest sizes of the database since its runtime increases with the growth of the data set, $O(m^2)$[8]. DBSCAN is sensitive to cluster datasets of widely varying densities[8].

Algorithms based out of swarm-based technique are developing as an alternative method to conventional techniques like hierarchical clustering and K-means. Deneubourg et al. initially developed Ant-based clustering and sorting algorithms, and later, Lumer and Faieta (1994) extended Deneubourg's initial work of the ant-based clustering algorithm to cluster data of a numerical type[25]. This study extended the algorithm applicability to a broad range of data types and clustering data mining.

Among the various swarm based clustering methods, Ant-based clustering is the most widely used and accepted technique. Deneubourg et al. initially developed Ant-based clustering and sorting algorithms, to explain different types of nature-inspired heuristics and also models the behavior of real ants [9,10,11,12]. The work of Deneubourg et al. mainly focuses on deriving a technique applicable to collective robotics and data analysis [9]. Lumer proposed an enhanced version of this algorithm and Faieta[25], which achieved good rankings when compared to other competing algorithms, but it creates small clusters failing to merge with other large clusters[14].

Several solution modifications were introduced, especially for addressing data mining problems such as noise elimination[15], clustering and topographic mapping[9]. There are also many algorithms to handle clustering problem such as improved ant-based clustering [16], improved entropy-based ant clustering[17]the , ant-means clustering[18], the combination of Ant-based clustering with fuzzy c-means and K-means algorithms proposed by Gu and Hall 2006[19], Kanade and Hall 2003, 2004[20, 21]; Monmarché et al., 1999[22] to improve the efficiency of the algorithm. According to [23,24], this algorithm has been modified with the simultaneous transportation of entire stacks of data items on the grid, which is further enhanced by Lumer and Faieta [25] by using short-term memory feature. Pheromone traces proposed by (Ramos and Merelo 2002[26]; Vizine et al., 2005b [27], (Montes de Oca et al., 2005[28]) to direct ant movement towards promising grid positions. Information exchange between agents and the replacement of picking and dropping probabilities by fuzzy rules have been used by (Schockaert et al., 2004a, 2004b [29,30] to improve the accuracy. In addition to that, due to its high flexibility and self-organization, Ant-based clustering has been widely applied in several other problematic applications such as web mining, text mining, DNA sequencing, speech processing, medical diagnosis, the stock market and so on[14].

2.3 LCS Distance for the Temporal Sequence of Events:

The LCS produces the most robust and intuitive correspondence between the data points. LCS is an improved version of the edit distance model; the primary purpose of this model is to match two given sequences by allowing them to stretch, without disturbing the format of the elements, but it also allows some elements to be unmatched based on the requirements [31]. Zhang et al., [32] analyse the treatment information of chronic kidney patients where hierarchical clustering based on LCS distance is introduced to cluster the temporal sequences of patients. Among the various approaches, LCS has been widely applied in biomedical research as a distance measure used in trajectory and protein sequence analysis [32]. DNA sequence clustering technique uses an advanced filtering method based on

the LCS to find sequence pair of the same type [33]. Chen Y. et al. proposed an improvised LCS method by adding semantic similarity feature, which can increase the accuracy of processing in Chinese disease mapping [34]. Park et al., [35] propose a classification method that adopts LCS for the similarity function for classifying abnormal human behavioral pattern.

3 METHODOLOGY

3.1 Data

The dataset used in our study has generated synthetic population aged 25 to 84 years with each 'synthetic person having a set of CVD risk factors based on New Zealand Framingham CVD risk equation and Canadian Framingham CVD risk equation. Synthetic population generation was carried out in three steps, first, a demographic framework generation from census data[3]; secondly, a non-demographic framework from health statistics[5]; finally hospitalization and readmission cohort is created based on an individual level CVD risk assessment, which is used to develop required temporal events using a Markov chain for all individuals generated in the previous steps. This framework allows the retention of correlations between variables to simulate the cohort on par with the actual population.

Each trajectory in the dataset must be an individual representation of patient information defined by the vector of binary and categorical variables. The data will be a set of N health trajectories corresponding to N distinct individuals, in which each trajectory is considered as a matrix with d columns. The d columns represent the patient's profile information and time series of length l_i . The variables required to model HF hospitalization include demographic details, such as age, gender, ethnicity, social deprivation were considered as the independent variable and generated randomly and non-demographic information such as smoking status, family history, systolic blood pressure (SBP), LDL, HDL, total cholesterol ratio, BMI, also generated randomly. Table 1 shows the distribution summary of demographic and non-demographic details of the synthetic population.

Diabetes is one of the risk factors of CVD, according to statistics the synthetic dataset represented this as 20% of the population having diabetes[5], and also it is associated with some of the other attributes which are considered as risk factors for diabetes. The process of generating a population with diabetes should select the individuals from the cohort, whose risk scores are high on the bases of the Framingham risk scores.

