Clinico-genomic Data Analytics for Precision Diagnosis and Disease Management

Anuja Kench*, Vandana P. Janeja*, Yelena Yesha*, Naphtali Rishe[†], Michael Grasso[‡], Amanda Niskar[§]
* University of Maryland Baltimore County, Baltimore, USA
Email: {akench1, vjaneja, yeyesha} @umbc.edu
[†] Florida International University, Florida, USA
Email: Rishe@fiu.edu
[‡] University of Maryland School of Medicine, Baltimore, USA
Email: mgrasso@em.umaryland.edu
[§] Arthritis Foundation, USA
Email: aniskar@arthritis.org

***Abstract*—Patient data can be present in clinical notes, lab results, genomic data sources, environmental and geospatial data sources and tissue banks to name a few. A holistic view of the patient's health can be achieved when relevant data from multiple heterogeneous sources are extracted and analyzed in a personalized manner. Moreover, comparative analysis of patients can be performed when multiple patient records are viewed across these heterogeneous data sources. To address this need, we propose clinico-genomic data analytics to enhance personalized medicine treatment decisions using heterogeneous, high dimensional, sparse and massive datasets. We utilize this framework to discover similar patients and overlaps among patients in a set of features towards the goals of: (1) better cohort discovery for clinical trials, (2) better disease management by studying peer group of patients with similar diagnosis but better prognosis, (3) early disease diagnosis by identifying similar features in patients with the existing diagnosis.**

We propose novel approach in two areas: (1) integrating clinical and genomic data of patients and (2) combined data analytics in such heterogeneous datasets. Our approach is modeled as a unified clustering algorithm for finding correlations among clinical and genomic factors of patients. We integrate data containing risk causing Single Nucleotide Polymorphism's (SNP's) known from literature with clinical records of patients. In such heterogeneous data, we propose a combined similarity measure for numeric and nominal data attributes, which we use in our clustering algorithm. Our results show compelling overlaps among patients in the same cluster. These patients had high pairwise similarity and emulated the real world similarities between patients with co-morbid diseases.

Keywords
Personalized Medicine, Single Nucleotide Polymorphism, correlations, clustering.

I. Introduction

Patient data can be present in clinical data, genomic data sources, environmental and geospatial data sources, tissue banks and personal genetic data. A complete view of the patient's health can be achieved when relevant data from all these sources are extracted and modeled in a personalized manner. Moreover, comparative analysis of patients can be performed when multiple patient records are viewed across these heterogeneous data sources. In this research, we propose clinico-genomic data analytics for precision diagnosis and better disease management utilizing datasets that are heterogeneous, high dimensional, sparse and massively large at the same time. The need for such a system is motivated by the following example.

Motivating Example: Let us consider a female, South east Asian, age 55 years with Diabetes. We extract the neighborhood information of this patient and public data available for this race and gender for various disease factors. We extract the clinical, genomic and environmental factors of this patient. Let us say this patient does not exhibit the SNP A1, which is associated with Diabetes. However, when we cluster this patient's data with other patients we find this patient has 80% overlap with other diabetic patients but does not exhibit the SNP A1, which other patients in her cluster possess. Other clinical factors are highly overlapping. This information can be viewed by the patient, her caregiver, public health researcher and clinical scientists in a different manner. For example, the caregiver can focus on better treatment based on the distinction of this patient from other diabetes patients so that the care can be better managed in a personalized manner. A clinical scientist may look at this information to further investigate potential environmental factors or demographic factors, which may have caused this genomic deviation for her. The public health researcher may consider evaluating neighborhood demographics which may have resulted in aggravating the clinical factors. The patient herself may use this knowledge to understand the disease prognosis and public sources of help available.

We propose a generalizable framework, which can accommodate several distinct features (identified in the example) coming into the system from clinical, genomic and environmental domains among others. We utilize this framework to discover similar patients and overlaps among patients in a set of these features which is useful for several applications including:

- better disease management by studying peer group of patients with similar diagnosis but better prognosis through case-based reasoning,
- better cohort discovery for clinical trials,

- early disease diagnosis by studying similar features in patients with existing diagnosis

We propose an approach, which utilizes patient similarities in multiple heterogeneous datasets. Such a framework emulates the knowledge of an expert who looks at multiple heterogeneous case characteristics to come up with matching cases. There are several challenges that emerge in designing such a framework: (a) Merging features from multiple heterogeneous datasets, (b) Reducing the search space in the massive datasets to zero in/on the relevant patients, (c) Discovering similarities in highly sparse feature sets, (d) Computing similarities in heterogeneous feature sets with mixed distance measures, and (e) Developing novel clustering methods to identify groups of similar feature sets. We address these challenges by proposing a combined similarity function that looks at similarity across numeric and categorical values and employing this function in a clustering algorithm to find similar patients across multiple heterogeneous datasets.

The rest of the paper is organized as follows:
In Section II we discuss related work, in Section III we explain our approach, in Section IV we introduce our data sets and present results and in Section V we discuss conclusions and future work.

II. Related Work

A. *Similarity-based Clustering*

There have been several studies on clustering algorithms in general. Traditional clustering algorithms employ distance measures for determining similar data objects. While these techniques are suitable for numeric data, they do not work well for categorical data due to its discrete nature. There exist hierarchical clustering methods which can be applied to mixed attributes data [1]. However they are computationally intensive. In this section we describe some of the efforts taken towards clustering datasets of mixed attributes.

