I.J. Modern Education and Computer Science, **2022, 6, 35-52**
Published Online on December 8, 2022 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2022.06.04

An Efficient and Scalable Technique for Clustering Comorbidity Patterns of Diabetic Patients from Clinical Datasets

Bramesh S M*
P. E. S. College of Engineering, Mandya, 571401, Karnataka, India
Email: smb@pesce.ac.in
ORCID iD: https://orcid.org/0000-0003-3924-657X
*Corresponding Author

Anil Kumar K M
JSS Science and Technology University, Mysuru, 570006, Karnataka, India
Email: anilkm@sjce.ac.in
ORCID iD: https://orcid.org/0000-0002-3236-1500

Received: 24 May, 2022; Revised: 16 June, 2022; Accepted: 19 August, 2022; Published: 08 December, 2022

Abstract: Clustering diabetic patients with comorbidity patterns are necessary to learn relationships between diabetes patients' clinical profiles and as an essential pre-processing stage for analysis tasks, like classification and categorization. Nevertheless, the heterogeneity of these data makes traditional clustering methods more difficult to apply, necessitating the development of novel clustering algorithms. In this paper, we recommend an effective and scalable clustering technique suitable for datasets made up of attributes which are atomic and set-valued. In these datasets, each record corresponds to a different diagnosis detail of a diabetic patient based on his or her hospital visit, where diagnosis details in each record are represented using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes. Our proposed technique involves three main stages. In the first stage, we selected the top-k diabetes-specific comorbidities patterns. In the second stage, we ensured that the co-occurring conditions in the selected top-k diabetes-specific comorbidities patterns really co-occur together or not using topic modeling and in the last stage, we constructed high quality clusters efficiently using average linkage agglomerative clustering with cosine similarity. Also, based on silhouette analysis, we assessed the efficiency and effectiveness of our proposed technique using a large, freely available MIMIC dataset (MIMIC-III and MIMIC-IV), comprised of over 14,222 and 68,118 distinct records, respectively. Our findings reveal that our technique finds clusters that: (i) preserve interrelations between demographics (age, gender) and diagnosis codes (ICD-9-CM codes), and (ii) are well-separated and compact. Finally, the founded clusters are beneficial for numerous investigative tasks like query answering, visualization, anonymization, classification etc.

Index Terms: Diabetes, Comorbidity patterns, Topic modeling, Clustering, and ICD-9-CM codes.

1. Introduction

An Electronic Health Record (EHR) is a computerized replica of a patient's paper chart [1], which includes information like the patient's demographics, diagnosis, prescriptions, and laboratory test results, among other things [2]. EHRs can assist in the delivery of healthcare by lowering paperwork time [3] and making patient information more accessible [4]. Furthermore, using data mining technologies [5,6], data-driven quality measurements and clinical research can both benefit from EHRs. Such technologies may be useful to guide the patient's intervention [7], by segmenting the data into meaningful clusters using clustering, and or by finding diagnoses codes that co-occur (comorbidities) that help prediction and assessment of quality care, via mining patterns. Nevertheless, due to the EHR data heterogeneity, numerous existing mining techniques are inappropriate to EHR data, necessitating the development of new ways to deal with data heterogeneity effectively [6].

1.1. Motivation

Consider the pre-processed Medical Information Mart for Intensive Care-IV (MIMIC-IV) dataset as shown in Table 1, here every record relates to a diabetic patient diagnosis detail and includes demographic information as well as a list of diagnosis codes. There are mainly two attribute types in the dataset: (i) Attributes with only one value per record (single-valued or atomic). Here, if the attribute is numerical, the value can be a number and if the attribute is categorical, the value can be a category. For instance, in Table 1, every diabetic patient's record comprises one value in the 'anchor_age' attribute (numerical attribute) and one more value in the 'gender' attribute (categorical attribute) (ii) Attributes which are set-valued on the other hand contains a list of values for each record. For instance, in Table 1, a diabetic patient's record comprises of diagnosis codes values in the 'icd_code' attribute (set_valued attribute). The same explanation will also hold good for MIMIC-III dataset. Relational Transaction datasets (RT-datasets) are datasets that contain both atomic and set-valued features or attributes and are useful in a variety of medical analysis applications [8-12]. Clustering an RT-dataset, made up of demographics (age, gender) and diagnosis codes (ICD-9-CM codes), focuses on generating meaningful clusters of records that share related demographics (age, gender) and diagnosis codes (ICD-9-CM codes). To put it another way, the goal of this work is to uncover natural, hidden data structures [13]. For instance, using the MIMIC-III dataset, our method produce clusters with male patients whose age falls under 15 - 52 associated with diabetes-specific disease comorbidities, clusters with female patients whose age falls under 15 - 52 associated with diabetes-specific disease comorbidities, clusters with male patients whose age falls under 53 - 71 associated with diabetes-specific disease comorbidities, clusters with female patients whose age falls under 53 - 71 associated with diabetes-specific disease comorbidities, clusters with male patients whose age falls under 71 - 93 associated with diabetes-specific disease comorbidities, clusters with female patients whose age falls under 72 - 93 associated with diabetes-specific disease comorbidities. Furthermore, each cluster's records contain correlated diagnosis codes that affect patients in large number, making cluster interpretation easier [14]. The generated clusters are beneficial for numerous investigative tasks, including: (i) query answering (for instance, to calculate aggregate statistics for subpopulations of patients in various clusters and utilize them to compare subpopulations), (ii) visualization (for instance, to examine the visible clusters in order to gain insights on patient subpopulations), (iii) anonymization [8] (for instance, to safeguard patient privacy, utilizing the clusters as input to algorithms that modify the values in every cluster), and (iv) classification (for instance, to pre-process a dataset in order to extract record classes that may then be used to do classification more efficiently and effectively [15,16]). As a result, an RT-dataset can be clustered and the clustering result can be useful to one or more of these tasks.

Table 1. An example of a pre-processed MIMIC-IV dataset

subject_id	hadm_id	icd_code	gender	anchor_age
17371341	28632974	25000, 36202, 64892, 64801, V270,6822	F	34
17371341	29258964	64801, V270, 25000, 65961, V4585,65661,65421,36201	F	34
19674536	22975928	5750, 5859, 2800, E9426, 78820, 53540, 5990, 3310, 532...	F	87
19674536	25678268	E9478, 5809, 2449, 5849, 3899, 25000, 41519, 29680, 29...	F	87
19674536	23999638	43310, 2948, 3899, 71690, 5990, V1582, 42833, 4280, 00...	F	87
...	...	...	...	...
11954282	23764552	71590, 31401, 2724, 2859, 7801, 25000, E9397, E9356,4...	F	57
17268513	26840060	5853, 42732, V4365,2729, 8020, E8859, 5693, 45981, 71...	M	86
11559004	27700446	81002, 80706, E8261, 2724, E8495, 25000, V180	M	72
17924864	22192246	311, 7919, 27651, 25000, 2809, 486, 2724, 5849, 53081, ...	F	73
10995091	20292470	V5867, 4019, 2851, 2724, 5781, V5866, V1051, 25000	F	63

Existing clustering methods, on the other hand, are not intended to cluster RT-datasets, which include demographics (age, gender) and diagnosis codes (ICD-9-CM codes). This is due to the following reasons as reported in the literature [8][17]:

(i) Many clustering methods utilise a single similarity measure, and creating one such measure that captures the records similarity, when each record is made up of two attribute types namely atomic and set-valued is difficult. The reason for this is that atomic attribute, like demographics, and set-valued attributes, like diagnosis codes, have different interpretations. i.e., a demographic attribute has one value for each record out of a limited number of possible values, whereas diagnosis codes have a large number of values for each record out of a vast number of likely diagnosis codes. Because of this it is hard to find one similarity metric that captures how similar two or more records' demographics and diagnosis codes are. For example, Euclidean distance, which is appropriate for numerical demographics, is ineffective for determining the distance between diagnosis codes sets. Also, the Jaccard distance, which is appropriate for calculating the distance between diagnosis codes sets, is not appropriate for assessing the distance between numerical demographics.

(ii) For RT-datasets, techniques for multi-objective clustering that seek to optimize numerous measures at the same time are ineffective. Using a compound measure like the product or weighted sum of Euclidean distance (used to demographics) and Jaccard distance (used to diagnosis codes), for example, could result in low-quality clusters. This is due to the fact that the distribution of values for each of these measures is different,

implying that the compound measure does not accurately reflect quality of cluster. Also, 2-level or hybrid optimization algorithms that first cluster demographics and then cluster diagnosis codes (for example, the strategy utilized in [8]) fail to find clusters with high-quality. Additionally, co-clustering (also known as bi-clustering) algorithms (e.g., [18,19]) are not intended for clustering an RT-dataset with demographics (age, gender) and diagnosis codes (ICD-9-CM codes). The reason for this is that they may generate clusters comprising portions of records, whereas in our case clustering technique must generate clusters, in such a way that the complete diabetic patient record must be there in any one generated cluster.

1.2. Contributions

We present an efficient and scalable clustering technique for an RT-dataset that includes demographics (age, gender) and diagnostic codes (ICD-9-CM codes). Our new technique fundamental idea is to build a record representation that lets us to compare records with reference to both demographics (age, gender) and diagnosis codes (ICD-9-CM codes).

To create such a representation, firstly we discretize [20] numerical demographics like age (i.e., substitute aggregate values for their values) and secondly select diagnosis codes subsets contained in the dataset, which are referred to as top-k diabetes-specific comorbidities patterns. Lastly, using one-hot encoding, we express every record of the RT-dataset as a binary representation, as shown in Table 2 (Shown for MIMIC-IV RT-dataset). The binary representation's features (columns) are as follows:

(i) each numerical demographic attribute's values, (ii) each category demographic attribute's values, and (iii) the selected top-k diabetes-specific comorbidities patterns (CP_1, CP_2, CP_3, CP_4, and CP_5). A binary representation feature with a value of 1 (or 0) indicates that the record has (or does not include) the feature. We use a clustering technique designed for data represented in binary to create clusters of related records based on the binary representation.

