2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Comparison of Distance Metrics for Hierarchical Data in Medical Databases

Diman Hassan, Uwe Aickelin and Christian Wagner

***Abstract*—Distance metrics are broadly used in different research areas and applications, such as bio-informatics, data mining and many other fields. However, there are some metrics, like *pq*-gram and Edit Distance used specifically for data with a hierarchical structure. Other metrics used for non-hierarchical data are the geometric and Hamming metrics. We have applied these metrics to The Health Improvement Network (THIN) database which has some hierarchical data. The THIN data has to be converted into a tree-like structure for the first group of metrics. For the second group of metrics, the data are converted into a frequency table or matrix, then for all metrics, all distances are found and normalised. Based on this particular data set, our research question: which of these metrics is useful for THIN data?. This paper compares the metrics, particularly the *pq*-gram metric on finding the similarities of patients' data. It also investigates the similar patients who have the same close distances as well as the metrics suitability for clustering the whole patient population. Our results show that the two groups of metrics perform differently as they represent different structures of the data. Nevertheless, all the metrics could represent some similar data of patients as well as discriminate sufficiently well in clustering the patient population using *k*-means clustering algorithm.**

I. INTRODUCTION

SINCE the representation of structured objects in large and modern databases like The Health Improvement Network (THIN) database becomes more complex and important, such structures should be considered when searching for similar objects. Therefore, finding an efficient measurement for discovering similar objects in data sets is the key feature when the task is to classify new objects or to cluster data objects. The *pq*-gram [1] and Edit Distance [2] metrics are known to be two good approaches that have been used to measure the similarity of the structured data objects, especially in Trees. The limitation of Edit Distance metric is related to the computational complexity which is considered very high [3] as compared to the *pq*-gram distance metric.

On the other hand, there are other metrics that are simple and implemented on non-structured data, such as Euclidean, Minkowski, Manhattan and Hamming Distance metrics [4]. Some of these metrics have been compared to other measures to find their efficiency. In [5], a comparison has been made between the geometric metrics and actual measures to estimate the distance in spatial analytical models. The results gave accurate distances for the actual distances than to the geometric metrics. Recently, using THIN database (www.thin-uk.com) which is belong to the general practice electronic healthcare database, some research [6] [7] have been performed using data mining techniques, such as association and sequential patterns. The purpose was to detect association between patient attributes (e.g. age, gender, medical history) and adverse events of drugs. No other data mining technique has been applied to the THIN database yet, such as clustering; this motivated us to use the unexplored clustering approach for the prediction and detection of negative side effects of drugs. The overarching aim of our research is to cluster hierarchical data to identify adverse side effects of drugs in the THIN database. However, clustering techniques need distance measures to represent the similarity between patients who have similar side effects. For this reason, this preliminary work aims to find the useful and suitable measure for our hierarchical data set in order to cluster patients. To achieve this aim, different metrics are considered and applied to the THIN data and their results compared. The investigation determines if these metrics can measure similarity and find similar patients (i.e. the patients who have similar side effects of drugs). Additionally, by looking at the whole patient population, is any of the metrics able to accurately represent similarity between patients when using, for example the *k*-means clustering algorithms [8]?.

The layout of this paper is as follows. In Section II, a background on the THIN database and the distance metrics is given. The data preparation for both groups of metrics, the calculation of the distances and the clustering using those metrics are explained in Section III followed by a discussion on the results in Section IV. Section V presents a summary and the conclusion of the work.