Huang et al. [2] introduced K-prototypes algorithm to address the problem of clustering large mixed datasets. In this approach, distinct similarity measures are used for different types of attributes. For numeric attributes squared Euclidean distance is used whereas for categorical attributes Hamming distance is used. These measures are then integrated and used for clustering the dataset. Weight is associated with categorical attributes which is based on average standard deviation on numeric attributes in a particular cluster. While this algorithm defines a combined similarity measure for mixed data sets, there are a few shortcomings. Firstly, the similarity measure for categorical attributes may not be the best measure for different kinds of data sets. Secondly, the weight parameter for numeric attributes is assumed to be one whereas for categorical attributes it depends on the distribution of numeric attributes. Thus this technique may not consider the influence of all the attributes which are significant in clustering.

He et al. [3] proposed an algorithmic framework called Cluster Ensemble Based Mixed Data Clustering (CEBMDC). In this framework, the original dataset is categorized into two data sets; one containing numeric attributes and another containing categorical attributes. For numeric dataset, existing algorithms like CURE [4] and CHAMELEON [5] are applied and for categorical dataset algorithms like ROCK [6] and Squeezer [7] are applied to obtain corresponding clusters. In the next step, individual clustering results are combined together as a categorical dataset and a suitable clustering algorithm is applied to obtain final results. The accuracy of the proposed algorithm was evaluated against the K-prototypes algorithm in terms of average clustering error by running the algorithms against two data sets. According to the experimental results, the CEBMDC algorithm had lower error that the K-prototypes. In this approach, the final results depend on the individual clustering techniques applied to different data sets and the integration of the intermediate results.

Reddy, Kavitha et. al. [8] proposed a clustering algorithm based on Similarity Weight and Collaborative filtering techniques. They use a methodology similar to that of He et al. where the original dataset is divided into pure numeric and categorical datasets. These datasets are then clustered using Similarity Weight method where a Similarity matrix is constructed using Jaccard Coefficient. In the last step, a filter method is applied to obtain final results. Ming-Yi Shih et al. [9] developed a Two-Step Method for Clustering Mixed Categorical and Numeric Data (TMCM). In this technique, the numeric attributes are normalized and the similarity between categorical attributes is determined based on their co-occurrence. In the next step, the categorical attributes are converted to numeric and existing algorithms are applied to the converted data set.

Ahmad et al. presented a clustering algorithm based on a new cost function and distance measure [10]. The proposed cost function consists of two parts out of which one is used for numeric data while other for categorical data. The numeric distance function is weighted squared Euclidean distance where the weight is determined from the input data itself. For categorical attributes, the proposed distance function takes into account the overall distribution of attributes and their co-occurrence with other attributes. The authors defined two evaluation metrics namely, micro-precision and micro-recall. The performance of the proposed algorithm was compared against various algorithms including K-prototypes, ROCK and traditional hierarchical clustering against different datasets.

As we can observe from the previous studies, many algorithms that have been developed employ a multi-step approach for mixed data clustering. Essentially different types of attributes are clustered separately and then combined to obtain final clusters. There are a few algorithms like K-prototypes that attempt to cluster heterogeneous, mixed data together. However, the similarity measures especially for categorical data may not truly represent the inherent nature of the datasets involved.

B. *Integration of clinical and genomic data*

Yorgos Goletsis [19] developed a framework for clinical decision support system (CDSS) based on profile extraction

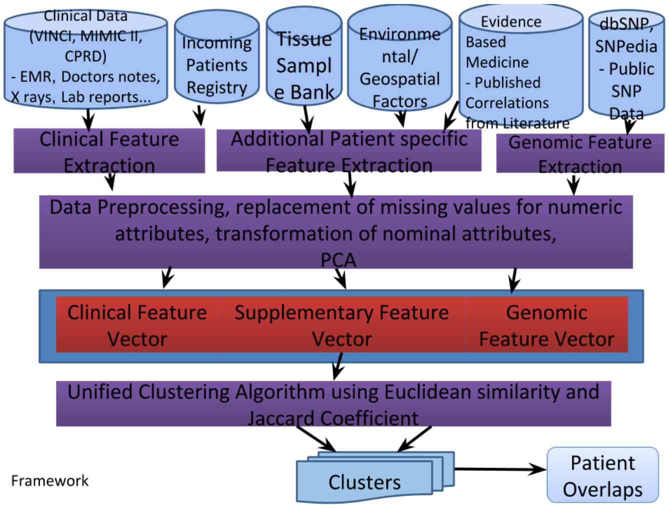
Clinical Data (VINCI, MIMIC II, CPRD) - EMR, Doctors notes, X rays, Lab reports...
Incoming Patients Registry
Tissue Sampl e Bank
Environme ntal/ Geospatial Factors
Evidence Based Medicine - Published Correlations from Literature
dbSNP, SNPedia - Public SNP Data
Clinical Feature Extraction
Additional Patient specific Feature Extraction
Genomic Feature Extraction
Data Preprocessing, replacement of missing values for numeric attributes, transformation of nominal attributes, PCA
Clinical Feature Vector
Supplementary Feature Vector
Genomic Feature Vector
Unified Clustering Algorithm using Euclidean similarity and Jaccard Coefficient
Clusters
Patient Overlaps
Framework

Fig. 1. System Architecture

by integrating clinical and genomic data for colon cancer. The clinical data included age, diet, obesity, diabetes information. The genomic information was represented in the form of SNP's related to colon cancer. Perez Rey D. [20] demonstrated the use ONTOFUSION system for integrating seven public web-based databases. Matthias Samwald [21], developed a semantic knowledge base for clinical pharmacogenetics using Web Ontology Language (OWL 2 DL), a Resource Description Framework (RDF) model and RDF conversion of relevant biomedical datasets. Emanuel Schwarz [22] used graph networks for integrating clinical and genomic data while studying psychiatric diseases.