The patient's journey of hospitalization and readmission for HF over a 36 months period is represented as temporal events or states. The states or events for HF comprises of 'not admitted' 'admitted', 'intensive care units(ICU)', 'discharged', 'discharged with home care' and 'mortality' denoted as ('A', 'B', 'C', 'D', 'E' 'F') as characters in the cohort. The cohort was designed as three segments of patients with low, moderate and high-risk scores. Framingham risk scores have been generated for all individual trajectories, based on this, individuals with low-risk considered to be healthy and moderate- risk individuals having very less chance of going to the state of intensive care unit and mortality. The individuals with high-risk scores represented as the highest chance of going to the ICU and mortality states.

Table:1 Summary of demographic and non-demographic details

Variables		Age	
		25-64	65-84
Number of Individuals		6695	3305
Gender	Male(%)	33.2	16.73
	Female(%)	33.75	16.32
Ethnicity(% of age)	Asian	9.2	4.8
	European	37.8	20.12
	Māori	9.1	4.98
	Pacific	8.9	5.1
NzDep(%)	1	11.82	6.18
	2	10.35	5.7
	3	11.2	5.8
	4	13.28	7.72
	5	18.35	9.6
Smoking status(%)	Current	21.55	11.81
	Former	21.65	11.81
	Never	21.8	11.38
Diabetes(%)		9.15	10.85
Family History(%)		32.9	17.65
Blood pressure(%)	Normal	15.4	7.9
	Hypertension	5.8	11.5
	Hypotension	21.28	38.8
Total Cholesterol(%)	Low	17.15	18.44
	Normal	30.42	14.57
	High	12.43	6.99
BMI(%)	Under weight	16.12	9.4
	Normal	16.28	8.36
	Over weight	16.43	8.4
	Obesity	16.17	8.84

Figure 1 represents the age and diabetic distribution of synthetic data set, which shows that the older are more diabetic than, the younger adults. Figure 2 describes the distribution of average Framingham risk scores for different age ranges. Age is one of the most important risk factor, when the age increases, according to the age group the risk scores also increases(figure.2).

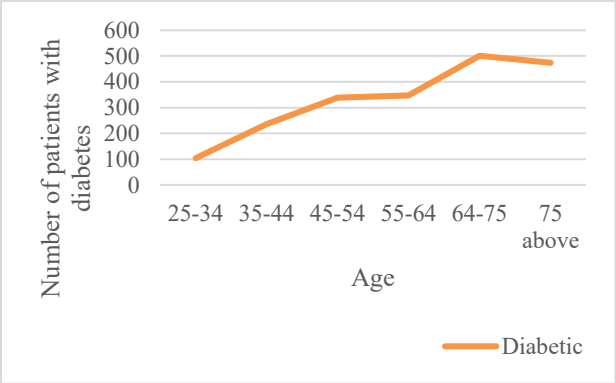

Fig: 1 Distribution of age and patients with diabetes

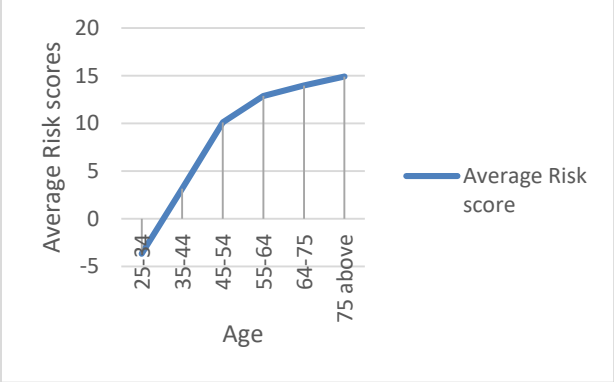

Fig. 2 Distribution of Age and Average risk scores

3.2 Distance Measure

One of the most critical aspects of clustering is to choose the appropriate distance measure, which is used to compare the overall features of data. Computing the similarity between sequences is a fundamental challenge for healthcare dataset. In the past similarity between the time series is modeled using various techniques which includes Euclidean distance, Dynamic time wrapping (DTW) distance, Edit distance and Longest Common Subsequence(LCS). Choosing the Euclidean distance as the similarity model is unrealistic since its performance degrades rapidly in the presence of noise. Dynamic Time Warping (DTW) has been used so far for one-dimensional time series. Euclidean matching completely disregards the variations in the time axis, while DTW performs excessive matchings, which distorts the exact distance between sequences. The fundamental purpose of LCS is to match the two given sequences by allowing them to stretch, without disturbing the format of the elements, but it also allows some elements to be unmatched based on the requirements[44].

3.3 Clustering

3.3.1 Proposed Model

The proposed work introduces some modified features to the version of the Ant-based clustering algorithm, which is intended to improve the cluster quality and to handle the temporal sequence of data. The proposed algorithm includes two phases: the clustering phase and the merging phase. The primary task of the first phase is to cluster the dataset where ants pick up data items from the grid and attempt to drop the data into the most similar classes. This phase forms clusters and assigns most of the data items into groups, but it creates many small clusters. The small clusters have been overcome during the merging phase, where small and similar classes are joined together to form new homogenous clusters.