II. MATERIALS AND METHODS

A. Background on THIN Database

The THIN database is one of the electronic health-care longitudinal databases that contains anonymous electronic medical records extracted directly from general practices throughout the United Kingdom. The database contains information of each patient registered within the general practice including personal details, such as gender, date of birth, date of registration and family history. In addition, the data on all the drug prescriptions and the associated set of symptoms based on which the drug is prescribed are also included. The individual medical record is represented in the THIN database by a reference code named as read code. The latter is an alphanumeric code that defines and groups illnesses using the hierarchical nosology system. The read codes are also comprehensive coded medical language developed in the UK

Diman Hassan, Uwe Aickelin and Christian Wagner are with the School of Computer Science, University of Nottingham, Nottingham, United Kingdom (email: {dsh, uxa, cxw}@cs.nott.ac.uk).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

and funded by the National Health Service (NHS). In this paper, we test our experiments on a group of patients between the age of 0 and 17 years old. The information shown in Table I was extracted from THIN for two kinds of drugs that have been chosen based on the number of prescriptions. The first drug DESLORATADINE has a large number of prescriptions and is used to treat allergies under the group of Antihistamines. The second drug has a smaller number of prescriptions and belongs to the family of Tricyclics that relate to antidepressant drugs [9]. For our experiments, a sample size of 9949 prescriptions after 30 days of taking the drug (representing 988 patients) out of 53,995 prescriptions (representing 18,293 patients) have been tested to find the similarity between them for the first drug. For the second drug we used all the prescriptions (1172) after 30 days of taking the drug for 42 patients.

TABLE I
A SUBSET OF INFORMATION FROM THE DATABASE FOR TWO KINDS OF DRUGS

	DESLORATADINE	DOXEPIN
All drug's codes in THIN data set	6	15
All prescription	358,768	72448
All patients	81,000	6152
All prescription(0-17 years)	53,995	2014
All patients (0-17 years)	18,293	60
All presc.(0-17) after 30 days	9949	1172
All Patients(0-17) after 30 days	988	42

B. Background on Distance Metrics

A metric space (X, d) is a set X that has the concept of distance $d(x, y)$ between any pair of points $x, y \in X$ and the metric is a function on the set X that satisfies the following properties for a distance [10] [11].
Definition: a metric d on a set X is a function d: $X \times X \rightarrow \mathbb{R}$ such that for all $x, y \in X$:
$d(\text{x, y}) \geq 0 \forall$ x, y $\in X$. (Non-negativity).
$d(\text{x, y}) = 0 \iff \text{x} = \text{y}$ (Identity).
$d(\text{x, y}) = d(\text{y, x}) \forall$ x, y $\in X$. (Symmetry).
$d(\text{x, y}) \leq d(\text{x, z}) + d(\text{z, y})$. (Triangle inequality) $\forall$ x, y *and* z $\in X$.
The following are the six distance metrics used in this study:

1) Euclidean Distance Metric: Euclidean metric is a distance d on the space $\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ which is defined as a distance between any two points in space $\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^n} \quad (1)$$

where $x = (x_1, x_2, ... ,x_n)$, $y = (y_1, y_2, ... ,y_n)$ [12].

2) Minkowski Distance Metric: Minkowski metric is a p-metric between n-dimensional points $x = (x_i)$ and $y = (y_i)$ defined as:

$$d(x, y) = \sqrt[p]{\sum_{i=1}^{n} |(x_i - y_i)^p|} \quad (2)$$

If $p = 2$, it is called Euclidean distance and if $p = 1$ it is called Manhattan or city block distance. If $p = \infty$, then it is called Chebyshev or maximum distance [4]. In our experiment, $p = 3$ has been used.

3) Manhattan Distance Metric: It is a special case of the Minkowski metric when $p = 1$ [4]:

$$d(x, y) = \sum_{i=1}^{n} |(x_i - y_i)| \quad (3)$$

where $x = (x_1, x_2,..., x_n)$ and $y = (y_1, y_2,..., y_n)$

4) Hamming Distance Metric: Hamming distance is used for the detection and correction of errors in digital communications. It is defined as the number of different symbols between two equal length sequences. For example, the hamming distance between "toned" and "roses" is 3 and between 217389 and 213379 is 2 [13].