III. Clinico-genomic Data Analytics Approach

In this section, we present our framework for performing clinico-genomic unified clustering. We propose a generalized methodology which can incorporate data coming from a variety of heterogeneous sources. This includes clinical, genomic, environmental and geospatial domains among others. We believe that this can help in discovery of better cohorts of patients where there is a clear demarcation based on factors like similar disease manifestations, demographic background, clinical parameters and so on. In addition to this, patients grouped together may differ across certain parameters which may be contributing to risk factors of the disease they are suffering from. Such cohorts can be selected for clinical trials. This can lead to better disease management by studying peer group of patients with similar diagnosis and better prognosis as well as early disease diagnosis by studying similar features in patients with existing diagnosis.

Thus we perform integration of heterogeneous clinical and genomic datasets as shown in Figure 1, transform them to homogeneous form, select relevant features and apply unsupervised machine learning techniques to obtain and study patient overlaps. We now explain in detail the steps performed for clustering patient records and studying overlaps across them.

A. *Data Extraction*

The overall approach as shown in Figure 1, looks at clinical data and performs clinical record extraction. The clinical data can consist of structured medical records, unstructured doctor's notes, X-rays, patient lab results among other data components. Data is extracted from relational and non-relational forms. The genomic data is in the form of SNP's and is extracted from large databases, prior clinical studies and patient specific genomic data. The extracted data can be noisy, can have missing or null values and can be inconsistent with respect to attribute labels. All these factors can hamper the efficiency of the results and data evaluation. Hence there is a need to clean the data to eliminate these discrepancies and make it suitable for performing the analysis.

B. *Data Pre-processing*

Data pre-processing consists of multiple steps including feature selection, data cleaning, data integration and transformation.

*1) **Feature selection**:* Feature selection is an important step in data mining and particularly in our approach as we have a vast amount of heterogeneous features. Feature selection deals with the selection of subset of features which increase the relevance of the features to the problem under study and help to reduce redundancy and noise in the data. There are different feature selection techniques depending on the learning problem which can be supervised or unsupervised or domain driven. We select pertinent features using domain knowledge from literature. We have used this approach to identify a set of features which are known to increase the susceptibility of diseases under study. Along-with this we also sought help of a domain expert in selecting additional features which may not have a strong disease association but can increase the risk factor. We believe that this might help us in identifying unknown overlaps which is the focus of study. Some of the features which we are interested in include Cholesterol and Triglyceride levels, Arterial Blood Pressure as well as Electrolyte levels of patients including Sodium and Potassium.

*2) **Data Cleaning**:* Once we have selected a set of features, the next step is to handle missing values in the data. In a given dataset, missing values can be represented as either Null or empty values. This can happen due to multiple reasons including lack of information, incorrect data measurement and it can impact the accuracy of data evaluation. We utilize mean or modal values from the class or group where the record with missing value belongs. So for example, for each patient record having empty values for certain attributes, we look at patient records having similar demographic information in terms of ethnicity and gender. Then we aggregate their values for the corresponding missing attribute and calculate the minimum value, maximum value, average value and standard deviation. We then replace the empty value with these four values. In case of Null values, we simply ignore them.

*3) **Data Integration and Transformation**:* In order to obtain clinico-genomic clusters, the clinical and genomic data of patients has to be integrated. For each patient record we identify the specific disease manifestation. From this we determine the risk inducing SNP's from large public databases as well from patient specific genomic data. We integrate the dbSNP [11] reference number of these SNP's with the clinical records of patients to obtain combined clinico-genomic patient data. This data being heterogenous in nature consists of different types of numeric attributes like Heart Rate and Blood Pressure as well as categorical attributes like Ethnicity and SNP reference ID's. Hence there is a need to transform the data in order to make it suitable for defining a combined similarity measure and to perform clustering.

In order to achieve this, we reconstruct the categorical features of patients as a binary feature vector. As a first step, we create a set of all possible values of all the categorical features present in the data. For each patient we create a binary feature vector initialized with same number of zeros as that of attribute values. We then analyze the patient records and set the binary value to one for corresponding value of each categorical feature present in the set created earlier and remaining values as zero. We thus transform the categorical feature vector into binary vector. Although Jaccard Coefficient can be applied directly to categorical features, we transformed them to binary for the sake of ease of calculation. At the end of this step, we obtain a universal feature vector by combining clinical features with binary feature vector for each patient.

C. Combined Similarity Function

An important element of a clustering algorithm is the underlying similarity measure which determines the semantics of the clusters. For data sets containing continuous variables, Euclidian distance is a commonly used distance metric. However the notion of similarity between categorical variables is difficult to understand. Moreover, it depends on the application domain where clustering is performed. A widely used approach is to first scale the categorical variables of data instances into binary variables and then apply a similarity measure for binary data. Some of the popular similarity measures include Simple Matching Coefficient, Jaccard Coefficient, Dice Coefficient, Rogers and Tanimoto [12]. The choice of the similarity measure depends on whether the number of matches is equivalent to the number of mismatches and whether zero-one or one-zero matches are to be considered significant while defining similarity. While Simple Matching Coefficient is useful in cases where zero-zero matches are as significant as one-zero matches, Jaccard Coefficient is suitable for calculating overlap between two data sets. As we are interested in identifying patient similarities and differences in terms of clinical and genomic overlaps, we decided to use Jaccard Coefficient as our similarity measure for categorical attributes. For continuous variables, we selected Euclidian distance and defined a distance function combining the two.