I. Clustering Phase

The environment of this algorithm consists of randomly placed high-dimensional data objects in a bi-dimensional grid. The size of the grid has to be large enough for the ants to roam and find data objects. The clusters formed are affected by the original spatial distribution of the data objects. The process begins with ants picking up data objects based on low density and similarity. The ants then try dropping the data objects at a suitable location in which similar objects exist already[13]. Assuming that an unloaded ant comes across an element, the probability of picking that element increases when low-density and similarity. The ants are then dropping that element at a suitable location where similar elements exist already. Accordingly, the probability of picking an element i is defined as

$$P_{pick(i)} = \left(\frac{k_p}{k_{p+f(i)}}\right)^2 \quad (1)$$

Where k_p is a constant and $f(i)$ a local estimation of the density of elements and their similarity. In the same way, the ant will drop a carried element should increase with the density of similar elements in its surrounding area.

$$P_{drop\ (i)} = \begin{cases} 2f(i) & f(i) < k^d \\ 1 & otherwise, \end{cases} \quad (2)$$

where k_p and k_d are constants, $f(i)$ is a neighborhood function.

$$f(i) = max(0, \frac{1}{\sigma^2}\sum(1 - \frac{d(i,j)}{\alpha}) \quad (3)$$

where, $d(i, j)$ is a measure of the similarity between data points i and j α [0, 1] is a data-dependent scaling parameter, and σ^2 is the size of the local neighbourhood.

II. Merging Phase:

This phase directly addresses the problem of generating the small clusters, in which data items from the same class occupy two different areas. We propose a merging mechanism to determine whether data items of two clusters should be combined into one cluster. According to Chaoji and et al.[43], merging any two clusters, which are closer to each other in the distance space

and also that have the highest similarity. We focus on defining a similarity threshold value, in order to determine the merging of two clusters into one cluster,

$$Dist\_thrs_{(i,j)} = m * Max(Avg\_dist_{(i)}, Avg\_dist_{(j)}) \quad (4)$$

Where m is a real number that is user input greater than one, $Avg\_dis_{(i)}$ is a function that calculates the average distance among all the data items in the cluster *i*. If the two clusters minimum distance is bigger than the value of *Thresh dis*(*i, j*), it should not be considered for merging anymore. Those two clusters considered for merging, if the selected clusters combined and result in the maximum value of silhouette index, then the merge should move ahead.

Pseudo code for merging phase

```
1  // Num_of_Clust: User input for selecting Number of clusters
2. // dist_thrs: user input for calculating the distance threshold value
3. //  Ci : ith cluster on the grid
4. max_no_clust = Current number of clusters in the grid after the clustering phase
5. for i=1 to max_no_clust && max_no_clust ≤ num_of_Clust
6.    for j=1 to max_no_clust
7       If (i != j)
8.         min_dist = minimum distance between cluster Ci and Cj ;
9.       avg_dist = minimum distance between  data items in cluster Ci ;
10.          avg_distj = minimum distance between data items in cluster Ci;
11.       dist_thershold = dist_thrs   x   max(avg_dist i, avg_dist j)
12.       // if the minimum distance is small, then it is considered to have a chance to merge.
13.       If ( min_dist ≤ dist_thershold)
14.          // check the similarity
15.          silhouette(Ci, Cj) = Ci, Cj considered to be a single cluster, then Silhoutte_Index has calculated
16.          If (silhouette = maximum value)
17.             Merge Ci, Cj into one cluster.
18.          End if
19.       End if
20.    End for
21. End for
22. End Procedure
```

III. Cluster Evaluation:

A common approach consists of running the clustering algorithm for different numbers of clusters and computing the validity index for evaluating the quality of the clusters. The number of clusters with the best index value is selected as optimal. These kind of techniques are useful in cases, where there is no prior knowledge related to the nature of the clusters. We considered some of the most commonly used validity indexes such as the Silhouette index[46], Dunn index[48], Davies-Bouldin(DB) index[47]. Silhouette index refers to a method based on, how tightly the data in each cluster are grouped, it achieves high values for nicely clustered data. It considers both cohesion and separation of data points for evaluating the clusters. The Dunn Index and DB Index are somewhat similar since both depend on the relative size of inter-cluster and intra-cluster distance. Dunn Index determines the minimum ratio between cluster diameter and inter-cluster distance for a given partitioning. DB Index determines the average similarity between each cluster and its most similar one, averaged over all the clusters. The best Dunn and Silhouette index corresponds to the highest value, while the best DB index corresponds to the lowest value.

IV. Prediction:

The primary objective of prediction in this part of the architecture is to classify patients trajectories based on the sequence of events and also predict patients future events. In order to classify patients trajectories, we look for clusters that include different sequences of events that happened in a particular time frame. For this purpose, we propose an approach using the LCS algorithm to classify a current set of events. According to the clustering algorithm in the previous section, there is a set of clusters $nc = (nc_1, nc_2, \ldots\ldots, nc_n)$. The next process after clustering algorithm is to identify different combinations of event sequence for each cluster, where $np_i = (p_1, p_2, \ldots\ldots\ldots, p_k)$ is a set of k events. This kind of sequences has been identified using the LCS algorithm.

