-----MEASURING CLINICAL HEALTH DATABASE SIMILARITY USING CLUSTERING AND CLASSIFICATION

Mrs.Amita Mishra[1], P.Vaishnavi[2], T.Nagavendar[2], Y.Shiva Reddy[2],
[1]*Asst.Professor,Computer Science and Engineering, CMR Engineering College, Medchal , T.S, India,*
[2]*B.Tech, Computer Science and Engineering, CMR Engineering College, Medchal, T.S, India*

ABSTRACT

Clustering data derived from Electronic Health Record (EHR) systems is important to discover relationships between the clinical profiles of patients and as a preprocessing step for analysis tasks, such as classification. However, the heterogeneity of these data makes the application of existing clustering methods difficult and calls for new clustering approaches. In this paper, we propose the first approach for clustering a dataset in which each record contains a patient's values in demographic attributes and their set of diagnosis codes. Our approach represents the dataset in a binary form in which the features are selected demographic values, as well as combinations (patterns) of frequent and correlated diagnosis codes. This representation enables measuring si- milarity between records using cosine similarity, an effective measure for binary-represented data, and finding compact, well-separated clusters through hierarchical clustering. Our experiments using two publicly available EHR datasets, comprised of over 26,000 and 52,000 records, demonstrate that our approach is able to construct clusters with correlated demographics and diagnosis codes, and that it is efficient and scalable.

1. Introduction

An Electronic Health Record (EHR) can be defined as an electronic record of the medical and treatment history of a patient [1], which contains (among others) a patient's demographics, diagnoses, medications, and laboratory results. EHRs can benefit healthcare delivery, by reducing documentation time [2] and facilitating the sharing of patient information [3]. In addition, EHRs can improve clinical research and data-driven quality measures through the application of data mining technologies [4]. Such technologies can be used to guide the treatment of patients [5], by partitioning the data into meaningful groups through clustering, or by identifying co-occurring diagnoses (*co- morbidities*) that help prognosis and quality of care assessment, through pattern mining. However, the heterogeneity of EHR data makes several existing mining methods inapplicable to EHR data, calling for new methods to properly deal with data heterogeneity [6].

1.1Motivation

Consider the dataset in Table 1. Each record corresponds to a different patient and contains their demographics and set of diagnosis codes. The dataset contains two types of attributes: (I) Single-valued (or atomic) attributes that contain one value per record[7]. The value can be a number (respectively, category), in which case the attribute is numerical (respectively, categorical). For example, in Table 1, each patient's record contains one value in the numerical attribute *Age* and another value in the categorical attribute *Gender*. (II) Set-valued attri- butes that contain a set of values per record. For example, in Table 1, a patient's record contains a set of diagnosis codes in the *Diagnosis codes* attribute. Datasets containing both single-valued and set-valued attributes are referred to as RT-datasets (for Relational Transaction datasets), and they are useful in several medical analysis application[8].

The task of clustering an RT-dataset, comprised of demographics and diagnosis codes, aims to create meaningful groups of records (*clusters*) that share similar demographics and diagnosis codes[8]. In other words, the task aims to find natural, hidden structures of the data [9]. For example, our method may produce a cluster with male patients under 60 associated with diseases of the respiratory system, another cluster with male patients over 60 associated with mental disorders, and a third cluster of female patients under 40 associated with complications of pregnancy. Furthermore, the records in each cluster contain correlated diagnosis codes affecting many patients, which helps the interpretability of clusters [10]. The created clusters are useful for several analytic tasks, including: (I) visualization (e.g., to obtain insights on patient subpopulations by examining the visualized clusters), (II) query answering (e.g., to derive aggregate statistics about patient sub- populations in different clusters and use them to compare the sub- populations), (III) anonymization [11].

Table 1

A (toy) example of an RT-dataset. *Gender*, F is for Female and M for Male. The diagnosis codes are represented as ICD-9 codes [12]. The attribute *ID* is for reference.