While the separate use of similarity measures for continuous and categorical variables is straightforward, a challenging task is to define a unified similarity for an integrated heterogeneous data set consisting of both these kinds of variables. This is due to the fact that the scale of continuous variables is different than that of the multivariate categorical variables. Hence we decided to transform the Eucledian distance into a similarity measure by adding one to it and taking its reciprocal. The reciprocal value is taken because eucledian distance measures the dissimilarity in the form of distance while Jaccard coefficient measures similarity, so to combine the measures we need to have them on the same level as either similarity or dissimilarity One is added to the distance to ensure that the resulting value lies between zero and one. Then we add this Euclidian-based similarity with the Jaccard Coefficient to obtain a combined similarity function. In this way we transform variables of different type and scale into homogenous form and then derive a combined similarity function which is suitable for clustering. Algorithm 2 demonstrates the calculation of the combined similarity measure.

D. Unified Clustering

In this section, we describe our algorithm which performs unified clustering on heterogeneous datasets.
As described in Algorithm 1, our methodology is an improvement over the simple K-means clustering to make it suitable for heterogenous data. The key differences lie in the use of a consolidated similarity measure by combining Euclidian-based similarity with Jaccard similarity. We also define a unique way to calculate cluster centroids which consist of mean/modal values depending on the type of data. The input to the algorithm is the number of data instances, a dataset consisting of numeric and categorical attributes, *K* which is the number of clusters to be formed and a list of initial centroids randomly sampled from data set. At the beginning, we perform initial assignment of data instances to the clusters based on similarity score between them and the cluster centroids as stated on line 3-4. For determining the similarity score, we use the GetSimilarity function which is described in Algorithm 2. After the initial assignment, we recalculate the centroid of all the clusters on line 8 using RecalculateCentroid procedure described in Algorithm 3. We then reassign the instances to the closest cluster based on similarity score on line 9. We repeat steps on lines 8-9 until there are no more reassignments of data instances across clusters.

Algorithm 2 describes a function to compute the similarity between two data instances. It takes as an input two data instances and returns similarity between them.

Algorithm 3 illustrates the process of calculating new centroids after the initial assignment of data instances to clusters. It takes as an input *clusters[i, j]* which stores the centroid for each cluster where i is the cluster number and j is the cluster centroid, *clusterMembers[i, j]* which is used to store the number of instances assigned to each cluster where i is the cluster number and j is an instance assigned to it. As indicated on lines 1-2 for each cluster C_i, a list of all instances assigned to it is extracted. Then we iterate over all the instances and collect numerical attributes and multivariate categorical

Algorithm 1 Unified Clustering algorithm

```
Inputs
  1) n: total number of instances.
  2) instancesList: list of instances $D_1$, $D_2$,...$D_n$ containing
     numeric and categorical attributes.
  3) K: number of clusters to be formed.
  4) centroidList: initial list of cluster centroids.
Require: $K >= 2$
for each instance $D_i$ in instancesList do
    for each centroid $C_i$ in centroidList do
        call GetSimilarity to obtain similarity score between
        $D_i$ and $C_i$.
    end for
    assign $D_i$ to the cluster such that the similarity score
    is maximum.
end for
repeat
    call RecalculateCentroid procedure
    reassign the data instances to new clusters based on
    their original and new similarity scores
until no movement of data instances across clusters =0
```

Algorithm 2 combined similarity

```
function GETSIMILARITY($D_i$, $D_j$)
catDist = similarity between categorical attributes of $D_i$,
    $D_j$ using Jaccard Coefficient
numericDist = distance between numerical attributes of
    $D_i$, $D_j$ using Euclidean distance
numericDist = $\frac{1}{1+numericDist}$
similarityScore = catDist + numericDist
return similarityScore
end function
```

Algorithm 3 Recalculate Centroid of clusters

```
Inputs
  1) $clusters[i, j]$: stores the centroid for each cluster.
  2) $clusterMembers[i, j]$: stores the instances belonging to
     clusters.
 1: for each cluster $C_i$ in clusterMembers do
 2:     get list of instances assigned to $C_i$ and store them in
        clusterInstances
 3:     for each inst in clusterInstances do
 4:         aggregate numeric attributes of inst in numericAttr
 5:         aggregate categorical attributes of inst in categorica-
            lAttr
 6:     end for
 7:     avgNumericAttr = average of numerical attributes in
        numericAttr.
 8:     modeCategoricalAttr = mode of categorical attributes in
        categoricalAttr.
 9:     $newCentroid$ = concatenate avgNumericAttr and mod-
        eCategoricalAttr.
10:     clusters[i] = $newCentroid$
11: end for=0
```

attributes separately as stated on line 4 and 5. On line 7, we calculate average of numerical attributes across all the instances. We also compute mode of multivariate categorical attributes across all the instances as stated on line 8. We then concatenate the average and the mode to obtain the new centroid of cluster *i*. We update *clusters[i]* to store this new centroid. The process is repeated for all the clusters.