Table 2. MIMIC-IV dataset represented in binary

18-51	52-70	71-91	Male	Female	CP_1	CP_2	CP_3	CP_4	CP_5
0	0	1	0	1	1	0	0	0	0
0	0	1	0	1	1	0	0	0	0
0	1	0	0	1	0	0	0	0	1
0	1	0	0	1	1	0	0	0	0
1	0	0	1	0	0	0	0	0	1
0	0	1	0	1	0	0	0	1	0
0	0	1	0	1	0	0	0	1	0
0	0	1	0	1	0	0	1	0	0
1	0	0	1	0	1	0	0	0	0
1	0	0	1	0	1	0	0	0	0
0	1	0	1	0	0	1	0	0	0
0	1	0	1	0	0	1	0	0	0

Yet, there are three issues that need to be attempted to appreciate our technique. First, we need a way to select top-k diabetes-specific comorbidities patterns that help us to construct clusters with high-quality. Second, there should be a way to ensure that the co-occurring conditions in the selected top-k diabetes-specific comorbidities patterns really co-occur together or not. Third, there should be a way to efficiently generate clusters with high-quality. We tackle these issues in three stages namely Pattern Selection (PS), Pattern Validation (PV) and Pattern-based Clustering (PC). They are described as follows:

1.2.1. The Pattern Selection

It involves selecting patterns that consists of interrelated diagnosis codes (specifically, any code in the pattern is highly likely to infer the other codes in the pattern) in diabetic patients from MIMIC-III and MIMIC-IV datasets respectively. The reason for selecting such patterns is that these patterns reveal diabetes-specific disease comorbidity patterns in large diabetic patients, which will favor clusters with interrelated diagnosis codes, consequently worthless clusters with diagnoses codes that coexist by accidental are circumvented and also favor well-separated and compact clusters in case of both MIMIC-III and MIMIC-IV dataset.

1.2.2. The Pattern Validation

Here, we have used Latent Dirichlet Allocation (LDA) model on both MIMIC-III and MIMIC-IV datasets to validate the co-occurring diagnoses codes or conditions of diabetic patients in the comorbidity's patterns selected by the PS stage from MIMIC-III and MIMIC-IV dataset respectively. i.e., to ensure that the co-occurring conditions in the comorbidity's patterns selected by the PS stage really co-occur together or not.

1.2.3. The Pattern-based Clustering

Here, we use an RT-dataset as input, as well as the selected and validated top-k diabetes-specific comorbidities patterns by the PS and PV stages respectively, to cluster the dataset. Pattern-based clustering accomplishes the following tasks:

(i) It represents RT-dataset in binary form: In this illustration, the records having the diagnosis codes in not less than one of the identified and validated top-k diabetes-specific comorbidities patterns by the PS and PV stages respectively make up the dataset, which needs to be clustered. This enables our technique to emphasis on records with frequently occurring and interrelated diagnosis codes, resulting in understandable clusters. The unclustered records (outstanding records) can be handled alone, using various methodologies based on the application's needs.

(ii) It clusters the dataset represented in binary to form records groups that are comparable in terms of both demographics (age, gender) and diagnosis codes (ICD-9-CM codes). Here, the average linkage agglomerative clustering technique [13] with cosine similarity [21], is used to cluster the data. We adopt average linkage agglomerative clustering with cosine similarity because it does not require the clustering representative notion, which is improper for data represented in binary [22,23], unlike density-based, partitional, as well as complete and single linkage hierarchical algorithms. We use cosine similarity because, unlike Jaccard or Euclidean distance, it works well for grouping binary-represented data with several features [24]. The advantage of proposed technique is that it generates clusters consist of interrelated diagnosis codes (because of the identified and validated patterns of top-k diabetes-specific comorbidities) that are also interrelated with demographics (because of the clustering).

We implemented our efficient and scalable technique, using PS, PV and then the PC stages. We assessed the efficiency and effectiveness of our technique using a large, freely-available MIMIC dataset (MIMIC-III and MIMIC-IV), comprised of over 14,222 and 68,118 distinct records, respectively. Our findings reveal that our technique finds clusters that: (i) preserve interrelations between demographics (age, gender) and diagnosis codes (ICD-9-CM codes), and (ii) are well-separated and compact.

1.3. Paper Organization

The remainder of the paper is laid out as follows. Section 2 discuss the Related work. Section 3 includes background information as well as a statement of the problem addressed in this paper. Section 4 presents materials and methods. Section 5 presents our technique for RT-datasets clustering. Section 6 presents experimentation and results. Lastly, Section 7 concludes the paper.

2. Related Work

In this section, we highlight the techniques that are related to clustering EHR and the research problem we intend to address. For widespread surveys of EHR data mining, publications by P.B. Jensen et al. [25] (2012), P. Yadav et al. [6] (2018) and Tabinda et al. [63] (2022) are very interesting. Section 2.1 discusses EHR data clustering, and Section 2.2 presents pattern-based clustering in brief.

2.1. EHR Data Clustering

Based on the data type, we categorize existing techniques for EHR data clustering. They are described as follows:

(i) Demographics: In contrast to RT-datasets, datasets composed of a list of demographic attributes are intrinsically low-dimensional (for instance, they naturally comprise less than twenty demographics). As a result, several existing clustering techniques can be used to cluster them, such as hierarchical (for instance, average-linkage, complete-linkage, and single-linkage) [13], partitional (for instance, k-medoids [28], k-means++ [27], & k-means [26]), and density-based (for instance, OPTICS [29] and DBSCAN [21]) techniques. For instance, an interesting work by B. Andreopoulos et al. [30] (2009) uses density-based clustering technique on patient data. The motive for the aforementioned techniques is not appropriate for RT-dataset clustering is that, the similarity measures they use cannot be enforced to an attribute of set-valued (for instance, the diagnosis codes attribute), as they unable-to capture similarity efficiently for attributes of set-valued. It is also observed from the literature that in case of high-dimensional data, similarity of records could not be captured successfully by distance measures employed in several existing techniques because attributes of set-valued (for instance, the diagnosis codes attribute) are intrinsically high-dimensional [22] [8]. This is referred to as the high-dimensionality curse [21].

(ii) Diagnosis codes: These are similar to RT-datasets, here every record contains a list of diagnosis codes that are intrinsically high-dimensional (i.e., the attribute of set-valued naturally consist of thousands of diverse diagnosis codes). We can cluster such datasets using techniques build for set-valued data. Examples of such techniques are SCALE [32], CLOPE [31], SV-k-modes [33], and ROCK [22]. For instance, working of SV-k-modes is identical to k-means, except instead of the centroid, it employs a list of values in an attribute of set-valued as the cluster representative. By employing centroids as representatives in atomic attributes, this approach can be used to datasets with atomic attributes

(i.e., RT-datasets). SV-k-modes is only relevant to attributes of set-valued with a very trivial domain size (for instance, attributes with ten to fifty different values) because of its exponential time complexity with reference to the set-valued attributes domain size. As a result, unlike the proposed technique, SV-k-modes is unsuitable for clustering an attribute of set-valued made up of diagnosis codes with size of domain in thousands. In the literature, there are a few noteworthy works on clustering EHR datasets constituted of sets of diagnosis codes by L. Kalankesh et al. [34] (2013) and F.S. Roque et al. [35] (2011). These techniques [31, 32, 22, 34, 35] could not be applied to cluster an RT-dataset since their similarity measures could not be applied to atomic attributes in an RT-dataset, such as demographic attributes.

(iii) Other high-dimensional data: There are several clustering techniques that are applied to datasets that made up of genomic sequences, trajectories, or text. For instance, F. Doshi-Velez et al. [36] (2014) and S. Ghassempour et al. [37] (2014) focused on clustering trajectory data, where trajectories denote diseases sequences, whereas C.E. Lopez et al. [38] (2018) and A. Ultsch et al. [39] (2017) studied clustering genomic data. H. Xu et al. [40] (2012) and M. Moradi et al. [41] (2018) studied medical text clustering. These techniques, obviously, cannot be regarded as substitutes to our technique since the data they employed has different meanings than the RT-dataset attributes.

2.2. Pattern-based Clustering

Pattern-based clustering techniques represent dataset records in binary form, in which the patterns are features, and clustering is applied to the data represented in binary [42,43]. As a result, our method is similar to pattern-based clustering algorithms. Pattern-based clustering is an efficient technique for clustering data of high-dimensional because the number of patterns is usually lesser than the number of different values in the dataset [44]. Pattern-based clustering algorithms also generate clusters that are not "too" trivial and easy to comprehend on the basis of the patterns, by concentrating on records comprising patterns with specified features, such as common patterns [14]. The following are some of the most popular types of data that are clustered using pattern-based approaches.

(i) Documents: There are techniques targeting to cluster documents dataset. Every dataset record represents a document, which is represented as a collection of words [45]. These techniques usually produce clusters with low quality, due to the high dimensionality of the data. The efforts in [46-48] are motivated by this limitation and use common itemsets as representational patterns for documents. The purpose of these studies is to lower the document representation's dimensionality in order to increase clustering accuracy and efficiency.

(ii) Gene expression data: They're commonly represented as a matrix of real-valued, with each row representing a gene and each column representing a condition [19]. However, clustering such a matrix is problematic [49] since a collection of genes can only show the same activation patterns under specified experimental conditions. Y. Cheng et al. [50] (2000) suggested Bi-clustering approaches to solve this problem, which cluster the rows and columns of the matrix at the same time. Bi-clustering approaches (for instance, [19, 49, 51, 52, 18]) produce, as clusters, submatrices in which genes subgroups show highly interrelated activities for conditions subgroups. Several of these approaches [18] also uses patterns for bi-clustering, such as association rules and frequent patterns. All bi-clustering approaches strive to cluster the matrix's rows and columns at the same time. As a result, they cannot be utilised to organise records into clusters, contrary to our strategy. By applying bi-clustering to the binary representation in pattern-based clustering, a row equivalent to an RT-dataset record could take part in numerous clusters.

3. Background and Problem Statement

In this section, we'll go over some basic principles as well as the problem we're trying to address.

3.1. RT-datasets

We study an RT-dataset, say D, in which each record r corresponds to a different diagnosis detail of a diabetic patient based on his or her hospital visit. More specifically, every record r in D has one or more attributes of demographic that might be numerical, and or categorical, as well as attributes of set-valued that contains diagnosis codes. Without losing generality, we presume that the first m attributes in D, represented with a^1, …, a^m, are attributes of demographic, and the last l attributes in D, represented with a^{m+1}, …, a^{m+l}, are attributes of set-valued. The diagnosis codes are represented using ICD-9-CM codes.

3.2. LDA

Here, we introduce the concept of LDA used in our approach.