4 RESULTS AND DISCUSSION

Experiments were conducted using a large synthetic data set with a population of 10,000 individuals aged from 25-84 years, representing demographic and non-demographic characteristics, and also including the patient's journey of HF hospitalization and readmission status over a 36 month period. In each of the K-fold cross-validations, the data set divided into training (90%) and evaluation (10%) datasets. The training set is used to generate the models based on clustering, while the evaluation set is used to test the generated model. The training set is used to generate the models based on clustering, while the evaluation set is used to test the generated model.

Clustering is then performed using an improved Ant-based clustering algorithm for the synthetic HF dataset and compared with standard k-medoids, and hierarchical clustering algorithms with LCS similarity measure as well as the original version of the Ant-based clustering algorithm. The ant-based clustering algorithm is an unsupervised procedure which automatically determines the optimal number of clusters. The improved Ant-based clustering has been restricted to form a specified number of clusters (like other clustering methods such as the Hierarchical and K-medoids). This number can be determined based on the nature of the data set and/or by evaluating the quality of clustering results and then specified.

The algorithm records the result of the analytical measures (Silhouette, Dunn, DB index) and runtime. At the end of the run, the algorithm calculates the mean(μ) and standard deviation(σ) for all previous evaluation measures. Some of the parameters need to be set for this algorithm such as the number of clusters, number of iterations, number of ants and the real number used for the threshold value of the merging phase. In order to find the best number of clusters; we used validity indexes discussed in section 3.3.1. In table 2 we show these indexes as a procedure to find the

number of clusters K varying from 3 to 6. It shows that K=3 is the optimal number of clusters, in which Dunn and silhouette indexes are maximized, while the DB index reaches the minimum value (with the sole exception of the Silhouette Index for original version Ant-based Clustering).

Table 2. The Silhouette, Dunn, DB index for synthetic heart failure data set

Algorithms	Number of Clusters	Silhouette Index	Dunn Index	DB Index
Ant based Clustering (improved)	**3**	**0.433**	**0.658**	**1.008**
	4	0.381	0.511	1.67
	5	0.296	0.435	2.09
	6	0.271	0.403	2.68
Hierarchical Clustering	**3**	**0.483**	**0.708**	**0.94**
	4	0.403	0.56	1.04
	5	0.305	0.463	2.82
	6	0.287	0.363	2.98
K-medoids	**3**	**0.385**	**0.559**	**0.577**
	4	0.364	0.425	1.38
	5	0.205	0.354	2.84
	6	0.187	0.333	3.68
Ant-based Clustering (original)	**3**	**0.171**	**0.359**	**1.36**
	4	0.183	0.285	1.38
	5	0.175	0.167	2.814
	6	0.161	0.169	2.168

Table3: 10-fold cross-validation of validity indexes and runtime

Algorithms	Silhouette	Dunn Index	DB Index	Runtim(seconds)
K-medoids	0.227 ± 0.34	0.596 ± 0.56	0.57 ± 0.45	600
Hierarchical clustering	0.483 ± 0.02	0.706 ± 0.0012	0.94 ± 0.0016	18000
Ant-based (improved)	0.432 ± 0.16	0.658 ± 0.25	1.010 ± 0.35	7740
Ant-based (original)	0.187 ± 0.48	0.197 ± 0.67	2.168 ± 0.57	5220

The ten-fold cross-validation results generated for the HF dataset and mean, and standard deviation of DB Index, Dunn Index and Silhouette index for k-medoids, hierarchical clustering, and ant-based clustering are shown in table 3. It shows that the hierarchical clustering algorithm is a standard algorithm which gives better results, but it is confined to small data sets: for a large population of health records, it is a time-consuming process to find the distance between every record. The speed obtained by the K-medoids clustering algorithm is better than the other clustering algorithms, but the limitations are the results are not stable. Considering all this, the improved Ant-based clustering algorithm gives better results, and it can provide useful results for large sets of population health records.

Since the clustering algorithms are unsupervised; they generate different results for every run of the 10-fold validation, and the generated clusters are investigated and sorted manually for comparison. Figure 3 clearly shows that patient with low scores are grouped in cluster1, whereas high-risk patients are clustered in cluster2, and the moderate risk population is grouped in cluster3. The detailed summary description and cluster-wise comparison of algorithms for the frequency of temporal events after clustering are shown in table 4. From the table, it is observed that the Cluster 1 has a substantial amount of healthy people and their state of hospitalization remain not admitted and it has little over to a lesser amount of people in the others states especially, Intensive care ('C') and mortality ('F').