E. Cluster Evaluation Metrics

After running the unified clustering algorithm across patient records to obtain clinico-genomic clusters, the next step is to assess the quality of clusters. We primarily use three different measures of cluster evaluation as described in the following sections.

*1) **Sum of Squared Errors(SSE):*** This is a common measure of cluster quality which calculates the error of a data point in a given cluster in terms of its distance from the cluster centroid. To obtain *SSE*, the errors of all the data points are squared and then added.

Given a set of *n* data instances D = $d_1, d_2, ...d_n$ where each data point has *m* dimensions. The data set is divided into *K* clusters such that C_i is the centroid of i^{th} cluster. Then the sum of squared error between each data instance and the cluster centroid as follows

$$\sum_{j=0}^{m}[(d_j - C_{ij})^2]$$

Euclidean distance is commonly used to calculate the distance of a data point from its cluster centroid. In our algorithm, we calculate sum of squared errors using the combined similarity function. For each cluster, we calculate the similarity of a data instance with its centroid using the GetSimilarity function. We subtract this similarity from one to obtain the error in terms of dissimilarity. Then we sum the squared dissimilarity of all the data instances and take average to obtain *SSE* for a particular cluster.

We have utilized the *SSE* values to select the ideal number of clusters for K-means and our clustering algorithm. For this we calculate *SSE* values by varying the number of clusters denoted by *K*. We choose the K value with the *SSE* value using the elbow method [13].

*2) **Silhouette Coefficient:*** Silhouette Coefficient is a cluster evaluation measure which is useful to check if the clusters are well-formed and robust [14]. Silhouette Coefficient is calculated as follows:

$$s(i) = \frac{b(i)-a(i)}{max\{b(i),a(i)\}}$$

where,
i: data instance
s(i): Silhouette Coefficient of instance *i*
a(i): average dissimilarity of *i* with all the data points belonging to same cluster.
b(i): lowest average dissimilarity of *i* with any cluster other than the one it belongs to.

For each data point, we first calculate the average similarity with all other data points in the same cluster using the combined similarity function. To obtain average dissimilarity we subtract it from one. We also calculate the lowest average similarity of each data point with data points of other clusters. We subtract it from one to obtain the lowest average dissimilarity. Then using the above formula we calculate Silhouette Coefficient of each data instance. We calculate the Silhouette Coefficient of a cluster by taking average of the coefficient values of all the data instances in the cluster. For the number of clusters *K* as selected in earlier step, we calculate Silhouette Coefficient of each cluster along with *SSE* values. We choose the clusters having lower *SSE* and higher Silhouette Coefficient for studying patient overlaps.

*3) **Patient Overlaps**:* For each of the clusters identified using evaluation measures described above, we generate a similarity matrix consisting of pairwise similarity of all the patient records. We then select the set of patient records having high similarity amongst themselves to study their clinical and genomic features and identify similarities and differences in terms of overlap. The partial overlaps, especially overlaps above a certain threshold, for example 60 to 70 percent, may be interesting. This is because these patients indicate a high level of similarity in majority of attributes but do not overlap in some attributes, which may be of interest to study especially if the overlap is in key clinical variables but not in genomic variables and vice versa.

IV. Experiments and Results

In this section we describe the datasets used for performing experiments. We also present our results and lay a path for future work.

A. *Data Sets*

Here we describe the clinical and genomic datasets which we have utilized for our analysis.

*1) **Clinical Dataset**:* We use MIMIC II [15] (Multiparameter Intelligent Monitoring in Intensive Care) clinical database for clinical records of patients. It consists of clinical information of ICU patients admitted in Beth Israel Deaconess Medical Center(BIDMC) in Boston. The data was collected over a period of seven years starting in 2001. It is available for access under data user agreement (DUA) and it is stored in relational database. The database consists of records of patients suffering from various diseases. We are currently focusing on Diabetes and Cardiovascular diseases but our framework can be extended to include other diseases as well. Our dataset consists of clinical records of 542 patients. Out of these, there are 260 Diabetic patients, 179 patients suffering from cardiovascular diseases whereas 103 patients have manifestations of both Diabetes and Cardiovascular diseases. Table 2 lists a set of features which we have selected for our study. The Arterial blood pressure represents the mean blood pressure. We have selected patient ID and disease description in order to assist in evaluating the results. While performing clustering we remove them from the feature vector.

TABLE I
LIST OF CLINICAL FEATURES

Number	Feature name
1	Cholesterol level(mg/dl)
2	Glucose level(mg/dl)
3	Triglyceride level(mg/dl)
4	Hemoglobin(gm/dl)
5	Gender
6	Arterial Blood Pressure (mm Hg)
7	Daily Weight(kg)
8	Heart Rate(beats per minute)
9	Potassium(mg)
10	Ethnicity
11	Sodium(mEq/l)
12	Calcium(mg/dl)
13	Marital Status
14	Disease description
15	patient ID

*2) **Genomic Databases**:* In order to gain insight into risk inducing *SNP's*, we performed a literature survey of genomic databases which are publicly known. Few examples include *dbSNP* and *SNPedia*. *dbSNP* stands for Short Genetic Variations database [11]. It maintains information about single nucleotide polymorphisms which may have disease associations. Users can submit information about *SNP's* to the *dbSNP* database. *dbSNP* assigns a reference *SNP ID* to all the *SNP's*. Users can search for genetic variations by querying the database online. The results consist of detailed report of genetic records including reference *SNP ID*, summary of the allele, the coordinates of the chromosome and so on.