ID	Gender	Age	Diagnosis codes
1	F	77	250.00, 272.4, 278.01, 401.9
2	M	71	244.9, 285.1, 530.81
3	F	46	421.0, 427.31, 584.9
4	F	78	250.00, 272.4, 401.9, 414.8
5	M	73	244.9, 530.81, 648.01, 661.11
6	F	48	285.1, 427.31, 584.9
7	F	80	196.6, 250.00, 272.4, 401.9
8	M	73	244.9, 401.9, 530.81
9	F	48	427.31, 584.9, 693.0
10	F	75	250.00, 272.4, 401.9, 560.1
11	M	73	218.0, 244.9, 530.81
12	F	49	427.31, 584.9, 995.91

to algorithms that transform the values in each cluster to protect patient privacy), and (IV) classification (e.g., to preprocess a dataset in order to derive classes of records, which can subsequently be used for per- forming classification more efficiently and effectively [13]). Thus, one can cluster an RT-dataset and then use the clustering result in one or more of these tasks[14].

However, existing clustering algorithms are not designed to cluster an RT-dataset comprised of demographics and diagnosis codes. This is because, as explained in [15]:

(I) Most clustering algorithms use a single similarity measure, and it is difficult to design a measure that captures the similarity of records with both single-valued and set-valued attributes[16]. The reason is that single-valued attributes, such as demographics, and set-valued attributes, such as the attribute comprised of diagnosis codes, have different semantics. That is, there is one value per record in a demographic attribute, among a relatively small number of possible values, while there is a large number of diagnosis codes per record, among a very large number of possible diagnosis codes[17]. This makes it difficult to find a single function (similarity measure) to capture how similar the demographics and diagnosis codes of two or more records are. For instance, Euclidean distance, which is applicable to numerical demographics, is not suitable for measuring distance between sets of diagnosis codes, and Jaccard distance, which is applicable to sets of diagnosis codes, is not suitable for measuring distance between numerical demographics[18].

(II) Multi-objective clustering algorithms that aim to optimize several measures simultaneously are not suitable for RT-datasets. For example, using a composite measure, such as the weighted sum or product of Euclidean distance (applied to demographics) and Jaccard distance (applied to diagnosis codes), may lead to low-quality clusters. This is because each of these measures has often a different distribution of values, which means that the composite measure does not capture cluster quality well. In addition, two-level (hybrid) optimization strategies, which first try to cluster demographics and then diagnosis codes (e.g., the strategy used in [19]) are not able to find high quality clusters, as shown in our experiments.

Contributions

We propose the first clustering approach that is designed for an RT- dataset comprised of demographics and diagnosis codes. The main idea of our approach is to construct a record representation that allows measuring similarity between records, based on both demographics and diagnosis codes.

To construct such a representation, we *discretize* [20] numerical demographics (i.e., replace their values with aggregate values) and select subsets of the diagnosis codes contained in the dataset, which are referred to as *patterns*. Then, we represent each record of the RT-dataset using one-hot encoding, producing a *binary representation* of the dataset (see Table 2). The features (columns) in the binary representation are:

(I) the values in each discretized numerical demographic attribute, (II) the values in each categorical demographic attribute, and (III) the selected patterns. A value of 1 (respectively, 0) in a feature of the binary representation implies that the record contains (respectively, does not contain) the feature. Based on the binary representation, we construct clusters comprised of similar records, by applying a clustering algorithm that is suitable for binary-represented data [21].

Yet, there are two challenges that need to be tackled to realize our approach. First, we need a way to select patterns that help constructing a high-quality clustering. Second, we need a way to construct high- quality clusters efficiently. We address these challenges by proposing two methods; Maximal-frequent All-confident pattern Selection (MAS) and Pattern-based Clustering (PC):

1. The MAS method selects patterns that:

Table 2

Binary representation of the RT-dataset in Table 1. The features in the binary representation are: { } F and { } M , the values in the categorical demographic attribute Gender; {45, ,49}, {70, ,74} ... ... , and {75, ,80} ... , the values of the discretized numerical attribute Age; and {250.00, 272.4, 401.9}, {244.9, 530.81}, and {427.31, 584.9}, the patterns comprised of diagnosis codes. The binary representation is clustered into three clusters, with Cluster IDs 1, 2, and 3. The attributes Cluster ID and ID are for reference[23].