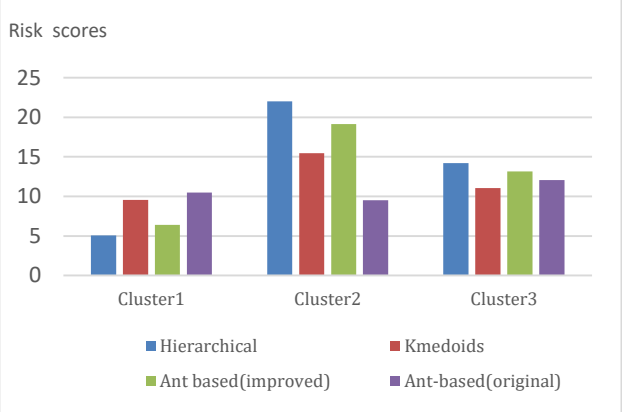

Fig:3 Distribution of Clusters based on the Average Risk Scores

Cluster 2 consist of high-risk patients group with more patients in the ICU('C'), discharged with home care ('E') and mortality ('F') states. Whereas Cluster 3 indicates the moderate risk population with more frequency in the admitted state ('B'). In these terms, the Hierarchical clustering and improved Ant-based clustering algorithms appear to provide better results than the other two algorithms in terms of stronger concentration of state frequencies for each cluster.

Each sequence of events np_i in the evaluation set is divided into two parts. The first, n sequences in the np_i are used for generating the prediction, whereas, the remaining part of the np_i is used to evaluate the prediction results. The first part of np_i, has been classified based on the clusters and concerning the sequence of events, np_i from co-occurrence matrix M is denoted as anp_i,

which is used to predict the sequences of events. The set of the predicted list can be compared with the remaining $|np - n|$ in np and denote this np as $eval_{np}$. The accuracy has been generated for $P(anp_i)$ as follows.

$$Accuracy = \frac{|\boldsymbol{p(anp_i)} \cap \boldsymbol{eval_{np}}|}{|\boldsymbol{p(anp_i)}|} \quad (5)$$

The evaluation set is taken into account, which consists of the patient's demographic, non-demographic information and temporal states or events. Heart failure hospitalization for 30 months of the period is taken for evaluation and the remaining six months of the period are predicted using LCS algorithm.

Table 4: Cluster-wise comparison of algorithms for the frequency of temporal events

Cluster1						
Algorithms	**A**	**B**	**C**	**D**	**E**	**F**
K-medoids	22.6 ± 9.02	4.8 ± 2.66	2.01 ±1.8	3.38 ± 2.4	3.87 ± 4.3	1.15 ± 1.28
Hierarchical	33.7 ± 0.6	0.54 ± 0.05	0.04 ± 0.008	0.54 ± 0.03	0.6 ± 0.24	0
Ant-based (improved)	28.69 ± 2.99	2.7 ± 0.95	0.78 ± 0.49	1.15 ± 0.38	2.26 ± 0.84	0.34 ± 0.37
Ant-based (Original)	19.6 ± 8.02	4.05 ± 3.66	2.05 ± 4.6	3.18 ± 4.38	3.87 ± 5.38	3.15 ± 1.42
Cluster 2						
Algorithms	**A**	**B**	**C**	**D**	**E**	**F**
K-medoids	7.68 ± 5.6	6.6 ± 4.6	7.57 ± 5.6	3.35 ± 2.16	7.05 ± 2.07	1.15 ± 1.5
Hierarchical	5.83 ± 0.6	5.53 ± 0.41	7.4 ± 0.035	4.1 ± 0.07	6.9 ± 0.02	4.09 ± 0.13
Ant-based (improved)	5.68 ± 0.63	8.7 ± 1.05	7.07 ± 1.07	3.68 ± 0.17	6.57 ± 0.48	4.03 ± 1.25
Ant-based (Original)	15.6 ± 5.8	4.5 ± 3.66	4.9 ± 3.27	4.45 ± 4.13	7.9 ± 2.19	1.32 ± 1.3
Cluster 3						
Algorithms	**A**	**B**	**C**	**D**	**E**	**F**
K-medoids	10.68 ± 5.6	8.9 ± 2.05	3.83 ± 2.39	2.73 ± 1.7	5.21 ± 2.59	1.7 ± 2.45
Hierarchical	7.7 ± 0.12	11.3 ± 0.04	2.63 ± 0.24	3.58 ± 0.2	7.4 ± 0.3	0.53 ± 0.003
Ant-based (improved)	11.6 ± 5.9	10.27 ± 2.64	4.83 ±1.15	2.59 ± 0.8	7.16 ± 1.4	1.11 ± 0.94
Ant-based (Original)	20.6 ± 8.05	4.86 ± 3.79	3.49 ± 2.06	5.73 ± 3.7	3.21 ± 2.1	1.03 ± 2.05

The accuracy of the prediction is compared with standard Hidden Markov model (HMM) and Recurrent Neural Network(RNN). Figure 4 represents the 10 runs of the algorithm for prediction accuracy. It shows that, comparing with HMM and RNN, the LCS algorithm gives moderate result, with prediction accuracy between 0.45 to 0.68

Learning temporal event patterns from population health data (e.g. as may be collected automatically in electronic health records, EHRs) allows can provide information to help in formulating various improvements for managing chronic conditions. This may be through predicting adverse events or better understanding the relationship of interventions and disease progression. Moreover finding patterns in temporal events where improvements take place in patients health, may help the clinicians to identify promising care models for the future course of action. Exploring temporal sequence of events also allows us to identify significant differences in the health record distribution of patients progression stages like, hospitalization, medications, treatment procedures and so on. And examining differences in demographics of cluster membership, may give us insight on the equity of a healthcare system.