SNPedia is like a Wikipedia of genetic information. It consists of information about 64945 *SNP's* [16]. It has a search facility where we can search for a specific *SNP* or *SNP's* associated with a particular disease. It provides detailed information about the mutations of the nucleotides, the associated risk factor, chromosome position, Gene, *dbSNP* reference number etc. It also lists a set of publications and peer-reviewed papers which have identified the mutations.

We use *SNPedia* to obtain the reference ID's of risk causing *SNP's* for diseases under study. For experimental purpose, we consider top 3 risk inducing *SNP's* for each of the diseases.

B. *Clustering and Results*

In this section we present a comparative study of the proposed unified clustering algorithm with the K-means algorithm. We have used *Weka 3.6.10* for executing Simple K-means clustering with seed value of 500 and with Eucledian distance as the distance metric. Our purpose is to determine the

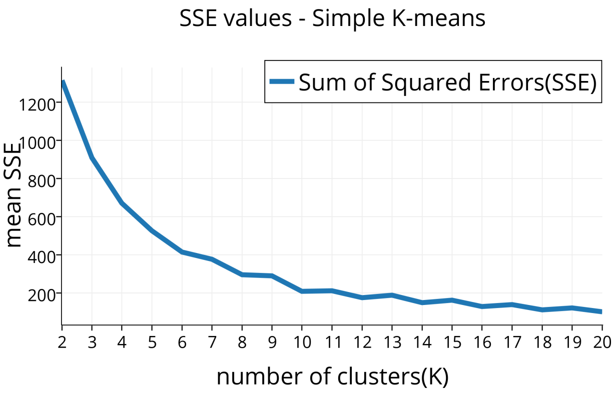

Fig. 2. A graph showing SSE values for Simple K-means

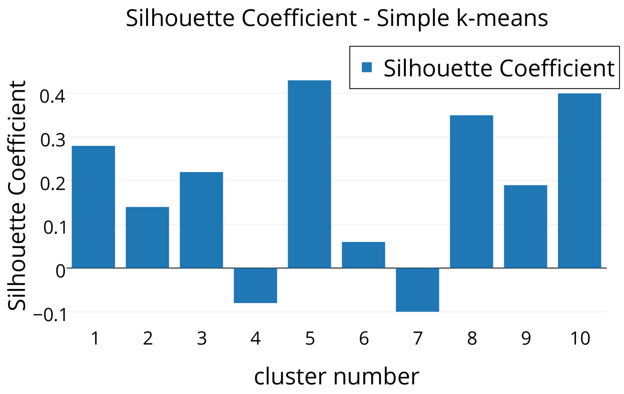

Fig. 4. Silhouette Coefficient values for K-means clustering

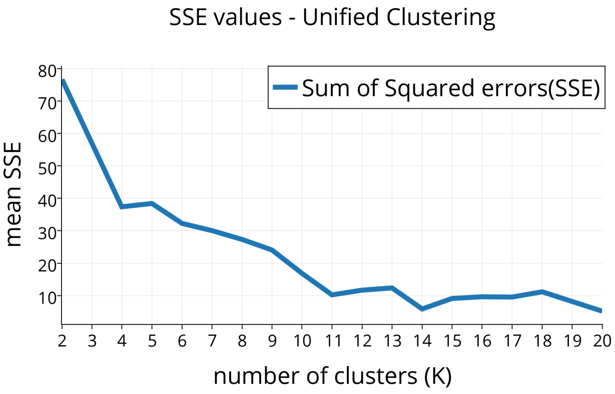

Fig. 3. A graph showing SSE values for unified clustering

effectiveness of a combined similarity function in a clustering algorithm as opposed to a single metric which Weka uses for identifying patient overlaps. We evaluate the algorithms primarily using the validation metrics explained in the previous section.

*1) **Validation using Sum of Squared Errors (SSE):*** As a first step in cluster evaluation, we use SSE to compare and analyze the two clustering algorithms. Figure 3 shows a graph of number of clusters denoted by K versus the average SSE for K-means clustering. As seen from the graph, initially the SSE values decrease significantly. For 10 clusters, the SSE is 209. As the number of clusters increase, the SSE values do not change significantly. Figure 4 shows a graph of average SSE values against number of clusters for unified clustering. For $K = 2$ *to* 10 we observe that SSE decrease rapidly. The SSE for 10 clusters was reported as 16. Using the elbow criterion with SSE as a measure for estimating the number of clusters, we decided to select ideal-K as 10 for both the algorithms.

*2) **Validation using Silhouette Coefficient:*** Figure 5 displays a plot of Silhouette index values for 10 clusters for Simple K-means. In addition to these, Table I contains SSE values for individual clusters. As we can observe, cluster 5 has the highest Silhouette value of 0.43 followed by cluster 10 whose value is 0.40. Cluster 4 and cluster 7 have lower values of -0.08 and -0.16 respectively. The average value of Silhouette over the entire data set is 0.20.

TABLE II
SSE AND SILHOUETTE COEFFICIENT VALUES FOR SIMPLE K-MEANS

cluster number	SSE	Silhouette Coefficient
1	399.73	0.28
2	149.83	0.14
3	324.00	0.22
4	250.35	-0.08
5	105.56	0.43
6	117.56	0.06
7	263.17	-0.16
8	90.51	0.35
9	238.64	0.19
10	151.23	0.40

The Silhouette values obtained for 10 clusters for the our approach with the combined similarity function is shown in Figure 6. The corresponding SSE values are shown in table II. Cluster 9 reported the highest Silhouette value of 0.61 followed by cluster 3 with 0.56 and cluster 10 with 0.55. The average Silhouette index for entire data set was reported as 0.46.