We use LDA to simulate the generation of diabetes patient records from a combination of K hidden topics, where a topic is a multinomial distribution across all of our vocabulary's coded conditions [53]. The steps in the generating process for each patient file are as follows:

First, by sampling from a Dirichlet distribution with α (parameter), a multinomial distribution for the t^{th} topic across V codes, denoted Φ_t ($K \geq t \geq 1$), is generated. The code conditional probability appearing in the t^{th} topic is represented by Φ_t. Next, a multinomial distribution across K topics, designated θ_i is sampled from a Dirichlet distribution with β (parameter) for each patient file, F_i. The file conditional probability being connected with each of the K topics is represented by θ_i.

Subsequently, for every code-position, j, in the file, F_i: (i) By sampling from θ_i, a topic is drawn; the nominated topic at j position in F_i, is represented as $Z_j^i \in \{1, \ldots, K\}$; (ii) Given: Z_j^i (the topic), C_j^i (a code) is drawn by sampling $\Phi_{Z_j^i}$ (the topic-code distribution).

Here, we have selected the value of K (No. of topics) as 2 for MIMIC-III and MIMIC-IV dataset based on the Coherence Score of the model on the respective datasets. Moreover, both datasets (MIMIC-III and MIMIC-IV) on which LDA has applied has records belonging to any one of the topics, i.e., diabetes or non-diabetes.

3.3. Problem Statement

The clustering problem that we are aiming to solve is now defined.

Given: D - an RT-dataset, $B = \{B_1, \ldots, B_{|D|}\}$ - a binary representation of D, and k – parameter. Create $C = \{c_1, \ldots, c_k\}$ - a partition of D with maximum $\sum_{i\in[1,k]} \sum_{r,r'\in c_i} cos(B_r, B_{r'})$, where c_i is a cluster and $cos(B_r, B_{r'}) = \frac{B_r B_{r'}}{\|B_r\| \|B_{r'}\|}$ is the cosine similarity metric between B_r and $B_{r'}$ records in B, which corresponds to the r and r' records, correspondingly, in c_i.

The problem that we are aiming to solve takes an RT-dataset D, specifically the binary representation of D, and the number of clusters k as inputs, and it necessitates finding k clusters using clustering, such that each cluster's records are similar. Precisely, it aims to maximize total cosine similarity, which is calculated using the D's binary representation records. The problem is NP-complete (this is straightforward from [53,54]), which defends the heuristics development.

4. Materials and Methods

4.1. Datasets

We used the MIMIC-III [54] and MIMIC-IV [55] datasets for the study. Both datasets, are large, de-identified and publicly accessible collection of medical records, which was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (BIDMC, Boston, MA, USA) and the Massachusetts Institute of Technology (MIT, Cambridge, MA, USA) [54,55]. Each record in the MIMIC-III dataset includes ICD-9-CM codes and some records in the MIMIC-IV dataset includes ICD-10-CM codes also, which identify diagnoses and procedures performed.

The key characteristics of the datasets is shown in Table 3.

Table 3. Key characteristics of the datasets

Key Characteristic	MIMIC-III Dataset [54]	MIMIC-IV Dataset [55]
Number of diagnostic records (only ICD-9-CM) of diabetic patients	1,99,964	9,47,244
Number of distinct hospital admissions of diabetic patients	14,222	68,118
Number of distinct people diagnosed with diabetes	10,318	26,115
Number of distinct male people diagnosed with diabetes	5,898	13,784
Number of distinct female people diagnosed with diabetes	4,420	12,331
Contains data from year	2001 - 2012	2008 - 2019

Table 4. Example of datasets

MIMIC - III

SUBJECT_ID	HADM_ID	ICD9_CODE	GENDER	AGE
117	140784	5715	F	49
117	140784	7895	F	49
117	140784	7054	F	49
117	140784	2875	F	49
117	140784	4280	F	49
...	...	...	...	...
97488	161999	414	M	67
97488	161999	30391	M	67
97488	161999	E8798	M	67
97488	161999	78791	M	67
97488	161999	V4986	M	67

MIMIC - IV

subject_id	hadm_id	icd_code	gender	anchor_age
17371341	28632974	25000	F	34
17371341	28632974	36202	F	34
17371341	28632974	64892	F	34
17371341	28632974	64801	F	34
17371341	28632974	V270	F	34
...	...	...	...	...
10995091	20292470	2724	F	63
10995091	20292470	5781	F	63
10995091	20292470	V5866	F	63
10995091	20292470	V1051	F	63
10995091	20292470	25000	F	63

4.2. Dataset Pre-processing

(i) Study Population: Since, we aimed at exposing patterns of commonly co-occurring diagnosis codes represented as ICD-9-CM codes in MIMIC datasets, we have selected the patient records, whose diagnosis details are represented in ICD-9-CM codes. Based on this criterion, 6,51,000 number of diagnostic records of patients were selected from MIMIC-III dataset and 30,90,370 number of diagnostic records of patients from MIMIC-IV dataset.

(ii) Dealing with Missing Values: From the exploratory data analysis, we observed that there were zero missing values in the MIMIC-IV dataset and very few missing values in the diagnosis codes columns in the MIMIC-III dataset. The records of this nature were removed from the datasets for further analysis.

(iii) Representing ICD-9-CM diagnosis codes defining diabetes: ICD-9-CM diagnosis code from 250.0x to 250.9x and from 250.x0 to 250.x3 represents diabetes complications [57]. In both MIMIC-III and MIMIC-IV dataset, ICD-9-CM codes are represented as a sequence of characters without decimal digits, i.e., 250.0x is represented as "2500x". Since our primary focus is to identify the comorbidities patterns in diabetic patients we represented the ICD-9-CM diagnosis codes representing diabetes complications as "25000" for further analysis.

(iv) Representing MIMIC datasets as RT-datasets: Both MIMIC-III and MIMIC-IV datasets, as shown in Table 4, contains diagnosis details of diabetic patients in several rows, where each row contains one diagnosis detail of a diabetic patient based on his / her hospital visit. So, to represent both MIMIC-III and MIMIC-IV datasets, as respective RT-datasets, we joined all the diagnosis details represented in several rows in original datasets into a single row containing all the diagnosis details of a diabetic patient based on his / her hospital visit. An example of the MIMIC-IV RT-dataset is shown in Table 1.

5. Proposed Methodology

This section presents our proposed technique for RT-dataset clustering, as specified in Problem Statement. Our technique applies: (i) the PS, which selects top-k diabetes-specific comorbidities patterns from MIMIC datasets, (ii) the PV, uses LDA model on MIMIC datasets to reveal and or to validate co-occurring diagnoses codes in the respective top-k diabetes-specific comorbidities patterns selected by the PS and (iii) the PC algorithm, first constructs the RT-datasets binary representation and then clusters the same, and produces the respective clustered RT-dataset. In the following, we briefly describe the operation of the PS, PV, and PC stages.

5.1. Dataset Pre-processing

Here, we have selected the top-k association patterns of diabetes-specific disease comorbidities identified by Bramesh S.M. et. al [59] from MIMIC-III and MIMIC-IV datasets as top-k diabetes-specific comorbidities patterns in respective datasets as shown in Table 5. They have implemented an effective rule-based approach for identification of comorbidity patterns namely CP_1 = {42731, 25000, 4280}, CP_2 = {25000, 2724, 4019}, and CP_3 = {25000, 41401, 4019} in diabetic patients from MIMIC-III dataset and comorbidity patterns namely CP_1 = {25000, 5859, 40390}, CP_2 = {25000, 42731, 4280}, CP_3 = {41401, 25000, 2724}, CP_4 = {25000, 53081, 2724}, and CP_5 = {25000, 4019, 2724} in diabetic patients from MIMIC-IV dataset.

Table 5. Selected Diabetes-specific comorbidities patterns

Dataset	Diabetes-specific comorbidities patterns represented as set of ICD-9-CM codes
MIMIC-III	CP_1 = {42731, 25000, 4280}, CP_2 = {25000, 2724, 4019}, and CP_3 = {25000, 41401, 4019}
MIMIC-IV	CP_1 = {25000, 5859, 40390}, CP_2 = {25000, 42731, 4280}, CP_3 = {41401, 25000, 2724}, CP_4 = {25000, 53081, 2724}, and CP_5 = {25000, 4019, 2724}

5.2. The Pattern Validation

Here, we have used LDA model to expose and or to validate the commonly co-occurring diagnosis codes in the selected top-k diabetes-specific comorbidities patterns shown in Table 5 from MIMIC-III and MIMIC-IV datasets respectively.

To create a data-matrix which can be processed through via topic modelling, we make some data representation and pre-processing choices, explanation for the data pre-processing is given in Section 4.2 and the explanation for the data representation is given below.

Data representation:

Here, every diagnosis detail of a diabetic patient based on his or her hospital visit (i.e., patient record), R_i (where i lies between 1 and M), is converted to a codes vector, $F_i = \langle C_1^i, \ldots, C_{N_i}^i \rangle$, where M represents the patient records count in the dataset and N_i is the total code occurrences count in the i^{th} patient record. Every code, C_j^i, in the vector is one of the ICD-9-CM codes in our vocabulary V (set of unique diagnosis codes in the respective datasets), seen as a value taken by a respective random variable, C_j (where j lies between 1 and N_i), indicating the code value found in the i^{th} patient record at j^{th} position. Any of the V codes in our vocabulary can appear at any position in a patient record, according to our findings.

To create a data-matrix, all patient records are processed, where every row denotes a patient, and every column denotes ICD-9-CM code contained in the vocabulary, so that every cell $\langle p, c \rangle$ denotes how many times a patient p was diagnosed with condition c. Every patient in the record set is thus assigned to a vector of V-dimensional, where V is the number of unique codes in our vocabulary and every vector element corresponds to the frequency with which a condition in the patient's record occurs. The corpus of all these vectors is called the patient-conditions corpus, and each of these vectors is called a patient-conditions record. Finally, we found that the data-matrix dimension of 58,929 × 6907 for the MIMIC-III dataset, and the data-matrix dimension of 3,35,378 × 2,87,566 for the MIMIC-IV dataset.