Learning temporal event patterns from population health data (e.g. as may be collected automatically in electronic health records, EHRs) allows can provide information to help in formulating various improvements for managing chronic conditions. This may be through predicting adverse events or better understanding the relationship of interventions and disease progression. Moreover finding patterns in temporal events where improvements take place in patients health, may help the clinicians to identify promising care models for the future course of action. Exploring temporal sequence of events also allows us to identify significant differences in the health record distribution of patients progression stages like, hospitalization, medications, treatment procedures and so on. And examining differences in demographics of cluster membership, may give us insight on the equity of a healthcare system.

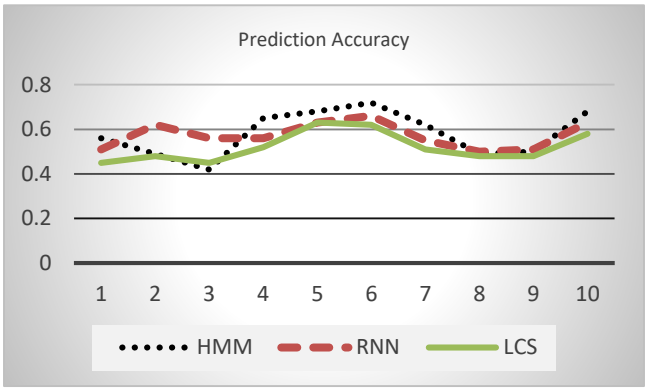

Fig 4: Prediction Accuracy for 10 Runs

The present research work mainly focuses on temporal sequence of events on CVD as an example of the management of chronic conditions and with this finding patterns in a patient's journey. Since publicly available data sets in UCR and MIMIC III primarily focus on radiology and in-patient, cardiology they lacked the relevant data related to the temporal sequence of events for long-term management CVD. Hence, we decide to make a synthetic dataset more closely related to the real world data sets in this domain, such as those collected by the national health system as in New Zealand or Australia..

It is considered that the Ant-based algorithms are more efficient in mining temporal events especially in identifying the user behavior pattern in web mining and text mining. Inspired by this paper explored whether they would be the better algorithm for finding the patients health phases in their health trajectory. The principal motivation and key future research for this work is to run the improved ANT based algorithm in New Zealand VIEW(Vasucular Informatics and Epidemiology using the Web) data. The intent is to use the VIEW linked decision support and national data collections database for finding the patient's journey in managing CVD to gain more insight about the different stages of CVD, Toward this goal, for the present research, we decided to simulate a dataset related to CVD based on the literature and health statistics, thus achieving something more relevant to the intended future direction.

5 CONCLUSION

Data mining applications are developed to identify different patterns in tracking the states of chronic disease, high-risk patients, design appropriate interventions, reduce the number of hospitalization and readmission. The primary objective of generating the synthetic data set is to check the performance of the new techniques before applying them to the actual data. Several experiments are conducted in the synthetic dataset by applying our improved Ant-based clustering algorithm, and it is compared with k-medoids and hierarchical clustering algorithms with LCS similarity measure. The runtime recorded by K-medoids is the shortest, but this method has a limitation that its performance is not stable when compared with the other approaches. The runtime of the hierarchical clustering algorithm is higher than other clustering algorithms, but the algorithm is stable in processing, and also gives better performance when compared with other algorithms. However, its runtime will become impractical for the large scale of population health records. The primary objective of this research work is to develop an alternative approach, on the basis of stability, performance, and run duration. In terms of this balance of criteria, the improved Ant-based clustering algorithm can provide better results for population health records.

Moreover, the LCS algorithm is also helpful in discovering health consumers' patterns to predict the future sequence of temporal events. The accuracy of the prediction process based on these clusters is somewhat less than that achieved with hidden Markov model(HMM) and recurrent neural network(RNN) – two methods ideally suited to the prediction of temporal sequences. However, the prevalence of states in sequences among the clusters differs significantly. The future result may reveal additional ways this sequence-based clustering can have a positive impact on the understanding of health consumer temporal event patterns.

The current research work has the following limitation mainly in the synthetic dataset which does not simulate all the features of the hospitalization and readmission; it covers some of the features and minimal temporal events of the research work. In the real dataset, there are various other opportunities to find a richer association of factors, sequences, and events happening in heart failure hospitalization and analyzing events pattern from the actual events. The principal motivation and key future research for this work are to run the improved ANT based algorithm in New Zealand VIEW(Vascular Informatics and Epidemiology using the Web) data. The intent is to use the VIEW linked decision support and national data collections database for finding the patient's journey in managing CVD to gain more insight about the different stages of CVD.

REFERENCES

[1] Song, S., Warren, J., & Riddle, P. (2014, May). Profiling Cardiovascular Disease Event Risk through Clustering of Classification Association Rules. In Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on (pp. 294-299). IEEE.