As we can observe the average Silhouette value of the entire data set for our unified clustering approach is greater than that of Simple K-means indicating a better cluster quality.

*3) **Study of patient overlaps:*** Once we have identified well formed clusters, the next step is to determine a group of highly similar patients and study their clinical and genomic features. For this purpose, we generate a similarity matrix by calculating pairwise similarity across all the patients in the desired cluster. We then select the patient records which have a high similarity value. Following are our observations:

- **Cluster 9:**
  This cluster comprised of 65 patients. These patients had very similar demographic background and all of

TABLE III
SSE AND SILHOUETTE COEFFICIENT VALUES FOR UNIFIED CLUSTERING

cluster number	SSE	Silhouette Coefficient
1	9.48	0.4
2	16.41	0.49
3	3.73	0.56
4	3.33	0.38
5	5.58	0.47
6	1.74	0.28
7	80.21	0.48
8	11.00	0.09
9	33.24	0.61
10	4.20	0.55

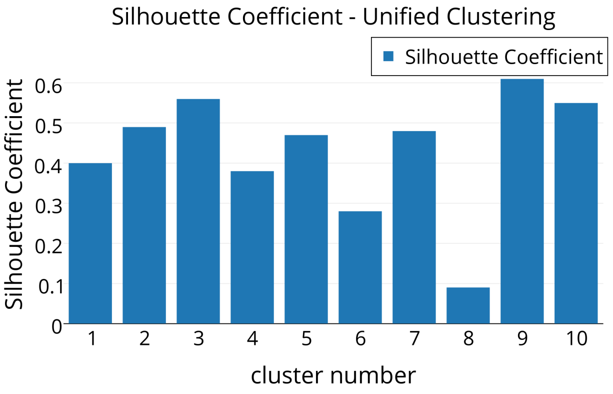

Fig. 5. Silhouette Coefficient values for Unified clustering

them demonstrated type II Diabetes manifestations. The patients were White males. 40% of the patients were widowed, 27% were married, 13% were single, 3% were divorced whereas for the rest Marital status was unknown. The glucose levels of the patients were varying with a few patients having higher levels. These patients had high Arterial Blood Pressure however, their Glucose and Triglyceride levels were in the normal range. The electrolyte levels of the patients were similar and normal. As we can observe patients in this cluster had high genomic overlap as they suffered from similar diseases whereas they had differences across certain clinical features. One of the unique features was that primary factors like Glucose and Triglyceride levels known for being a predominant risk factors for Diabetes were normal.

- **Cluster 3:**
  There were 105 patients present in this cluster. Around 70% of patients in this group were suffering from Chronic Heart disease and its manifestations. Two patients were diagnosed with Diabetes. They had high clinical overlap in terms of high Glucose levels, normal Cholesterol and Triglyceride levels and normal Electrolyte levels. An unusual observation was that these patients had normal Heart Rate.
- **Cluster 10:**
  This group consisted of 21 patients. The patients in this cluster belonged to different Ethnic groups. Roughly 43% of the patients were White, for 23% of patients ethnicity was not known or unspecified. 19% of patients were African and there was 1 Asian and 1 Hispanic/ latino patient. There was 1 male patient whereas rest were female. The patients also differed across their marital status with 14 patients were Widowed, 4 were Divorced, 1 was married whereas for 2 patients it was missing. On the other hand, these patients had similar values across all the clinical features. An interesting observation was that some patients were diagnosed with Type II Diabetes with Family history of Chronic Heart disease whereas few other only had a Family history of these diseases. Such a group of people having different Ethnic background and disease manifestations can be a good candidate for cohort study. In this case people having similar clinical observations may be diagnosed earlier and prescribed better medication.

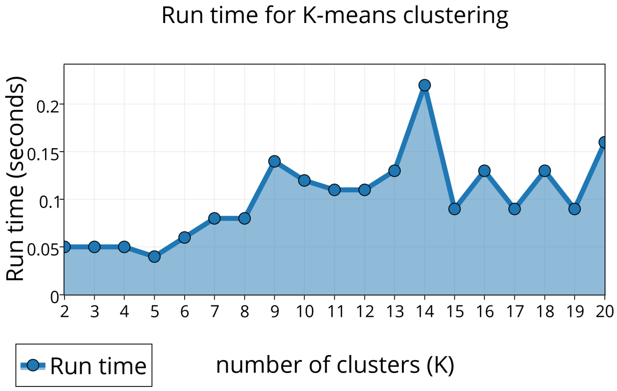

Fig. 6. Run Time for Simple K-means

*4) **Comparison of Algorithmic Run time:*** Figure 7 demonstrates a plot of run time of Simple K-means against varying number of clusters. As we can observe, the time taken to execute the algorithm increases gradually from 2 to 13 clusters. For 14 clusters, the algorithm spends maximum time in execution which is 0.22 seconds.

Figure 8 shows the run time of the proposed algorithm. As the number of clusters increases the run time also increases. For 20 clusters, it takes around 33 seconds for the algorithm to complete execution. Thus the running time for our algorithm is higher than that compared to Simple K-means. We have utilized Python for the algorithm implementation and we suspect that might have added to the running time.