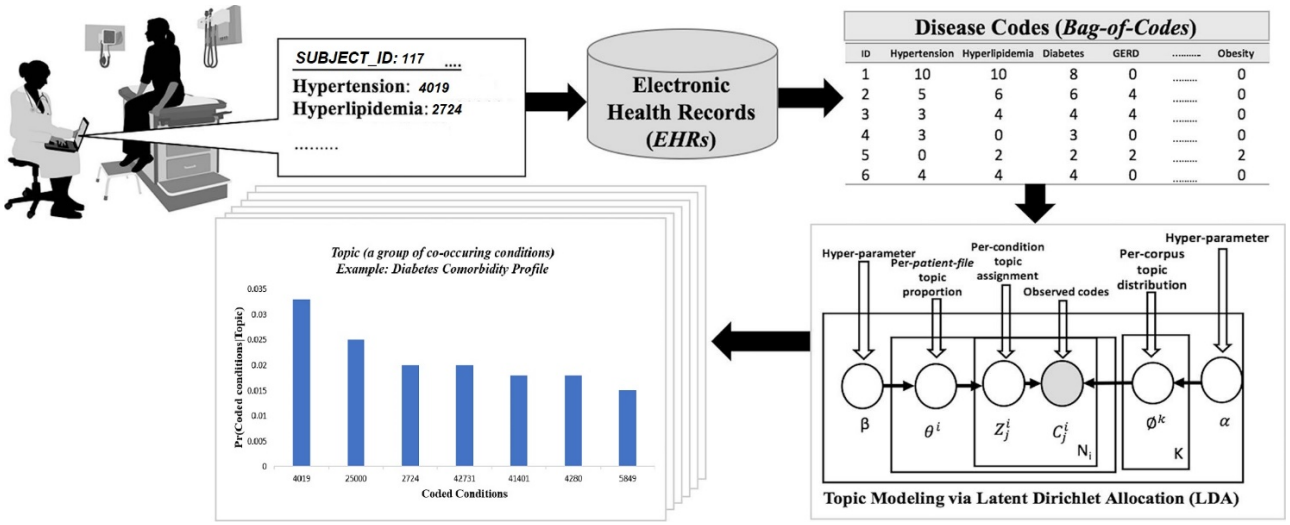

Fig. 1. Architecture of the LDA model used for pattern validation.

We use topic modelling in an unconventional way to uncover associations between ICD-9-CM codes issued and recorded in both MIMIC-III and MIMIC-IV datasets, as shown in Fig. 1. Unlike most previous work on topic modeling, we used the model on diagnosis codes instead of on the natural language, in order to expose and or to validate the commonly co-occurring diagnosis codes in selected top-k diabetes-specific comorbidities patterns. We use the Python Gensim package [60] to learn the parameters of the model based on our data and to build our model. We use the default values, set in the Gensim library, for the model parameters, except for the parameters like number of topics = 2 and the number of passes = 15.

5.3. The Pattern-based Clustering

The pattern-based clustering gets as input an RT-dataset, the top-k diabetes-specific comorbidities patterns selected and validated by PS and PV stages respectively, as well as k (number of clusters), and it works as shown in the Fig. 2:

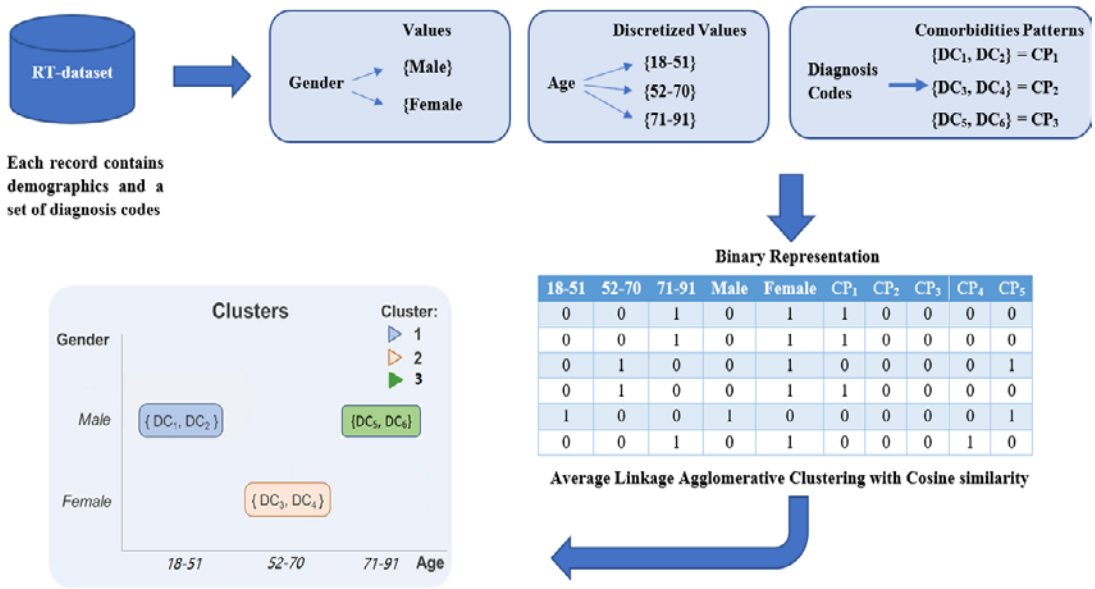

Fig. 2. Architecture of the Pattern-based Clustering.

As shown in the Fig. 2, the pattern-based clustering works in two phases:

(i) Creating binary representation: This phase focuses on representing the input RT-dataset in binary form based on the top-k diabetes-specific comorbidities patterns that were selected and validated by PS and PV stages respectively.

(ii) Hierarchical clustering: This phase focuses on clustering the binary represented RT-dataset using average linkage agglomerative clustering with cosine similarity.

We now discuss each phase in detail.

(i) Creating binary representation: Here, binary representation (i.e., *M*), of the RT-dataset (i.e., *D*), is created. As previously stated, the records in *M* and *D* have a 1-to-1 correspondence, and specifically the columns (attributes) of *M* correspond to the numerical demographic attributes discretized values, the categorical demographic attributes discretized values, and the input or selected and validated top-k diabetes-specific comorbidities patterns. The *k*-means++ technique [27] is utilized to discretize each numerical demographic attribute. *k*-means++ is fundamentally a *k*-means variant in which the first centroid is chosen at random and each subsequent centroid is chosen with a likelihood proportionate to its involvement to the overall Within-Cluster Sum of Squares (WCSS) measure.

Notably, *k*-means++ generates a discretization from the optimal discretization in O(log(k)) time [27]. In terms of the overall WCSS measure, this means that the discretized values are made up of relatively similar numerical values. The Elbow technique [61] chooses the number of discretized values in the attribute automatically. The Elbow technique is a useful heuristic [61] that employs *k*-means++ with *k* values ranging from 1 to 10 in our proposed technique and automatically selects the *k* value = 3 for MIMIC-III and MIMIC-IV dataset. Because at *k* = 3, sum of squared distances / inertia falls suddenly as shown in Fig. 3 and Fig. 4 for MIMIC-III and MIMIC-IV datasets respectively that results in better outcome in terms of the overall WCSS measure.

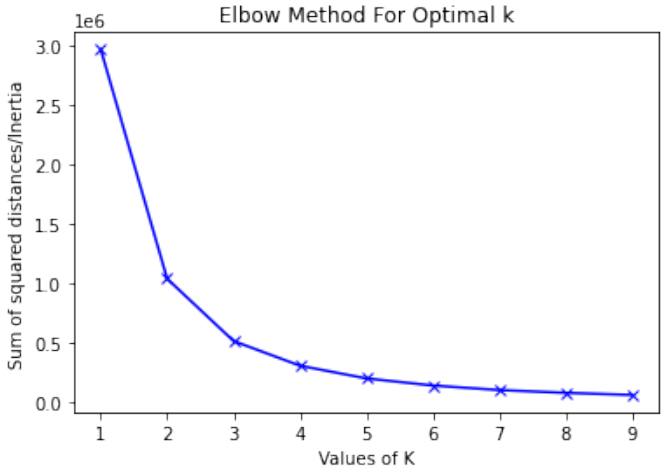

Fig. 3. Elbow Method on MIMIC-III dataset for finding optimal k for discretizing numerical attribute.

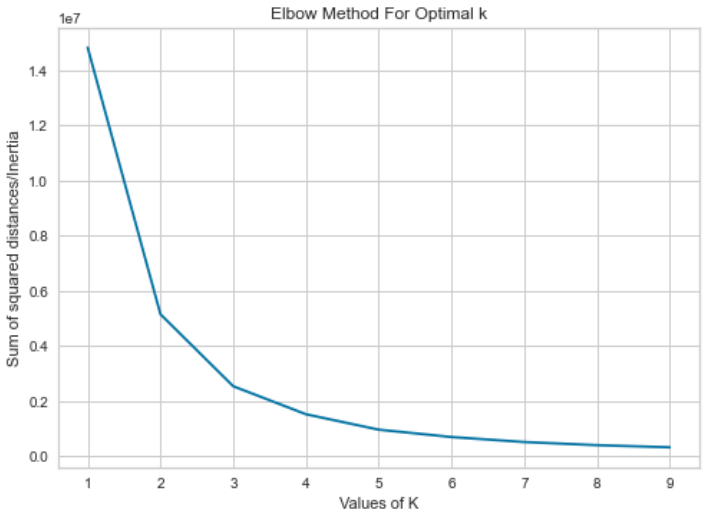

Fig. 4. Elbow Method on MIMIC-IV dataset for finding optimal k for discretizing numerical attribute.

We employ the Elbow approach since it allows for no-parameter discretization. Users can also discretize numerical demographic attributes like age using different ways [62] and send the discretized numbers to the pattern-based clustering as input.

(ii) Hierarchical clustering: In this step, the proposed technique uses average linkage agglomerative clustering with cosine similarity on the RT-dataset *D*'s binary representation *M*. This creates *k* clusters from *M*, with *k* being the input parameter for pattern-based clustering. We adopted hierarchical clustering since it has been demonstrated to be fruitful for EHR and binary-represented data clustering [36], and it is not necessary to build cluster representatives, which is troublesome for data like set-valued, as discussed in Section 1.1. We have selected similarity measure as cosine, since it

is shown to be fruitful for clustering data (represented in binary) with several features for example *M* [24]. It's worth noting that some records in *D* may lack the diagnosis codes found in any diabetes-specific comorbidity patterns in *M*. These records are known as un-clustered records, and they are fundamentally different from all other records represented in *M*. As a result, they are deleted from *M* (i.e., these data are not included to any clusters), which improves clustering quality as discussed in [14]. The clustered RT-dataset is created using pattern-based clustering, which clusters the dataset records according to how they were grouped in *M*. (i.e., creating a cluster with every record of the RT-dataset that are in the identical cluster after average linkage agglomerative clustering with cosine similarity has been used on *M*). Lastly, validating the clustering quality of our technique using silhouette analysis.