[2] Jay, N., Nuemi, G., Gadreau, M., & Quantin, C. (2013). A data mining approach for grouping and analyzing trajectories of care using claim data: The example of breast cancer.BMC Medical Informatics and Decision Making,13(1). doi:10.1186/1472-6947-13-130

[3] Ministry of Health (2015) Mortality and Demographic data 2013 provisional. Wellington: Ministry of Health

[4] Liao, M., Li, Y., Kianifard, F., Obi, E., & Arcona, S. (2016). Cluster analysis and its application to healthcare claims data: A study of end-stage renal disease patients who initiated hemodialysis. BMC Nephrology,17(1). doi:10.1186/s12882-016-0238-2

[5] Statistics NewZealand. (2013) Census data user guide. Wellington, NewZealand statistics., Newzealand.

[6] Dilts, D., Khamalah, J., & Plotkin, A. (1995). Using cluster analysis for medical resource decision making. Medical Decision Making, 15(4), 333-346.

[7] Bosomworth, N. J. (2011). Practical use of the Framingham risk score in primary prevention: a Canadian perspective. Canadian Family Physician, 57(4), 417-423.

[8] Nordhausen, K. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. International Statistical Review,77(3), 482-482. doi:10.1111/j.1751-5823.2009.00095_18.x.

[9] Handl, J., Knowles, J., & Dorigo, M. (2006). Ant-based clustering and topographic mapping. Artificial life, 12(1), 35-62.

[10] Bonabeau E, Dorigo M, Theraulaz G. (1999). Swarm intelligence, From natural to artificial systems. New York, Oxford University Press.

[11] Deneubourg, J.-L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., & Chre´tien, L. (1991). The dynamics of collective sorting: Robot-like ants and ant-like robots. In Proceedings of the First International Conference on Simulation of Adaptive Behavior: From animals to animats 1 (pp. 356 –365). Cambridge, MA: MIT Press.

[12] Dorigo, M., Bonabeau, E., Theraulaz, G. (2000). Ant algorithms and stigmergy. Future Generation Computer Systems, 16(8), 851 –871.

[13] Engelbrecht, A. P. (2006). Fundamentals of computational swarm intelligence. John Wiley & Sons.

[14] Martens, D., Baesens, B., & Fawcett, T. (2011). Editorial survey: swarm intelligence for data mining. Machine Learning, 82(1), 1-42.

[15] Zaharie, D., & Zamfirache, F. (2005, September). Dealing with noise in ant-based clustering. In Evolutionary Computation, 2005. The 2005 IEEE Congress on (Vol. 3, pp. 2395-2401). IEEE.

[16] Boryczka, U. (2008). Ant clustering algorithm. Intelligent Information Systems, 1998, 455-458.

[17] Weili, Z. (2009, July). An improved entropy-based ant clustering algorithm. In 2009 WASE International Conference on Information Engineering (pp. 41-44). IEEE.

[18] Hameurlaine, M., Moussaoui, A., & Cherroun, H. (2012).Ant means: A new hybrid algorithm based on ant colonies for complex data mining. International Journal of Computer Applications (0975–8887) Volume.

[19] Gu, Y., & Hall, L. O. (2006). Kernel-based fuzzy ant clustering with partition validity. In Proceedings of the IEEE international conference on fuzzy systems (pp. 263–267). Piscataway: IEEE Press.

[20] Kanade, P. M., & Hall, L. O. (2003). Fuzzy ants as a clustering concept. In NAFIPS 2003: 22nd international conference of the North American fuzzy information processing society (pp. 227–232). Piscataway: IEEE Press.

[21] Kanade, P. M., & Hall, L. O. (2004). Fuzzy ant clustering by centroid positioning. In Proceedings of the IEEE international conference on fuzzy systems (Vol. 1, pp. 371–376). Piscataway: IEEE Press.

[22] Monmarché, N., Slimane, M., & Venturini, G. (1999). On improving clustering in numerical databases with artificial ants. In D. Floreano, J.-D. Nicoud, & F. Mondada (Eds.), Lecture notes in artificial intelligence: Vol. 1674. Advances in artificial life: 5th European conference, ECAL 99 (pp. 626–635). Berlin: Springer.

[23] Monmarché, N., Ramat, E., Desbarats, L., & Venturini, G. (2000). Probabilistic search with genetic algorithms and ant colonies. In A. S. Wu (Ed.), Workshop on optimization by building and using probabilistic models, GECCO 2000 (pp. 209–211).

[24] Li, Q., Shi, Z., Shi, J., & Shi, Z. (2005). Swarm intelligence clustering algorithm based on attractor. In L. Wang, K. Chen, & Y.-S. Ong (Eds.), Lecture notes in computer science: Vol. 3612. Advances in natural computation, first international conference, ICNC 2005 (pp. 496–504). Berlin: Springer.

[25] Lumer, E., & Faieta, B. (1994). Diversity and adaptation in populations of clustering ants. In Proceedings of the Third International Conference on Simulation of Adaptive Behavior: From Animals to Animats 3 (pp. 501 –508). Cambridge, MA: MIT Press.

[26] Ramos, V., & Merelo, J. (2002). Self-organized stigmergic document maps: environments as a mechanism for context learning. In Proceedings of the first Spanish conference on evolutionary and bio-inspired algorithms (pp. 284–293). Mérida: Centro Univ. Mérida

[27] Vizine, A. L., de Castro, L. N., Hruschka, E. R., & Gudwin, R. R. (2005b). Towards improving clustering ants: an adaptive ant clustering algorithm. Informatica, 29, 143–154.