5) Edit Distance Metric: According to Kialing et al. [2], the definition of the Edit Distance measure between two trees T_1 and T_2 is the minimum cost of all edit sequences that transform T_1 to T_2: Edit Distance$(T_1, T_2) = \min\{c(S) \backslash S$ a sequence of edit operations transformations T_1 into $T_2\}$. Kialing et al. claimed the advantage of using the edit distance as a similarity measure provided the mapping between the nodes in two trees during the term of edit sequence (Insertion, Deletion and Relabeling nodes in a tree T).

6) PQ-Gram Distance Metric: The pq-gram distance has been proposed by Augsten et al. [1] and is mainly used for computing distances between ordered labeled trees. The pq-grams of a tree are all its sub-trees of a specific shape. The specific shape of the pq-gram is based on the values of two parameters p and q. The tree T shown in Fig. 1 is expanded by inserting dummy nodes (*) to make sure that each node appears at least in one pq-gram. The expansion of each tree is done by inserting p-1 before the root node, insert q-1 before the first and after the last child of each non-leaf node and insert q nodes to each leaf node, for example $p = 2$, $q = 3$ in Fig. 2. After the expansion process, the 2, 3-grams are extracted to produce the list of pq-grams. An example of a single 2, 3-gram is given in Fig. 2 where p = (*, a6706022p) is the stem and q = (*, *, 1) is the base. The trees that have a large number of common pq-grams are considered more similar than those trees that have less; Furthermore, the pq-gram distance is used to approximately match hierarchical data of large sources using the following equations:

$$dist^{(p,q)}(T_1, T_2) = |I_1 \uplus I_2| - 2|I_1 \Cap I_2| \quad (4)$$

Where T_1, T_2 are the two trees, and p and q are the two parameters that specify the shape of the pq-gram. The pq-gram indexes, I_1 and I_2 are the bags of Label-tuples of all pq-grams of T_1 and T_2, respectively. In addition, the $\uplus$ refers to the bag union between I_1 and I_2 and the $\Cap$ refers to the bag intersection between the same indexes. The normalisation of the pq-gram distances is as follows:

$$dist\_norm^{(p,q)}(T_1, T_2) = \frac{dist^{(p,q)}(T_1, T_2)}{|I_1 \uplus I_2| - |I_1 \Cap I_2|} \tag{5}$$

The pq-gram metric has been proposed originally to approximately match similar hierarchical information from autonomous sources that may have different representation in the sources [1]. The pq-gram metric has the advantage of computational efficiency and can be computed in $O(n \log n)$ time and $O(n)$ space. Another advantage of the pq-gram distance is that it can be tuned by adjusting the two parameters p and q [14]. The determination of p and q values depends on the underlying semantics of the data. In general, increasing the values of p and q makes the distance between two trees more sensitive to the structure of the trees rather than to the data, while decreasing them makes the distance sensitive to the data. As an example, in our experiments we have used different values of p and q: for $p = 1$ and $q = 3$ and for $p = 2$, $q = 3$, the results of pq-gram distances are shown in Table III and Table V. The results reveal that better distances are obtained when $p = 1$ and $q = 3$.