As mentioned above, we have used the average linkage to measure the distance between two sub-clusters of data points. In average linkage agglomerative clustering, the distance between two clusters like "*r*" and "*s*" (shown in Fig. 5) is defined as the average distance between each point in one cluster and each point in the other cluster. Mathematically, it is defined as shown in Eq. (1).

$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D\left(x_{ri}, x_{sj}\right) \tag{1}$$

Where,
n_r – Number of data points in *r*.
n_s – Number of data points in *s*.

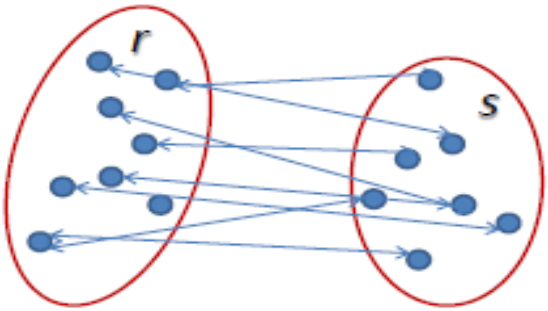

Fig. 5. Average Linkage.

6. Experimentation and Results

6.1. Experimental Setup

Table 6 shows the parameter values that we employed in our experiments. Specifically, the value for *k* (No. of Clusters) is selected using the Silhouette Score Elbow Method for *k*-means clustering technique and the value for the parameter *K* (No. of Topics) is selected as 2 for MIMIC-III and MIMIC-IV dataset based on the Coherence Score of the model on the respective datasets. Moreover, both datasets on which LDA has applied has records belonging to any one of the topics, i.e., diabetes or non-diabetes.

Table 6. The value for experimental parameters for each dataset

Dataset	Experimental Parameters		
	k (No. of Clusters)	*K (No. of Topics)*	*V (Vocabulary size)*
MIMIC - III	12	2	6, 907
MIMIC - IV	16	2	9, 495

As shown in the Fig. 6, we have applied *k*-means model for a range of *k* values from 2 to 30 on a MIMIC-III dataset to find an optimal value of *k*. When the model is fit with 12 clusters, we can see a line annotating the "elbow" in the graph, which in this case is known to be the optimal value.

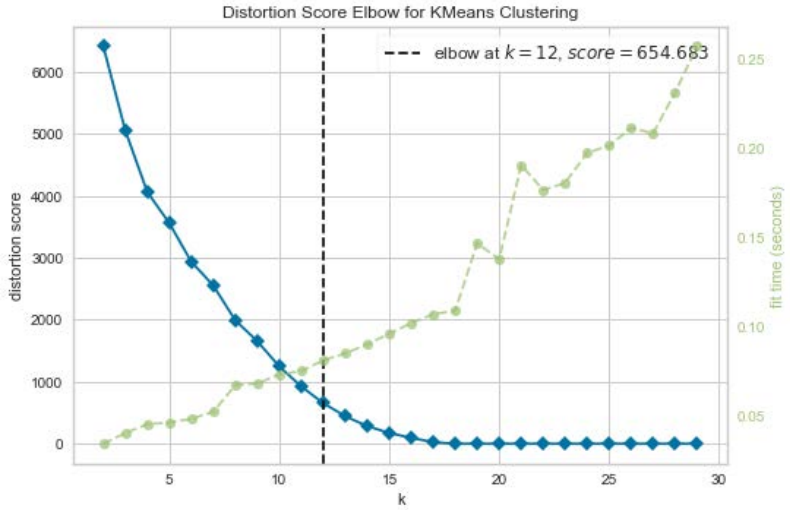

Fig. 6. Silhouette Score Elbow Method for finding optimal *k* on MIMIC-III dataset.

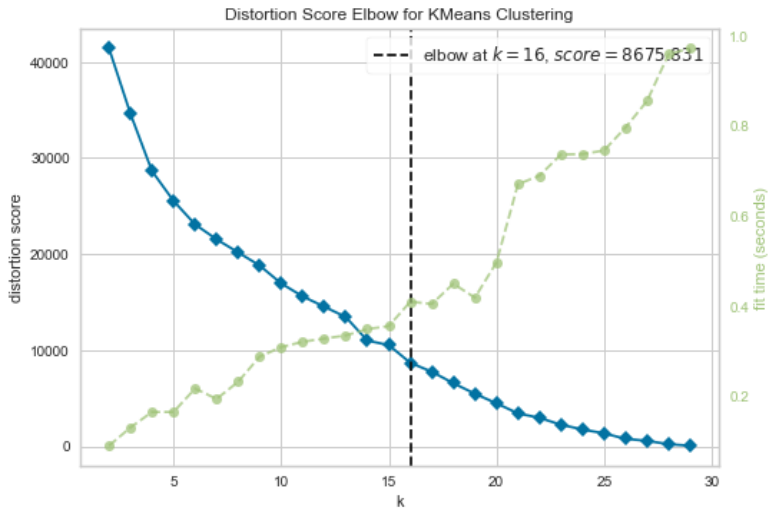

Fig. 7. Silhouette Score Elbow Method for finding optimal *k* on MIMIC-IV dataset.

As shown in the Fig. 7, we have applied *k*-means model for a range of *k* values from 2 to 30 on a MIMIC-IV dataset to find an optimal value of *k*. When the model is fit with 16 clusters, we can see a line annotating the "elbow" in the graph, which in this case is known to be the optimal value.

6.2. Experimental Results

In this section, we present the results obtained from our experiments. Firstly, we selected the top-k association patterns of diabetes-specific disease comorbidities identified by [59] from MIMIC-III and MIMIC-IV datasets as diabetes-specific disease comorbidities patterns in respective datasets as shown in Table 5.

Table 7. Example of one characteristic topic identified by our LDA model over the MIMIC-III dataset

Topic 1	$Pr(C\|T_i)$
4019	0.033
25000	0.025
2724	0.020
42731	0.020
41401	0.018
4280	0.018
5849	0.015

Table 7 reveals the conditions (represented as ICD-9-CM code) that co-occur in diabetic patients, along with the corresponding probability, $Pr(C|T_i)$, where C represents a condition and T_i represents the i^{th} topic. Table 7 has listed only the conditions, whose association probability with the particular topic is greater than 0.013. Also, Table 6 reveals that the diabetes-specific comorbidities patterns listed in Table 5 for MIMIC-III dataset is found to be valid as most of the conditions have really co-occurred in diabetic patients. We used only these patterns i.e., selected and validated patterns for clustering MIMIC-III dataset, hence the name pattern-based clustering.

Table 8. Example of one characteristic topic identified by our LDA model over the MIMIC-IV dataset

Topic 1	$Pr(C\|T_i)$
4019	0.038
2724	0.030
25000	0.029
V1582	0.023
53081	0.020
V5866	0.015
42731	0.014
41401	0.014
4280	0.013

Table 8 reveals the conditions (represented in ICD-9-CM codes) that co-occur in diabetic patients, along with the corresponding probability, $Pr(C|T_i)$, where C represents a condition and T_i represents the i^{th} topic. Table 8 has listed only the conditions, whose association probability with the particular topic is greater than 0.013. Also, Table 8 reveals that the diabetes-specific comorbidities patterns listed in Table 5 for MIMIC-IV dataset is found to be valid as most of the conditions have co-occurred in diabetic patients. We used only these patterns i.e., selected and validated patterns for clustering MIMIC-IV dataset, hence the name pattern-based clustering.

Table 9. Clusters of MIMIC-III dataset obtained by our technique

	Gender		Discretization of Age			Comorbidities Patterns selected from MIMIC-III as explained in Section 5		
Cluster ID	Male	Female	15 - 52	53 - 71	72 - 93	CP_1	CP_2	CP_3
0	542	0	0	542	0	0	542	0
1	487	0	0	487	0	487	0	0
2	0	604	0	0	604	604	0	0
3	663	0	0	0	663	663	0	0
4	509	0	0	509	0	0	0	509
5	316	0	0	0	316	0	0	316
6	0	241	0	241	0	0	0	241
7	0	266	0	0	266	0	0	266
8	0	68	68	0	0	0	68	0
9	105	0	105	0	0	0	0	105
10	143	0	143	0	0	0	143	0
11	0	41	41	0	0	0	0	41
12	0	13	13	0	0	13	0	0
13	42	0	42	0	0	42	0	0
14	295	0	0	0	295	0	295	0
15	0	309	0	0	309	0	309	0
16	0	342	0	342	0	0	342	0
17	0	267	0	267	0	267	0	0

Table 9 indicates that our clustering method has clustered the MIMIC-III dataset very well as we can find the records in every cluster that belongs to a certain age, gender and a comorbidity pattern. We also observed that the comorbidities patterns (CP_1, CP_2 & CP_3) depend on the age in the case of both male and female diabetic patients. It is also observed that the number of diabetic patients with comorbidity pattern CP_1 is less when compared with comorbidity patterns CP_2 & CP_3 in the age group 15-52 and the number of diabetic patients with comorbidity pattern CP_1 is more in the age group 72-93 when compared with comorbidity patterns CP_2 & CP_3 in the same age group. In other words, in younger age groups, the number of diabetic patients with comorbidity pattern CP_2 or CP_3 is more, whereas the diabetic patients older than 72 and above will be associated with comorbidity pattern CP_1. In the age group 53-71, the number of diabetic patients with comorbidity pattern CP_2 is more compared with comorbidity patterns CP_1 & CP_3 in the respective age group. In other words, more number of middle aged diabetic patients will be associated with comorbidity pattern CP_2. Lastly, it is also observed that, more number of male diabetic patients will be associated with the comorbidities patterns compared to female diabetic patients.