[28] Montes de Oca, M. A., Garrido, L., & Aguirre, J. L. (2005). Effects of inter-agent communication in Ant-based clustering algorithms: a case study on communication policies in swarm systems. In A. Gelbukh & H. Terashima (Eds.), Lecture notes in artificial intelligence: Vol. 3789. MICAI 2005: advances in artificial intelligence: 4th Mexican international conference on artificial intelligence (pp. 254–263). Berlin: Springer.

[29] Schockaert, S., Cock, M. D., Cornelis, C., & Kerre, E. E. (2004a). Efficient clustering with fuzzy ants. In D. Ruan, P. D'hondt, M. D. Cock, M. Nachtegael, & E. E. Kerre (Eds.), Applied computational intelligence, proceedings of the 6th international FLINS conference (pp. 195–200). River Edge: World Scientific.

[30] Schockaert, S., Cock, M. D., Cornelis, C., & Kerre, E. E. (2004b). Fuzzy Ant-based clustering. In M. Dorigo, M. Birattari, C. Blum, L. M. Gambardella, F. Mondada, & T. Stützle (Eds.), Lecture notes in computer science: Vol. 3172. Ant colony optimization and swarm intelligence, 4th international workshop, ANTS 2004 (pp. 342–349). Berlin: Springer

[31] Das, Gunopulos, Mannila(1997). Finding Similar Time Series. In: Proceeding of the First PKDD Symptoms.

[32] Zhang, Y., Padman, R., & Wasserman (2014). On Learning and Visualizing Practice-based Clinical Pathways for Chronic Kidney Disease. AMIA Annual Symposium Proceedings.

[33] Namiki, Y., Ishida, T., Akiyama, Y. (2013). Acceleration of sequence clustering using longest common subsequence filtering. BMC Bioinformatics, 14(Suppl 8), S7. http://doi.org/10.1186/1471-2105-14-S8-S7.

[34] Chen, Y., Lu, H., & Li, L. (2017). Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. PLoS ONE, 12(3), e0173410. http://doi.org/10.1371/journal.pone.0173410.

[35] Park, K., Lin, Y., Metsis, V., Le, Z., & Makedon, F. (2010). Abnormal human behavioral pattern detection in assisted living environments. In Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments (p. 9). ACM.

[36] Bannink, L., Wells, S., Broad, J., Riddell, T., & Jackson, R. (2006, November 17). Web-based assessment of cardiovascular disease risk in routine primary care practice in New Zealand: The first 18,000 patients (PREDICT CVD-1). Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17146488.

[37] P.W, Wilson D, R.B, Levy, Belanger, Silbershatz, Kannel, W.B. (1998). Prediction of coronary heart disease using risk factor categories. PubMed .97 (18): 1837–1847. PMID 9603539.

[38] Reddy, R. K. Y., Mahendra, J., & Gurumurthy, P. (2015). Identification of predictable biomarkers in conjunction to Framingham risk score to predict the risk for cardiovascular disease (CVD) in Non cardiac subjects. Journal of clinical and diagnostic research: JCDR, *9*(2), BC23.

[39] Anderson KM, Odell PM, Wilson PWF, Kannel WB (1991). Cardiovascular disease risk profiles. American Heart Journal,121(1, Part 2):293.

[40] Lloyd-Jones, D. M., Wilson, P. W., Larson, M. G., Beiser, A., Leip, E. P., D'Agostino, R. B., & Levy, D. (2004). Framingham risk score and prediction of lifetime risk for coronary heart disease. The American journal of cardiology, 94(1), 20-24.

[41] Tsipouras, Karvounis EC, Tzallas AT, Katertsidis NS, Goletsis Y, Frigerio M, Verde A, Trivella MG, Fotiadis DI (2013). Adverse event prediction in patients with left ventricular assist devices. In: Proceedings IEEE Medical Biological Society.

[42] New Zealand Guidelines Group(2003). Assessment and Management of Cardiovascular Risk. Wellington. https://www.health.govt.nz/system/files/documents/publications/cvd_risk_summary.pdf

[43] Chaoji, V., Al Hasan, M., Salem, S., & Zaki, M. J. (2008, December). Sparcl: Efficient and effective shape-based clustering. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 93-102). IEEE.

[44] Das, Gunopulos, Mannila(1997). Finding Similar Time Series. In: Proceeding of the First PKDD Symptoms.

[45] Knight,J., Wells, S., Marshall, R., Exeter, D., Jackson, R. (2017). Developing a synthetic national population to investigate the impact of different cardiovascular disease risk management strategies: A derivation and validation study. PLOS ONE, 12(4), p.e0173170.

[46] Rousseeuw J, Silhouettes P (1987). A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, Vol.20, pp.53-65.

[47] Davies D, Bouldin D (1979). A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2, pp.224-227.

[48] Dunn J (1974). Well-separated clusters and optimal fuzzy partitions, Journal of Cybernetics, Vol.4, pp.95-104.