Cluster ID	ID	{F}	{M}	{45,...,49}	{70,...,74}	{75,...,80}	{250.00,272.4,401.9}	{244.9,530.81}	{427.31,584.9}
1	1	1	0	0	0	1	1	0	0
1	4	1	0	0	0	1	1	0	0
1	7	1	0	0	0	1	1	0	0
1	10	1	0	0	0	1	1	0	0
2	2	0	1	0	1	0	0	1	0
2	5	0	1	0	1	0	0	1	0
2	8	0	1	0	1	0	0	1	0
2	11	0	1	0	1	0	0	1	0
3	3	1	0	1	0	0	0	0	1
3	6	1	0	1	0	0	0	0	1
3									

(I) occur in a large number of records,

(II) are comprised of correlated diagnosis codes (i.e., any codes in the pattern imply the other codes in the pattern with high probability), and

(III) co-occur in a large number of records, when the patterns share diagnosis codes[22].

hypertension", respectively) and appears in records with IDs 1, 4, 7, and 10. Whenever 250.00 is contained in a record,

Table 3

A clustered RT-dataset produced from the binary representation in Table 2. The attributes *Cluster ID* and *ID* are for reference.

Cluster ID	ID	Gender	Age	Diagnosis codes
1	1	F	77	250.00, 272.4, 278.01, 401.9
1	4	F	78	250.00, 272.4, 401.9, 414.8
1	7	F	80	196.6, 250.00, 272.4, 401.9
1	10	F	75	250.00, 272.4, 401.9, 560.1
2	2	M	71	244.9, 285.1, 530.81
2	5	M	73	244.9, 530.81, 648.01, 661.11
2	8	M	73	244.9, 401.9, 530.81
2	11	M	73	218.0, 244.9, 530.81
3	3	F	46	421.0, 427.31, 584.9
3	6	F	48	285.1, 427.31, 584.9
3	9	F	48	427.31, 584.9, 693.0
3	12	F	49	427.31, 584.9, 995.91

diagnosis codes 250.00, 272.4, and 401.9 (among others), corresponds to the record with ID 1 in

Table 2, which contains 1 in the feature {*F*} for *Gender*, the discretized value {75, ..., 80} for *Age*, and the pattern
{250.00, 272.4, 401.9} for *Diagnosis codes*. The binary-represented data in Table 2 are then grouped into three clusters, so that records in the same cluster share similar features (which implies that they have similar demographics and diagnosis codes). For example, the cluster with Cluster ID 1 in Table 2 is comprised of the records with IDs 1, 4, 7 and 10. These records correspond to female patients between 75 and 80 and contain all diagnosis codes in the pattern {250.00, 272.4, 401.9}[24]. Next, the clustered RT-dataset in Table 3 is constructed, by simply adding into each cluster the records from the RT-dataset in Table 1 which were clustered together in the binary representation in Table 2.

2. Related work

In this section, we discuss the methods that are closer to EHR clustering and the problem we study. For extensive surveys on mining EHR data, the reader is referred to. Section 2.1 discusses clus- tering for EHR data, Section 2.2 provides a brief overview of pattern- based clustering, and Section 2.3 discusses pattern mining on EHR data[25].

EHR data clustering

We categorize existing methods for clustering EHR data, based on the type of data they are applied to. of demographic attributes are inherently low-dimensional (e.g., they typically include fewer than 20 demographics). Thus, they can be clustered using many existing clustering algorithms,clustering algorithms, such as hierarchical (e.g., single-linkage, average-linkage, and complete linkage)partitional (e.g., *k*-means [26], *k*-means++ [27], and *k*-medoids, and density-based (e.g., DBSCAN and OPTICS [28]) algorithms. For example, an interesting recent work [29] applies density- based clustering on patient data. The reason that the aforementioned algorithms are not suitable for clustering an RT-dataset is that their similarity measures cannot be applied to a set-valued attribute, such as the attribute containing diagnosis codes. This is because their measures cannot capture similarity effectively for set-valued attributes. The reason is that set-valued attributes (such as the diagnosis codes attribute) are inherently high dimensional and the similarity between high dimensional data records cannot be captured effectively by distance measures used in many existing algorithms, as explained in [30]. This is known as the curse of high dimensionality.

Diagnosis codes Similar to RT-datasets, datasets in which each re- cord is comprised of a set of diagnosis codes are inherently high-dimensional (i.e., the set-valued attribute typically contains thousands of different diagnosis codes). Such datasets can be clustered using algorithms developed for set-valued (also referred to as transaction) data. Examples of such algorithms are CLOPE [31], SCALE, ROCK, and SV-*k*-modes. For example, SV-*k*-modes works similar to *k*- means, but it uses a set of values in a set-valued attribute as cluster representative, instead of the centroid used in *k*-means. This algorithm can be applied to datasets that also have single-valued attributes (i.e., RT-datasets), by using centroids as representatives in single-valued at- tributes. However, SV-*k*-modes is applicable only to set-valued attributes with a very small domain size (i.e., attributes with 10–50 distinct values), due to its exponential time complexity with respect to the domain size of set-valued attributes. Consequently, unlike our approach, SV-*k*-modes is not suitable to cluster a set-valued attribute comprised of diagnosis codes, whose domain size is in the order of thousands.