V. CONCLUSION AND FUTURE WORK

We provide a generalized framework to integrate clinical data of patients with genetic data from public databases to study overlaps between them. Our proposed similarity measure

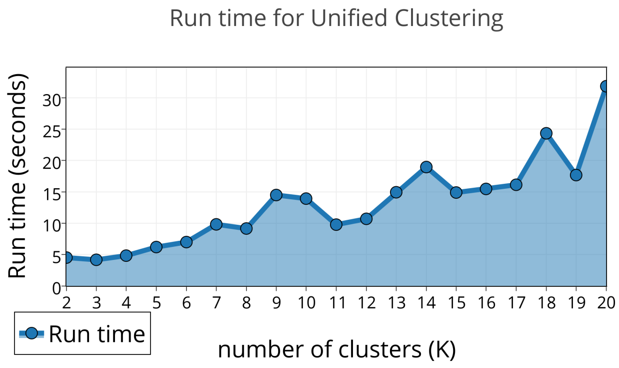

Fig. 7. Run time for Unified clustering

takes into account the heterogeneity in the data to obtain well formed clusters. As we can observe from the Silhouette Coefficient calculation, the clusters obtained by our algorithm are better in quality and are meaningful. Going forward it would be interesting to study highly specific genetic records of patients along with clinical records for integration which can also lead to discovery of potentially less known risk inducing SNP's. Our study can be extended to utilize other similarity measures like Dice Coefficient to test their usefulness in clustering. We are currently using MIMIC II database for obtaining clinical data of patients. In future, we will extend our scope to include patient data from other databases like VA Informatics and Computing Infrastructure (VINCI), Clinical Practice Research Datalink(CPRD) and so on. The combined similarity function may not follow triangle inequality theorem which may have issues with algorithm convergence. This can be investigated in future to ensure that our algorithm always converges. While studying the overlaps we looked at the mean/modal values of each patient record for direct comparison. As an ongoing work, we can also look at the min/max values to study variations in clinical features of patients in a cluster.

References

[1] J. C. Gower, *A General Coefficient of Similarity and Some of Its Properties*, BioMetrics, 27, pp.857-874 (1971)

[2] Huang Z, *Clustering large data sets with mixed numeric and categorical values*, Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore, pp. 21-34, 1997

[3] Zengyou He, Xiaofe i Xu, Shengchun Deng, *Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach*, arXiv:cs/0509011, 2005.

[4] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, *CURE: An Efficient Clustering Algorithm for Large Data sets*, Published in the Proceedings of the ACM SIGMOD Conference, 1998.

[5] George Karypis, Eui-Hong Han, Vipin Kumar, *CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling*, IEEE Computer Volume 32 Issue 8, 68-75, 1999.

[6] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, *ROCK: A Robust Clustering Algorithm for Categorical Attributes*, In Proceedings of the 15th International Conference on DataEngineering, 2000.

[7] Zengyou He, Xiaofe i Xu, Shengchun Deng, *Squeezer: An Efficient Algorithm for Clustering Categorical Data*, Journal of Computer Science and Technology, Volume 17 Issue 5, May 2002, Pages 611-624.

[8] M. V. Jagannatha Reddy, B. Kavitha, *Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method*, International Journal of Database Theory and Application Vol. 5, No. 1, March, 2012.

[9] Ming-Yi Shih, Jar-Wen Jheng, Lien-Fu Lai, *A Two-Step Method for Clustering Mixed Categroical and Numeric Data*, Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 11-19 2010.

[10] Amir Ahmad, Lipika Dey, *A k-mean clustering algorithm for mixed numeric and categorical data*, Data and Knowledge Engineering Journal, Volume 63 Issue 2, November, 2007 Pages 503-527.

[11] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K, *dbSNP: the NCBI database of genetic variation*, Nucleic Acids Res. 2001 Jan 1;29(1):308-11.

[12] Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl, *Cluster Analysis 5th Edition*, WILEY SERIES IN PROBABILITY AND STATISTICS.

[13] *The Elbow Method*, en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set.

[14] *Silhouette (clustering)*, http://en.wikipedia.org/wiki/Silhouette_clustering.

[15] Saeed, Mohammed and Villarroel, Mauricio and Reisner, Andrew T. and Clifford, Gari and Lehman, Li-Wei and Moody, George and Heldt, Thomas and Kyaw, Tin H. and Moody, Benjamin and Mark, Roger G, *Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database*, Critical Care Medicine, Volume 39, Pages 952-960, May 2011.

[16] Michael Cariaso, Greg Lennon, *SNPedia: a wiki supporting personal genome annotation, interpretation and analysis* Nucleic Acids Research 2011; doi: 10.1093/nar/gkr798.

[17] Schwarz E "Clinical bioinformatics for complex disorders: a schizophrenia case study." BMC bioinformatics 10.Suppl 12 (2009): S6.

[18]

[19] Goletsis Y, *Integration of clinical and genomic data for decision support in cancer.* Encyclopedia of Healthcare Information Systems, Idea Group Publishing, USA,(to be published) (2007).

[20] Prez-Rey D, *ONTOFUSION: Ontology-based integration of genomic and clinical databases.* Computers in biology and medicine 36.7 (2006): 712-730.

[21] Samwald M, *An RDF/OWL knowledge base for query answering and decision support in clinical pharmacogenetics.* Studies in health technology and informatics 192 (2013): 539.

[22] Schwarz E "Clinical bioinformatics for complex disorders: a schizophrenia case study." BMC bioinformatics 10.Suppl 12 (2009): S6.