Table 10. Clusters of MIMIC-IV dataset obtained by our technique

	Gender		Discretization of Age			Comorbidities Patterns selected from MIMIC-IV as explained in Section 5				
Cluster ID	Male	Female	18 - 51	52 - 70	71 - 91	CP_1	CP_2	CP_3	CP_4	CP_5
0	1892	0	0	1892	0	0	0	0	0	1892
1	926	0	0	926	0	0	0	0	926	0
2	483	0	0	0	483	0	0	0	483	0
3	0	1811	0	1811	0	0	0	0	0	1811
4	0	1044	0	1044	0	1044	0	0	0	0
5	0	1109	0	1109	0	0	0	0	1109	0
6	2129	0	0	0	2129	2129	0	0	0	0
7	0	209	209	0	0	0	0	209	0	0
8	0	1151	0	0	1151	0	0	1151	0	0
9	0	526	526	0	0	0	0	0	0	526
10	0	1671	0	0	1671	1671	0	0	0	0
11	1978	0	0	1978	0	0	0	1978	0	0
12	0	1103	0	1103	0	0	0	1103	0	0
13	1224	0	0	0	1224	0	0	1224	0	0
14	1936	0	0	1936	0	1936	0	0	0	0
15	0	768	0	768	0	0	768	0	0	0
16	334	0	334	0	0	0	0	0	334	0
17	0	189	189	0	0	189	0	0	0	0
18	0	370	370	0	0	0	0	0	370	0
19	1121	0	0	1121	0	0	1121	0	0	0
20	105	0	105	0	0	0	105	0	0	0
21	354	0	354	0	0	0	0	354	0	0
22	1212	0	0	0	1212	0	0	0	0	1212
23	1619	0	0	0	1619	0	1619	0	0	0
24	412	0	412	0	0	412	0	0	0	0
25	698	0	698	0	0	0	0	0	0	698
26	0	703	0	0	703	0	0	0	703	0
27	0	28	28	0	0	0	28	0	0	0
28	0	1581	0	0	1581	0	1581	0	0	0
29	0	1420	0	0	1420	0	0	0	0	1420

Table 10 shows that our clustering method has clustered the MIMIC-IV dataset very well as we can find the records in every cluster that belongs to a certain age, gender and a comorbidity pattern. We also observed that comorbidities patterns (CP_1, CP_2, CP_3, CP_4 & CP_5) depend on the age in case of both male and female diabetic patients. It is also observed that the number of diabetic patients with comorbidity pattern CP_5 is more when compared with the other four comorbidity patterns in the all the three age groups. In the age group 52-70, the number of female diabetic patients with comorbidity pattern CP_4 is more when compared with male diabetic patients with comorbidity pattern CP_4 but in case of other four comorbidities patterns in the respective age groups, the number of male diabetic patients with the corresponding comorbidity pattern is more, when compared with the female diabetic patients with the corresponding comorbidity pattern.

6.3. Validating the Clustering Quality of our Technique using Silhouette Analysis

Cluster quality is assessed using one of the widely used quality indices, such as Silhouette Index (SI). This is due to the fact that the datasets we use lack information on a grouping ground truth, and we are unaware of RT-datasets containing demographics (age, gender) and diagnosis codes (ICD-9-CM codes) that include such a clustering.

The separation distance between the created clusters can be investigated using silhouette analysis. The plot of silhouette shows how close each point in one cluster is to points in adjacent clusters, allowing you to visually investigate aspects like cluster count. This measure ranges from −1 to +1, where a high value implies that the object is matched well to its own cluster and matched poorly to adjacent clusters.

$$Silhouette\ Score\ /\ Silhouette\ Coefficient = \frac{(b-a)}{\max(a,b)} \tag{2}$$

Where,

a = average intra-cluster distance.

b = average inter-cluster distance.

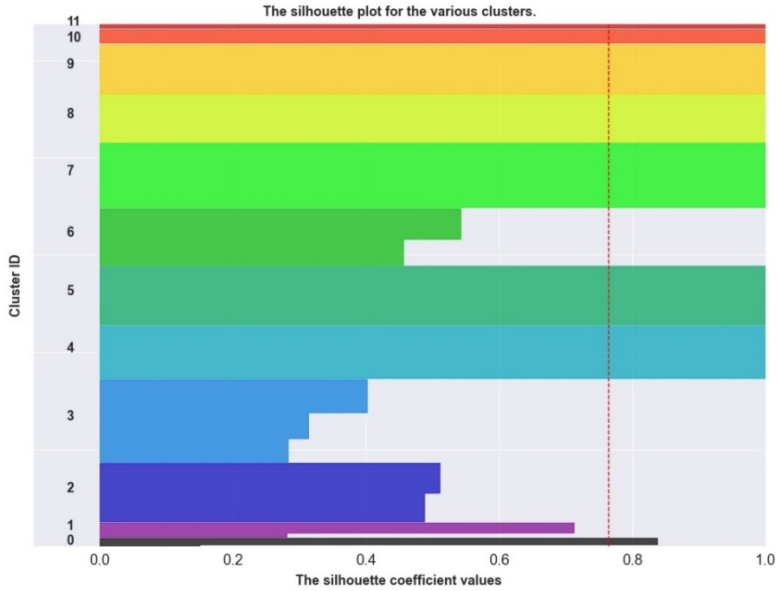

Fig. 8. Validating the clustering quality of our technique on MIMIC - III dataset with k = 12.

As shown in Fig. 8, the silhouette coefficients values for k = 12 (which is selected as explained Section 6.1) are near to 1, which specifies that the sample is far away from the neighboring clusters. We can also observe that not even a single cluster has negative silhouette coefficients values, however for some of the clusters (Cluster 1, Cluster 2, Cluster 3 and Cluster 6) silhouette coefficients values are still less than the average silhouette coefficients value, which is highlighted in red dashed vertical line as shown in Fig. 8. Hence, we can say that k = 12 is not an actual optimal value, so in order to find the actual optimal k value, silhouette analysis was conducted for different k values and from this experiment we found the actual optimal value k =18 and the result obtained for k = 18 is shown in Fig. 9.

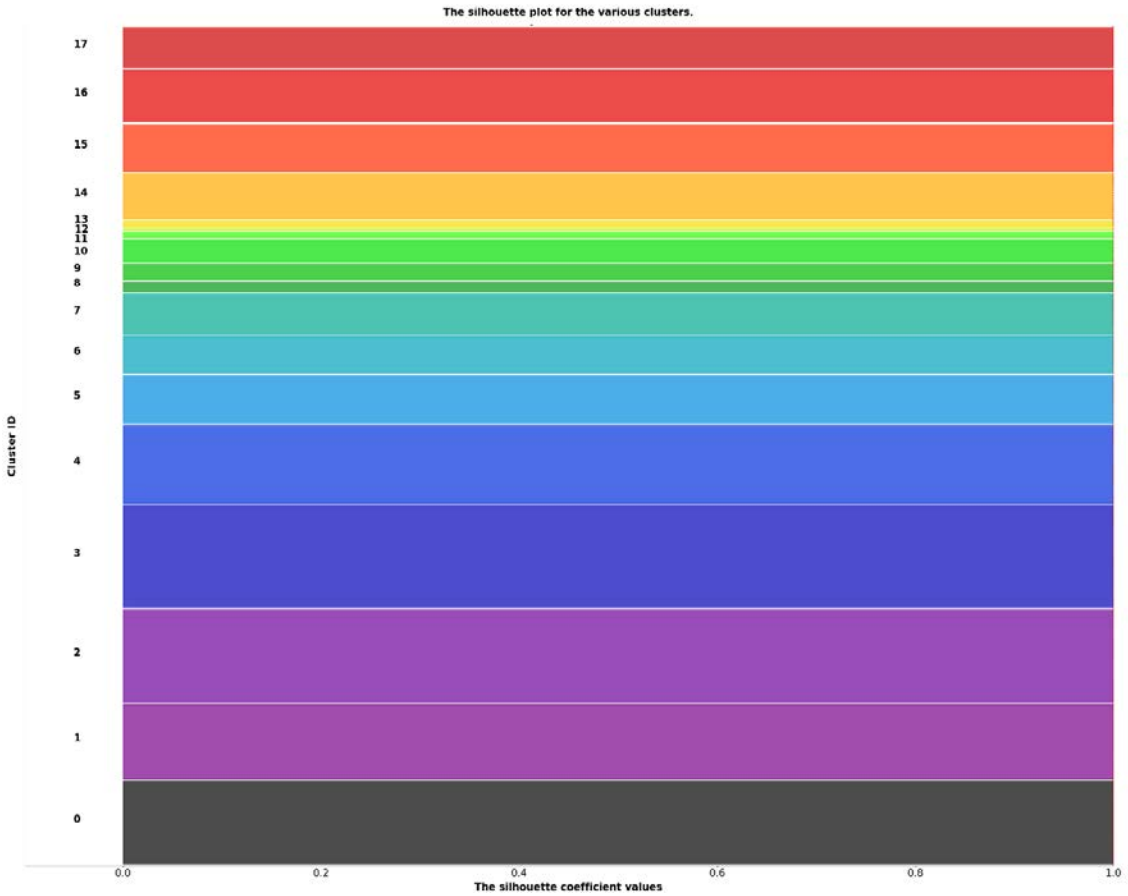

Fig. 9. Validating the clustering quality of our technique on MIMIC - III dataset with k = 18.

Now from the Fig. 9, we can observe that the silhouette coefficients values of all the 18 clusters is equal to 1 i.e., not even a single cluster whose silhouette coefficient value less than 1. As a result, we can claim that the record is matched well to its own cluster and matched poorly to adjacent clusters. The resulting grouping of the MIMIC-III dataset for k = 18 is shown in the Table 9.

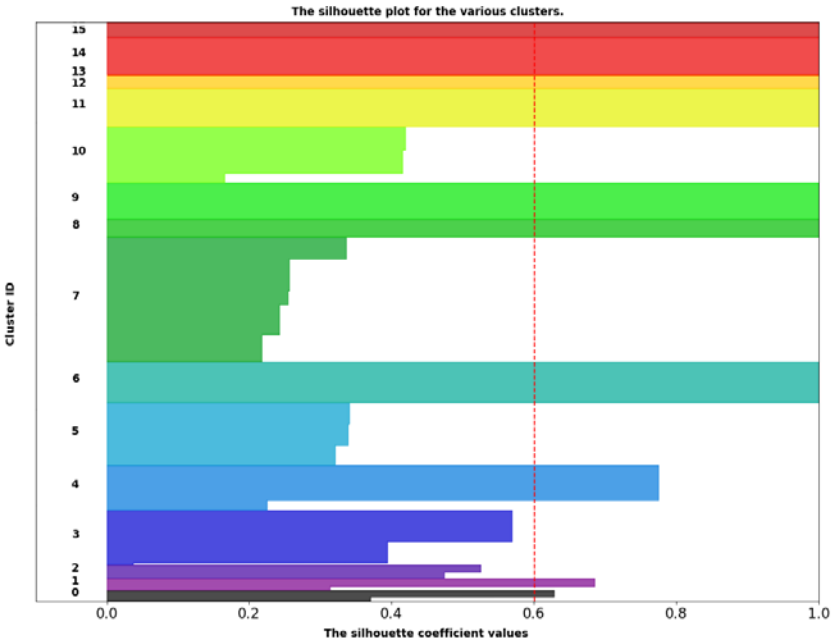

Fig. 10. Validating the clustering quality of our technique on MIMIC - IV dataset with k = 16.