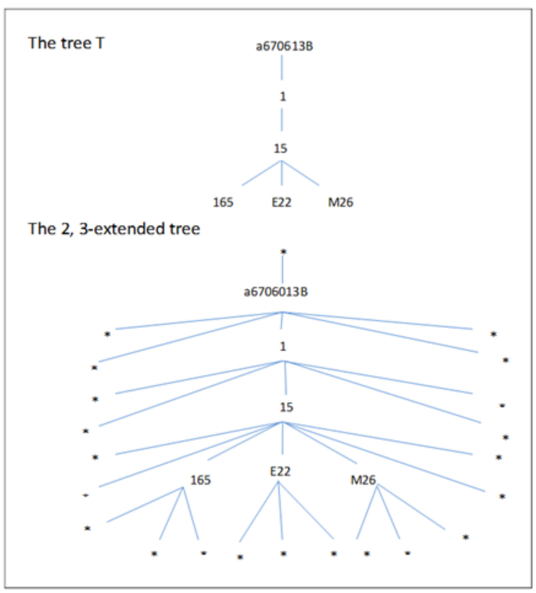


Fig. 1. An example of a tree T and its 2, 3-Extended tree

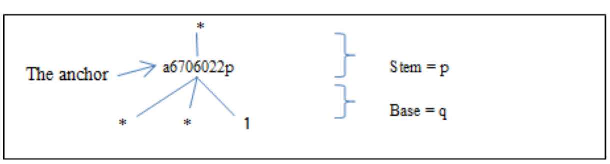


Fig. 2. An example of single pq-gram from a THIN data tree

III. EXPERIMENTS AND RESULTS

A. Data Preprocessing

The THIN data is converted into trees before applying the pq-gram and Edit Distance metrics, while the data for the geometric and Hamming distance metrics is converted into a frequency table. The data extracted from THIN is based on different patient's attributes such as the patient's unique ID, the gender, the age of first taking the specified drug and the medical codes related to the drug. The medical events are chosen at level 3 (the first three digits of the read codes like H33). Fig. 3 shows part of this information represented in THIN for three patients which have unique identifiers in the database (a6706013B, a6706015R, a670601o8):

Combid	Sex	Age	Medcode	Description
a6706013B	1	15	165	Feels hot/feverish
a6706013B	1	15	E22	Erectile dysfunction
a6706013B	1	15	M26	Acne vulgaris
a6706015R	2	10	168	Tired all the time
a6706015R	2	10	195	Gastric reflux
a6706015R	2	10	197	Epigastric pain
a6706015R	2	10	1B1	C/O - a headache
a6706015R	2	10	1BK	Worried
a6706015R	2	10	1C2	Tinnitus symptoms
a6706015R	2	10	730	Syringe ear to remove wax
a6706015R	2	10	C38	Localised adiposity - fat pad
a6706015R	2	10	F50	other otitis externa
a6706015R	2	10	F51	Eustachian tube dysfunction
a6706015R	2	10	F58	Ear pain
a6706015R	2	10	F58	Otalgia
a670601o8	1	12	171	Cough
a670601o8	1	12	19C	Constipation
a670601o8	1	12	1A5	Pain in penis

Fig. 3. Part of the THIN data extracted based on specific attributes

1) PQ-Gram and Edit Distance Preparation: From the data in Fig. 3, we have converted each patient's records into a tree as depicted in Fig. 4 to enable the computation of both pq-gram and Edit Distance metrics. For the pq-gram metric each tree is expanded in the same way as in Fig. 1. For our experiments, we use ($p = 1$, $q = 3$) and ($p = 2$, $q = 3$). After the process of tree expansion, the pq-grams are extracted for each tree; Fig. 5 shows the 2, 3-grams for the tree in Fig. 4. The pq-gram distance between two trees is formed by all the common pq-grams between them and computed using equation (4), while the calculation of the distances for the Edit Distance is performed by inserting, deleting or re-labeling nodes to convert one tree to another. The single edit operation has cost 1 and the Edit Distance between two trees is equal to the minimum cost or minimum number of edit operations to convert one tree to another.

```
A single tree
-------------------
a6706013B
  ---->1
      ---->15
          ---->165
          ---->E22
          ---->M26
```

Fig. 4. A tree representation from THIN data

2) Geometric and Hamming Metrics preparation: The THIN data for the Euclidean, Minkowski, Manhattan and Hamming metrics has been converted into a frequency table as shown in Table II. The table represents how many times each patient had a specific symptom after taking the specified drug. The table also contains additional columns,