Other high-dimensional data There are clustering methods that are applied to a dataset comprised of trajectories, genomic sequences, or text. For example, focus on clustering trajectory data, in which trajectories represent sequences of diseases, while study clustering genomic data. study clustering medical text. Clearly, these methods cannot be considered as alternatives to our approach, because the data they are applied to have very different semantics compared to those of the attributes in an RT-dataset.

Pattern-based clustering

Pattern-based clustering methods represent the records of a dataset to be clustered in a binary form, in which the features are patterns, and then apply clustering to the binary-represented data. Thus, our approach is related to pattern-based clustering methods. Since the number of patterns is typically smaller than that of the distinct values in the dataset, pattern-based clustering is an effective method for clustering high dimensional data.

Gene expression data Gene expression data are typically re- presented as a real-valued matrix, where each row corresponds to a gene and each column to a condition. However, there is a sig- nificant difficulty to cluster such a matrix because, only under specific experimental conditions, a group of genes can show the same activation patterns. For handling this difficulty, bi-clustering methods, which simultaneously cluster the rows and columns of the matrix, have been proposed. Bi-clustering methods (e.g.,) produce, as clusters, submatrices in which subgroups of genes exhibit highly correlated activities for subgroups of conditions. Some of these methods also employ patterns for bi-clustering, such as *frequent patterns* and *association rules* (see Section 3 for details). All bi-clustering methods aim to simultaneously cluster the rows and columns of the matrix. Thus, contrary to our approach, they cannot be used to group records into clusters. Specifically, a row corresponding to a record in an RT-dataset could participate in multiple clusters, if we applied bi-clustering to the binary representation in PC.

Pattern mining on EHR data

Pattern mining is an important task aiming to discover associated attribute values, often in a dataset comprised of one set-valued attribute. The values in the set-valued attribute are referred to as *items*. There are different ways of modeling associations, leading to different representations of patterns, such as *frequent*, *maximal-frequent*, and *all-confident* patterns (see Section 3 for details).

3. Background and problem statement

In this section, we introduce some preliminary concepts and the problem we aim to solve. Table 4 summarizes the acronyms used in the paper.

3.1.RT-datasets

We consider an RT-dataset $\mathcal{D}$, in which every record corresponds to a distinct patient. Each record *r* in is comprised of one or more demographic attributes that can be numerical or categorical, and of a set- valued attribute containing diagnosis codes. Without loss of generality, we assume that the first *l* attributes in $\mathcal{D}$, denoted with $\mathcal{A}^1,\ldots\mathcal{A}^l$, are demographic attributes, and the last attribute $\mathcal{A}^{l+1}$ is a set-valued attribute. The diagnosis codes can be represented in different formats. For example, they can be ICD-9 codes or ICD-10 codes. It is also easy to convert ICD-10 codes into ICD-9 codes, using General Equivalence Mapping. Extensions to RT-datasets comprised of more than one set-valued attributes are straightforward (see Section 7).

3.2. Maximal-frequent all-confident itemsets

We introduce the concept of Maximal-frequent all-confident itemset (MFA) used in our approach. In the following, we present the definitions and refer the reader to C for examples.

A subset $I \subseteq \mathcal{A}^{l+1}$ is called an itemset. The number of items in I is denoted with I and referred to as the length of I. An itemset $\not\subset$ I′ is a subitemset of I′ , and I′ is a superitemset of I. In our case, each record r of $\mathcal{D}$ contains an itemset that is comprised of the diagnosis codes contained in r. The frequency of an itemset I in $\mathcal{D}$ is defined as the number of records in $\mathcal{D}$ that have I as their subitemset, and it is denoted with $fr_{\mathcal{D}}(I)$. The support (relative frequency) of an itemset I in $\mathcal{D}$ is defined as $sup_{\mathcal{D}}(I) =$ [illegible] where $\mathcal{D}$ is the number of records in $\mathcal{D}$.