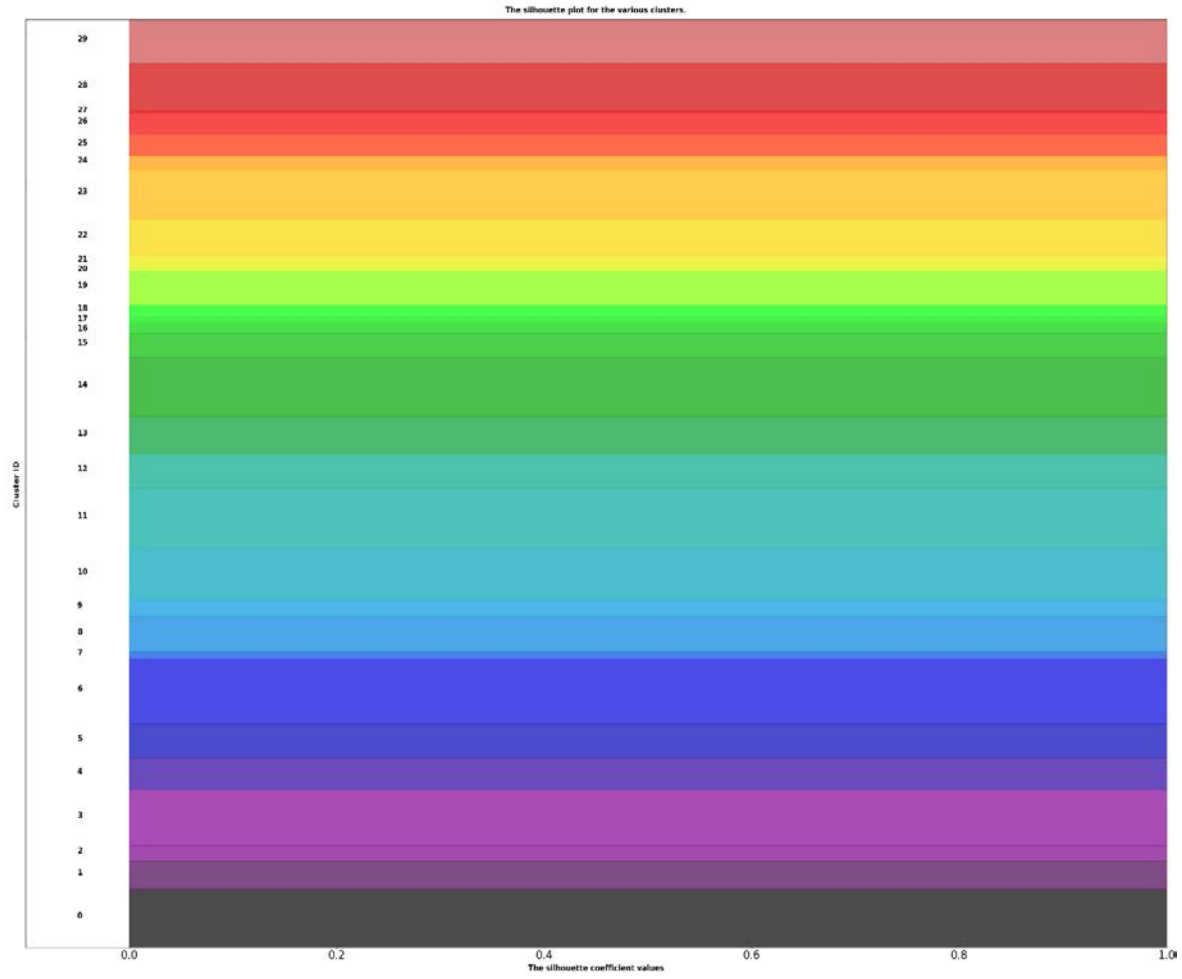

Fig. 11. Validating the clustering quality of our technique on MIMIC - IV dataset with k = 30.

As shown in Fig. 10, the Silhouette coefficients values for k = 16 (which is selected as explained Section 6.1) are near to 1, which specifies that the sample is far away from the neighboring clusters. We can also observe that not even a single cluster have negative silhouette coefficients values, however some of the clusters (Cluster 2, Cluster 3, Cluster 5, Cluster 7 & Cluster 10) silhouette coefficients values are still less than the average silhouette coefficients value, which is highlighted in red dashed vertical line in Fig. 10. Hence, we can say that k = 16 is not an actual optimal value, so in order to find the actual optimal k value, silhouette analysis was conducted for different k values and from this experiment we found the actual optimal value k =30 and the result obtained for k = 30 is shown in Fig. 11.

Now from the Fig. 11 we can observe that the silhouette coefficients values of all the 30 clusters is equal to 1 i.e., not even a single cluster whose silhouette coefficient value less than 1. As a result, we can claim that the record is matched well to its own cluster and matched poorly to adjacent clusters. The resulting grouping of the MIMIC-IV dataset for k = 30 is shown in the Table 10.

7. Conclusion

The special requirements for EHR data clustering, such as data with atomic and set-valued attributes, are not addressed by most traditional clustering methods. This paper proposes an efficient and scalable technique to cluster the data with atomic and set-valued attributes by combining pattern mining (to select top-k diabetes-specific comorbidities patterns), topic modeling (to validate the selected top-k diabetes-specific comorbidities patterns) and pattern-based clustering (to create clusters). The novelty of this technique is that, here we have used the low-dimensional feature vector, which is composed of validated top-k diabetes-specific comorbidities patterns, in place of the original high-dimensional document vector. Because the proposed technique utilizes only the validated top-k diabetes-specific comorbidities patterns instead of all the comorbidity patterns, and low-dimensional feature vector instead of vector space model, our technique is more efficient and scalable than others. Also, our experimental results show that our technique can create a well separated and interpretable clusters with correlated demographics (age, gender) and diagnosis codes (ICD-9-CM codes). Notably, the founded clusters are beneficial for numerous investigative tasks like query answering, visualization, anonymization, classification, etc. and also assist clinical practitioners to prescribe an optimal care for diabetic patients. Finally, as conditions linked with other disease are coded in a similar fashion within electronic health records, our proposed approach can be implemented to similar research concerns for diseases other than diabetes.

References

[1] https://www.healthit.gov/faq/what-electronic-health-record-ehr.

[2] Healthcare Information and Management Systems Society (HIMSS), https://www.himss.org/library/ehr, 2016.

[3] P. Campanella, E. Lovato, C. Marone, L. Fallacara, A. Mancuso, W. Ricciardi, M.L. Specchia, “The impact of electronic health records on healthcare quality: a systematic review and meta-analysis”, Eur. J. Public Health 26 (1) 60–64, 2015.

[4] C. Rinner, S.K. Sauter, G. Endel, G. Heinze, S. Thurner, P. Klimek, G. Duftschmid, “Improving the informational continuity of care in diabetes mellitus treatment with a nationwide shared EHR system: estimates from austrian claims data”, Int. J. Med. Inform. 92 44–53, 2016.

[5] D. Gotz, J. Sun, N. Cao, S. Ebadollahi, “Visual cluster analysis in support of clinical decision intelligence”, in: AMIA Annual Symposium Proceedings, Vol. 2011, pp. 481–490, 2011.

[6] P. Yadav, M. Steinbach, V. Kumar, G. Simon, “Mining electronic health records (EHRs): a survey”, ACM Comput. Surv. 50 (6) 85, 2018.

[7] R.J. Carroll, A.E. Eyler, J.C. Denny, “Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis”, Exp. Rev. Clin. Immunol. 11 (3), 329–337, 2015.

[8] G. Poulis, G. Loukides, S. Skiadopoulos, A. Gkoulalas-Divanis, “Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints”, J. Biomed. Inform. 65, 76–96, 2017.

[9] Centers for Medicare & Medicaid Services, Proposed changes to the CMS-HCC risk adjustment model for payment year 2017, 2015.

[10] A. Kemp, D.B. Preen, C. Saunders, C.D.J. Holman, M. Bulsara, K. Rogers, E.E. Roughead, “Ascertaining invasive breast cancer cases; the validity of administrative and self-reported data sources in Australia”, BMC Med. Res. Methodol. 13 (1) 17, 2013.

[11] G. Tsoumakas, I. Katakis, Multi-label classification: an overview, Int. J. Data Warehouse. Min. (IJDWM) 3 (3) 1–13, 2007.

[12] N. Mohammed, X. Jiang, R. Chen, B.C. Fung, L. Ohno-Machado, “Privacy-preserving heterogeneous health data sharing”, J. Am. Med. Inform. Assoc. 20 (3), 462–469, 2012.

[13] R. Xu, D.C. Wunsch, “Survey of clustering algorithms”, IEEE Trans. Neural Networks 16 (3), 645–678, 2005.

[14] V. Guralnik, G. Karypis, “A scalable algorithm for clustering sequential data”, Proceedings of the 2001 IEEE International Conference on Data Mining, pp. 179–186, 2001.

[15] N. Sokolovska, O. Cappé, F. Yvon, “The asymptotics of semi-supervised learning in discriminative probabilistic models”, Proceedings of the 25th international conference on Machine learning, ACM, pp. 984–991, 2008.

[16] V. Nouri, M.-R. Akbarzadeh-T, A. Rowhanimanesh, “A hybrid type-2 fuzzy clustering technique for input data preprocessing of classification algorithms”, in: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2014, pp. 1131–1138, 2014.

[17] G. Poulis, G. Loukides, A. Gkoulalas-Divanis, S. Skiadopoulos, “Anonymizing data with relational and transaction attributes, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases”, pp. 353–369, 2013.