Given itemsets *X* and *Y* such that the itemset $I = X \cup Y$ has length $X \cup Y| \geq 2$ and a dataset $\mathcal{D}$, the *all-confidence* of *I* in $\mathcal{D}$ can be defined as follows:

There are several algorithms for mining MFIs. In our work, we employ the FPMAX algorithm, because it is more efficient than the algorithms in, as explained in. We then construct the set of MFAs by keeping each MFI *I* with
$allConf_{\mathcal{D}}(I) \geq minAc$.

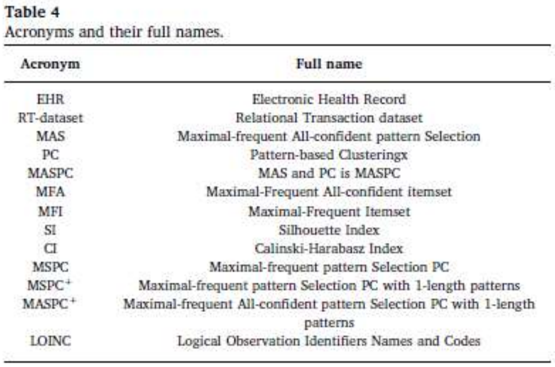


Table 4
Acronyms and their full names.

Acronym	Full name
EHR	Electronic Health Record
RT-dataset	Relational Transaction dataset
MAS	Maximal-frequent All-confident pattern Selection
PC	Pattern-based Clusteringx
MASPC	MAS and PC is MASPC
MFA	Maximal-Frequent All-confident itemset
MFI	Maximal-Frequent Itemset
SI	Silhouette Index
CI	Calinski-Harabasz Index
MSPC	Maximal-frequent pattern Selection PC
MSPC⁺	Maximal-frequent pattern Selection PC with 1-length patterns
MASPC⁺	Maximal-frequent All-confident pattern Selection PC with 1-length patterns
LOINC	Logical Observation Identifiers Names and Codes

3.3 Problem statement

We now formally define the clustering problem that we aim to solve.

Problem1. Given an RT-dataset $\mathcal{D}$, a binary representation $=\{B_1, \ldots, B_{|\ |}\}$ of $\mathcal{D}$, and a parameter *k*, construct a partition $C = \{c_1, \ldots, c_k\}$ of $\mathcal{D}$ 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 $cos(B_r, B_{r'}) = \frac{B_r \cdot B_{r'}}{\|B_r\| \|B_{r'}\|}$ is the cosine similarity measure between the records B_r $\|B_r\|\|B_{r'}\|$ $_{r'}$ in $\mathcal{B}$, which correspond to the records and r , respectively,′ in c_i.

The problem takes as input an RT-dataset $\mathcal{D}$, the number of clusters k, and the binary representation of , and it requires finding a clustering of k clusters for such that the records in each cluster are similar. Specifically, it seeks to maximize the total cosine similarity, which is measured on the records of the binary representation of $\mathcal{D}$. The problem is NP-complete (this follows easily from), which justifies the development of heuristics.

4. Clustering RT-datasets using MASPC

MASPC applies: (I) the *MAS* algorithm, which discovers maximal-frequent all-confident patterns (MFAs), and (II) the *PC* algorithm, which constructs and clusters the binary representation of the RT-dataset, and produces the clustered RT-dataset.

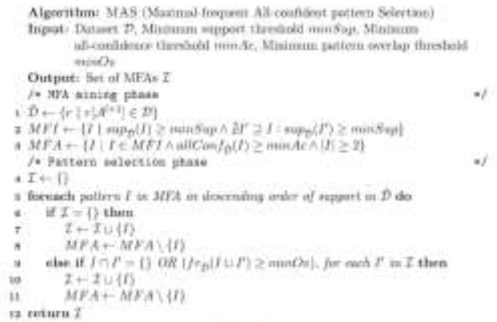
In the following, we explain the operation of the MAS and PC algorithms.

4.1 *The MAS algorithm*

MAS works in two phases:

(I) **MFA mining**: In this phase, all MFAs are mined from the input RTdataset.

(II) **Pattern selection**: In this phase, a subset of MFAs that help the subsequent clustering by the PC algorithm to construct clusters of
high quality are selected.



```
Algorithm: MAS (Maximal-frequent All-confident pattern Selection)
Input: Dataset $\mathcal{D}$, Minimum support threshold $minSup$, Minimum
       all-confidence threshold $minAc$, Minimum pattern overlap threshold
       $minOv$
Output: Set of MFAs $\mathcal{I}$
/* MFA mining phase                                              */
$\tilde{\mathcal{D}} \leftarrow \{r \mid r[\mathcal{A}^{|r|}] \in \mathcal{D}\}$
$MFI \leftarrow \{I \mid sup_{\tilde{\mathcal{D}}}(I) \geq minSup \wedge \nexists I' \supset I : sup_{\tilde{\mathcal{D}}}(I') \geq minSup\}$
$MFA \leftarrow \{I \mid I \in MFI \wedge allConf_{\tilde{\mathcal{D}}}(I) \geq minAc \wedge |I| \geq 2\}$
/* Pattern selection phase                                       */
$\mathcal{I} \leftarrow \{\}$
foreach pattern $I$ in $MFA$ in descending order of support in $\tilde{\mathcal{D}}$ do
    if $\mathcal{I} = \{\}$ then
        $\mathcal{I} \leftarrow \mathcal{I} \cup \{I\}$
        $MFA \leftarrow MFA \setminus \{I\}$
    else if $I \cap I' = \{\}$ OR $(fr_{\tilde{\mathcal{D}}}(I \cup I') \geq minOv)$, for each $I'$ in $\mathcal{I}$ then
        $\mathcal{I} \leftarrow \mathcal{I} \cup \{I\}$
        $MFA \leftarrow MFA \setminus \{I\}$
return $\mathcal{I}$
```

Fig. 1. Algorithm mas

5. Baselines for clustering RT-datasets

1. **Binary representation construction**: This phase is similar in the binary representation), k-PC; the difference is that each diagnosis code in is considered as a feature (column whereas PC considers each MFA as a feature instead. The resultant binary-represented dataset is denoted with $\mathcal{M}_{Hybrid}$.
2. **Medoids clustering**: In this phase, $\mathcal{M}_{Hybrid}$ is partitioned into p clusters, by applying an efficient to the binary representation construction phase of implementation of the k-medoids algorithm with cosine similarity multiple times. First, we apply the algorithm of Park et al. to the projection of $\boldsymbol{M}_{Hybrid}$ on the features (columns) corresponding to demographics, setting the number of clusters to a threshold a. Then, we apply the the projection of *Hybrid* on the

features corresponding to diagnosis codes of the cluster and partition it into *b* smaller clusters. Therefore, at the end of this phase, $a{\cdot}b = p$ clusters are created. algorithm to each of the resultant *a* clusters separately. Specifically, we consider the projection of *Hybrid* on the features corresponding to diagnosis codes of the cluster and partition it into *b* smaller clusters. Therefore, at the end of this phase, $a{\cdot}b = p$ clusters are created.

6. Experimental evaluation

6.1. Datasets

We used two publicly available RT-datasets, comprised of demographics and diagnosis codes:

- VERMONT . The dataset contains de-identified inpatient discharge data in Vermont during 2015.
- INFORMS . The dataset contains de-identified patient data and was used in the Informs Data Mining Contest 2008.

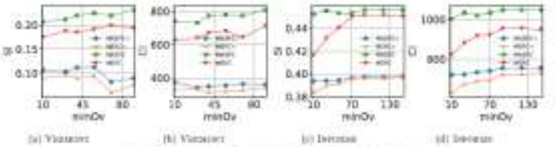
6.2. Experimental setup

We compared our approach against the four baseline methods in Section 5, because no existing clustering algorithms can be applied to an RT-dataset comprised of demographics and diagnosis codes (see Section 2).

6.3. Clustering quality measurement

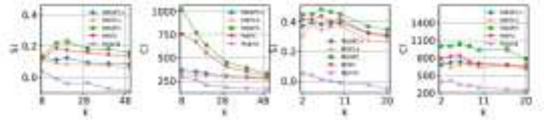
In this section, we show the superiority of MASPC over its 3 variations and HYBRID in terms of being able to construct a high-quality clustering, comprised of compact and well-separated clusters. We consider the impact of parameters *minSup*, *minAc*, *minOv*, and *k*. We omit HYBRID from the results of all experiments in which it performed much worse than all other methods.