[18] R. Henriques, F.L. Ferreira, S.C. Madeira, "BicPAMS: software for biological data analysis with pattern-based biclustering", BMC Bioinform. 18 (1) 82, 2017.
[19] A. Zhang, C. Tang, D. Jiang, "Cluster analysis for gene expression data: a survey", IEEE Trans. Knowl. Data Eng. (11) 1370–1386, 2004.
[20] J. Lustgarten, V. Gopalakrishnan, H. Grover, S.V. S, "Improving classification performance with discretization on biomedical datasets", AMIA Annual Symposium Proceedings, pp. 445–449, 2008.
[21] M.J. Zaki, W.M. Jr., W. Meira, "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2014.
[22] S. Guha, R. Rastogi, K. Shim, "ROCK: a robust clustering algorithm for categorical attributes", Inform. Syst. 25 (5) 345–366, 2000.
[23] F. Giannotti, C. Gozzi, G. Manco, "Clustering transactional data", Proceedings of the 2002 European Conference on Principles of Data Mining and Knowledge Discovery, 2002.
[24] A.S. Shirkhorshidi, S. Aghabozorgi, T.Y. Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data", PLOS ONE 10.
[25] P.B. Jensen, L.J. Jensen, S. Brunak, "Mining electronic health records: towards better research applications and clinical care", Nat. Rev. Genet. 13 (6) 395, 2012.
[26] J.A. Hartigan, M.A. Wong, "Algorithm AS 136: a K-means clustering algorithm", J. Roy. Stat. Soc. 28 (1) 100–108, 1979.
[27] D. Arthur, S. Vassilvitskii, "k-means++: The advantages of careful seeding", Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, pp. 1027–1035, 2007.
[28] H. Park, C. Jun, "A simple and fast algorithm for K-medoids clustering", Exp. Syst. Appl. 36 (2) 3336–3341, 2009.
[29] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, "OPTICS: ordering points to identify the clustering structure", in: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, Vol. 28, ACM, pp. 49–60, 1999.
[30] B. Andreopoulos, A. An, X. Wang, D. Labudde, "Efficient layered density-based clustering of categorical data", J. Biomed. Inform. 42 (2) 365–376, 2009.
[31] Y. Yang, X. Guan, J. You, "Clope a fast and effective clustering algorithm for transactional data", Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 682–687, 2002.
[32] H. Yan, K. Chen, L. Liu, "Efficiently clustering transactional data with weighted coverage density", Proceedings of the 15th ACM international conference on Information and knowledge management, ACM, pp. 367–376, 2006.
[33] F. Cao, J.Z. Huang, J. Liang, X. Zhao, Y. Meng, K. Feng, Y. Qian, "An algorithm for clustering categorical data with set-valued features", IEEE Trans. Neural Networks Learn. Syst. 29 (10) 4593–4606, 2018.
[34] L. Kalankesh, J. Weatherall, T. Ba-Dhfari, I.E. Buchan, A. Brass, "Taming EHR data: using semantic similarity to reduce dimensionality", Stud. Health Technol. Inform. 192 52–56, 2013.
[35] F.S. Roque, P.B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge, L.J. Jensen, S. Brunak, "Using electronic patient records to discover disease correlations and stratify patient cohorts", PLOS Comput. Biol. 7 (8) 1–10, 2011.
[36] F. Doshi-Velez, Y. Ge, I. Kohane, "Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis", Pediatrics 133 (1) e54–e63, 2014.
[37] S. Ghassempour, F. Girosi, A. Maeder, "Clustering multivariate time series using hidden markov models", Int. J. Environ. Res. Public Health 11 (3) 2741–2763, 2014.
[38] C.E. Lopez, S. Tucker, T. Salameh, C.S. Tucker, "An unsupervised machine learning method for discovering patient clusters based on genetic signatures", J. Biomed. Inform. 85 30–39, 2018.
[39] A. Ultsch, J. Loetsch, "Machine-learned cluster identification in high-dimensional data", J. Biomed. Inform. 66 95–104, 2017.
[40] H. Xu, Y. Wu, N. Elhadad, P.D. Stetson, C. Friedman, "A new clustering method for detecting rare senses of abbreviations in clinical notes", J. Biomed. Inform. 45 (6) 1075–1083, 2012.
[41] M. Moradi, "CIBS: a biomedical text summarizer using topic-based sentence clustering", J. Biomed. Inform. 88 53–61, 2018.
[42] L. Parsons, E. Haque, H. Liu, "Subspace clustering for high dimensional data: a review", ACM SIGKDD Explor. Newslett. 6 (1) 90–105, 2004.
[43] R. Gwadera, "Pattern-based solution risk model for strategic it outsourcing", in: Industrial Conference on Data Mining, Vol. 7987, pp. 55–69, 2013.
[44] H.-P. Kriegel, P. Kröger, A. Zimek, "Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering", ACM Trans. Knowl. Discov. Data (TKDD) 3 (1) 1, 2009.
[45] C.C. Aggarwal, C. Zhai, "A survey of text clustering algorithms, Mining Text Data", Springer, pp. 77–128, 2012.
[46] B.C. Fung, K. Wang, M. Ester, "Hierarchical document clustering using frequent itemsets", in, Proceedings of the 2003 SIAM international conference on data mining, SIAM, pp. 59–70, 2003.
[47] C. Su, Q. Chen, X. Wang, X. Meng, "Text clustering approach based on maximal frequent term sets", 2009 IEEE International Conference on Systems, Man and Cybernetics, IEEE, pp. 1551–1556, 2009.
[48] G. Kiran, R. Shankar, V. Pudi, "Frequent itemset based hierarchical document clustering using Wikipedia as external knowledge", International Conference on Knowledge-based and Intelligent Information and Engineering Systems, Springer, pp. 11–20, 2010.
[49] S.C. Madeira, A.L. Oliveira, "Biclustering algorithms for biological data analysis: a survey", IEEE/ACM Trans. Comput. Biol. Bioinf. 1 (1) 24–45, 2004.
[50] Y. Cheng, G.M. Church, "Biclustering of expression data", in: International Conference on Intelligent Systems for Molecular Biology, Vol. 8, pp. 93–103, 2000.
[51] I.V. Mechelen, H.-H. Bock, P.D. Boeck, "Two-mode clustering methods: a structured overview", Stat. Methods Med. Res. 13 (5) 363–394, 2004.
[52] A. Tanay, R. Sharan, R. Shamir, "Handbook of computational molecular biology" 9 (1–20) 122–124, 2005.
[53] D.M. Blei, A.Y. Ng, M.I. Jordan, "Latent Dirichlet allocation", J. Mach. Learn. research. 3 993–1022, 2003.
[54] S. Sahni, T. Gonzalez, "P-complete approximation problems", J. ACM (JACM) 23 (3) 555–565, 1976.

[55] A. Czumaj, C. Sohler, “Small space representations for metric min-sum k-clustering and their applications”, Annual Symposium on Theoretical Aspects of Computer Science, Springer, pp. 536–548, 2007.
[56] Johnson, A., Pollard, T., & Mark III, R. (2016). MIMIC-III clinical database. Physio Net, 10, C2XW26.
[57] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark IV, R. (2020). Mimic-iv (version 0.4). PhysioNet.
[58] https://www.cdc.gov/nchs/icd/icd9cm.htm
[59] Bramesh, S. M., & KM, A. K. “An Effective Rule Based Approach for Identification of Comorbidity Patterns in Diabetic Patients”, Indian Journal of Computer Science and Engineering (IJCSE), Vol. 13, No. 4, pp. 1067- 1082, Jul-Aug 2022. DOI: 10.21817/indjcse/2022/v13i4/221304054.
[60] https://radimrehurek.com/gensim/models/ldamodel.html
[61] T.M. Kodinariya, P.R. Makwana, “Review on determining number of clusters in Kmeans clustering”, Int. J. 1 (6) 90–95, 2013.
[62] L. Peng, W. Qing, G. Yujia, “Study on comparison of discretization methods”, in: 2009 International Conference on Artificial Intelligence and Computational Intelligence, Vol. 4, IEEE, pp. 380–384, 2009.
[63] Tabinda Sarwar, Sattar Seifollahi, Jeffrey Chan, Xiuzhen Zhang, Vural Aksakalli, Irene Hudson, Karin Ver-spoor, and Lawrence Cavedon. (2022). “The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges”. ACM Comput. Surv. 55, 2, Article 33, https://doi.org/10.1145/3490234, (January 2022).
[64] Sudhir Anakal, P Sandhya, “Decision Support System for Drug-Drug Interaction Pertaining to COPD and its Comorbidities”, International Journal of Education and Management Engineering, Vol.12, No.2, pp. 1-6, 2022.
[65] Arnold Adimabua Ojugo, Elohor Ekurume, “Predictive Intelligent Decision Support Model in Forecasting of the Diabetes Pandemic Using a Reinforcement Deep Learning Approach”, International Journal of Education and Management Engineering, Vol.11, No.2, pp. 40-48, 2021.

Authors’ Profiles

Bramesh S.M is currently doing Ph.D. under the supervision of Prof. Anil Kumar K.M in part-time, in the research center SJCE, Mysuru under the Visvesvaraya Technological University, Belagavi. He is working as an Assistant Professor, in the Department of Information Science & Engineering, P. E. S. College of Engineering, Mandya, Karnataka, India. He has teaching experience of 14 years and research experience of 5 years. His research interest includes Text mining, and Data mining. He has published nearly 8 Research papers in National and International proceedings.

Anil Kumar K.M is currently working as an Associate Professor, in the Department of Computer Science & Engineering, JSS Science and Technology University, Mysuru, Karnataka, India. He did his postdoc at Deakin University under Professor Jemal Abawajy and Ph.D. from the University of Mysore under the supervision of Prof. Suresha, Chairman, DOS in Computer Science. He has teaching experience of 23 years and research experience of 16 years. His research interest includes Text mining, Sentiment Analysis, Data mining, Opinion mining, Web Mining, Data Analytics, Computer Networks, and Cyber Security. He has received 5 grants from different Government and Private funding agencies for Research & Development. He has published nearly 50 Research papers in National and International proceedings.

How to cite this paper: Bramesh S M, Anil Kumar K M, "An Efficient and Scalable Technique for Clustering Comorbidity Patterns of Diabetic Patients from Clinical Datasets", International Journal of Modern Education and Computer Science(IJMECS), Vol.14, No.6, pp. 35-52, 2022. DOI:10.5815/ijmecs.2022.06.04