6.4. Efficiency of computation



(a) VERMONT (b) VERMONT (c) INFORMS (d) INFORMS

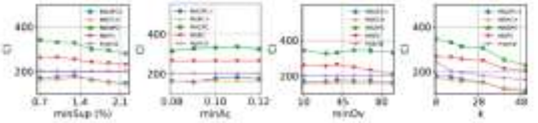
Fig. 4. SI and CI scores vs. minimum pattern overlap threshold minOv.



(a) VERMONT (b) VERMONT (c) INFORMS (d) INFORMS

Fig. 5. SI and CI scores vs. number of clusters k.

6.5. Overview of demographics and patterns in clusters



(a) VERMONT (b) VERMONT (c) VERMONT (d) VERMONT

Fig. 6. CI scores vs. minSup, minAc, minOv and k for VERMONT when unclustered records are clustered together.

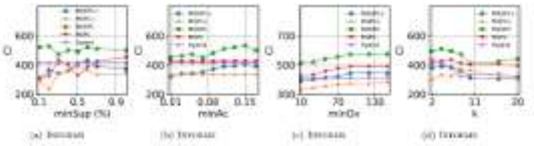


Fig. 7. CI scores vs. minSup, minAc, minDv, and k [illegible]

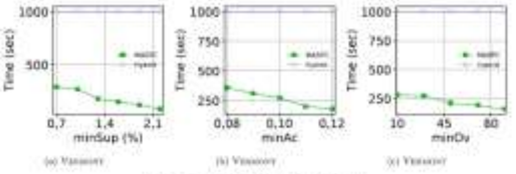
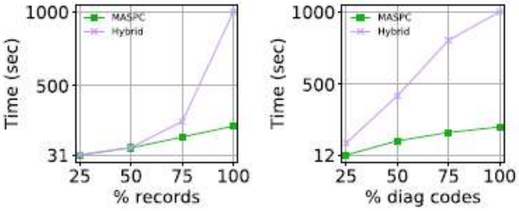


Fig. 8. Runtime vs. (a) minSup, (b) minAc, and (c) minDv.

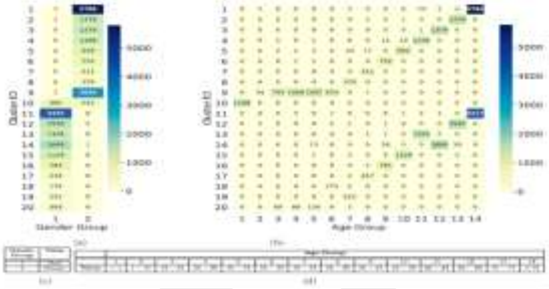


(a) VERMONT (b) VERMONT

Fig. 9. Runtime vs. (a) dataset size, and (b) number of diagnosis codes.

7. Extensions and limitations

We have shown, through extensive experiments, that our approach significantly outperforms baselines in terms of compactness and separation of clustering as well as in terms of runtime. We have also shown that the clusters created by our approach allow finding correlations between diagnosis codes and between diagnosis codes and demographics that have been documented in the medical literature.



8. Conclusion

The task of clustering an RT-dataset whose records are comprised of demographic values and sets of diagnosis codes is important, to discover relationships between the clinical profiles of patients and as a preprocessing step for classification and anonymization. Motivated by this and by the fact that existing clustering (or bi clustering) algorithms are not appropriate for this task, we proposed a new clustering approach.

Our approach constructs a binary representation of the input RT-dataset, in which the features corresponding to diagnosis codes are MFAs (maximal-frequent all-confident patterns), and it

applies clustering to that representation. Our experiments with two large, publicly-available datasets containing about 26,000 and 53,000 records respectively demonstrate
the effectiveness and efficiency of our approach. In particular, they show that our approach outperforms four baselines in terms of clustering quality, it is able to construct clusters with correlated demographics and diagnosis codes, and it is efficient and scalable.

References

[1] Healthcare Information and Management Systems Society (HIMSS), <https:// www.himss.org/library/ehr>, 2016.
[2] Srihari Rao Nidamanuru, Rajesh Tiwari, Bhaskar Koriginja, Jincy Denny, J Ramesh Babu, “Identifying the Websites that Maintain Operational Standards through Obligation Links to Website-Standards Approval Body”, Turkish Journal of Computer and Mathematics Education, Vol. 12, Issue 10, 2021, pp 4456-4461, e-ISSN: 1309-4653.
[3] 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 (2016) 44–53.
[4] Srihari Rao Nidamanuru, Rajesh Tiwari, Bhaskar Koriginja, Jincy Denny, J Ramesh Babu, “Identifying the Websites that Maintain Operational Standards through Obligation Links to Website-Standards Approval Body”, Turkish Journal of Computer and Mathematics Education, Vol. 12, Issue 10, 2021, pp 4456-4461, e-ISSN: 1309-4653.
[5] P. Yadav, M. Steinbach, V. Kumar, G. Simon, Mining electronic health records
(EHRs): a survey, ACM Comput. Surv. 50 (6) (2018) 85.
[6] 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) (2015) 329–337.
[7] 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 (2017) 76–96.
[8] Centers for Medicare & Medicaid Services, Proposed changes to the CMS-HCC risk adjustment model for payment year 2017, 2015.
[9] 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)
(2013) 17.
[10] G. Tsoumakas, I. Katakis, Multi-label classification: an overview, Int. J. Data
Warehous. Min. (IJDWM) 3 (3) (2007) 1–13.
[11] 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) (2012) 462–469.
[12] R. Xu, D.C. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Networks 16 (3) (2005) 645–678.
[13] V. Guralnik, G. Karypis, A scalable algorithm for clustering sequential data,
Proceedings of the 2001 IEEE International Conference on Data Mining, 2001, pp.
179–186.
[14] N. Sokolovska, O. Cappe, F. Yvon, The asymptotics of semi-supervised learning in
discriminative probabilistic models, Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 984–991.

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

[16] 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, 2013, pp. 353–369.

[17] R. Henriques, F.L. Ferreira, S.C. Madeira, BicPAMS: software for biological data analysis with pattern-based biclustering, BMC Bioinform. 18 (1) (2017) 82.

[18] A. Zhang, C. Tang, D. Jiang, Cluster analysis for gene expression data: a survey, IEEE Trans. Knowl. Data Eng. (11) (2004) 1370–1386.

[19] National Center for Health Statistics, International Classification of Diseases – Ninth Revision, <https://www.cdc.gov/nchs/icd/icd9cm.htm>, 2015.

[20] J. Lustgarten, V. Gopalakrishnan, H. Grover, S.V. S, Improving classification performance with discretization on biomedical datasets, AMIA Annual Symposium Proceedings, 2008, pp. 445–449.

[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) (2000) 345–366.

[23] Dr. C.N. RAVi , D. Palanivel Rajan, Desa Uma Vishweshwar, Edem Sureshbabu (CMREC)'A Review on Various Cloud-Based Electronic Health Record Maintenance System for COVID-19 Patients'Name of the Journal with ISSN:*Advances in Cognitive Science and Communications*, Cognitive Science and Technology - Springer Nature Singapore Pte Ltd. 2023 (CMREC Conference- ICCCE-2022)Vol. / Issue /PP. No. / Date/Month & Year of Publication:Impact Factor: https://doi.org/10.1007/978-981-19-8086-2_15, April 2023.

[24] Mrs.G.Sumalatha1 , Y.Jaideep Naidu M.Srividya, Karra karthik Reddy , D.Niharika, 'Smart OCR for Document Digitization', JASC: Journal of Applied Science and Computations, ISSN NO: 1076-5131, Volume VIII, Issue III, Macrh/2021.

[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) (2012) 395.

[26] J.A. Hartigan, M.A. Wong, Algorithm AS 136: a K-means clustering algorithm, J. Roy. Stat. Soc. 28 (1) (1979) 100–108.

[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, 2007, pp. 1027–1035.

[28] H. Park, C. Jun, A simple and fast algorithm for K-medoids clustering, Exp. Syst. Appl. 36 (2) (2009) 3336–3341.

[29] Rajesh Tiwari, Manisha Sharma and Kamal K. Mehta, "Improve the Execution Time by using GPU for Complex Application with SIMD " , International Journal of Scientific Research(IJSR), special issue March 2018, pp 227 – 232 , ISSN: 0976 – 2876. List Sr. No. 1240, Journal No. 20876.

[30] B. Andreopoulos, A. An, X. Wang, D. Labudde, Efficient layered density-based clustering of categorical data, J. Biomed. Inform. 42 (2) (2009) 365–376.

[31]. A Constructive Feature Grouping Approach for Analyzing the Feature Dominance to Predict Cardiovascular Disease | https://link.springer.com/chapter/10.1007/978-981-19-8086-2_62