TABLE II

THE FREQUENCY TABLE FROM THIN DATA

Patient's ID	The medical events															Patient's ages				Sex
	168	171	195	19C	1A5	730	F58	H17	M0.	M26	N24	N32	SD.	SL.	ZL5	10	11	12	15	
a6706013B	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1
a6706015R	1	0	1	0	0	1	2	0	0	0	0	0	0	0	0	1	0	0	0	2
a670601o8	0	1	0	1	1	0	0	1	1	1	1	1	1	1	0	0	0	1	0	1
a670601yJ	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	2

a6706013B
====================
(*, a6706013B,*,*, 1)
(*, a6706013B,*, 1,*)
(*, a6706013B, 1,*,*)
(1, 15,*,*, 165)
(1, 15,*, 165, E22)
(1, 15, 165, E22, M26)
(1, 15, E22, M26, *)
(1, 15, M26, *, *)
(15, 165, *, *, *)
(15, E22, *, *, *)
(15, M26,*,*,*)
(a6706013B, 1,*,*, 15)
(a6706013B, 1,*, 15,*)
(a6706013B, 1, 15,*,*)

Fig. 5. The 2, 3-grams of a tree T

one for the patient's gender (In THIN, 1 = male, 2 = female) and others for the different ages of each patient taking the drug. In Table II, the ages of the patients are 10, 11, 12 and 15.

3) Distances Calculation: The distances using all the six metrics applied to the THIN data are calculated and normalised. The normalisation of the distances is to demonstrate that the small distances that are close to 0 indicate similar patients, while the large distances that are near to 1 indicate dissimilar patients. In the case of Euclidean, Minkowski and Manhattan metrics, the data in Table II has been used to calculate the distances using equations (1), (2) and (3), respectively. For the calculation of Hamming distances, the number of different values between two of equal length sequences from Table II has been taken into account. The normalisation of the distances has been calculated using the formula: *norm-dist.* $(x) = x - \min(x)/\max(x) - \min(x)$ where x refers to the distance between two patients. Regarding the *pq*-gram metric, the distance between two trees of patients is defined as a symmetric difference between the two sets of *pq*-grams using equation (4), while the normalisation of the *pq*-gram distances is calculated using equation (5). On the other hand, the Edit Distance distances are equal to the minimum number of edit operations (insert, delete or rename nodes) when converting one tree to another. Each edit operation has cost 1 and based on the distance being equal to the minimum cost of converting T_1 to T_2. The Tree Edit Distance Normalisation (TED_NORM) is:

$$TED\_NORM(T_1, T_2) = \frac{TED(T_1, T_2)}{(|T_1| + |T_2|)} \tag{6}$$

Where $(|T1|+|T2|)$ means the sum of the two trees' nodes. The results of calculating the distances using all the six metrics are summarised in Table III and Table V for DESLORATADINE and DOXEPIN, respectively. The tables contain all the smallest normalized distances for patients (the most similar data) among the other distances.

The results for the first drug show that geometric and hamming metrics could find similar patients as the distance between two patients equal to zero. In contrast, the *pq*-gram and Edit Distance metrics produced a very few similar patients, like (a670605Up, a670602uS) and (a67340327, a681001KN) besides others who have some similarity or close distances to the identical level between patients. The reason behind that is related to the structure of the data which is a hierarchical tree structure.

On the other hand, the experiment for the second drug also produced a number of similar patients in their medical events based on the geometric and hamming metrics as shown in Table V, while for the *pq*-gram and Edit Distance metrics the table shows no similar distances. The reason behind that could be the lack of data for the second drug.

4) Clustering the Distances: The results in Table III and Table V show the similarities and closest distances between patients using the previously mentioned metrics. The following step of this work has been to use a clustering method to verify our results, to give the first insight on how the data looks like and to find which distance metric can represent similar distances better than the others. The clustering process has been also used to show whether all the similar distances in Tables III and V fall in one cluster or are distributed over all or some clusters. In this work, we used the *k*-means method and we chose the number of clusters to be equal to three clusters. For the first drug, two figures are reported to show the clusters of patients (Fig. 6(a) and Fig. 6(b)) using Euclidean and *pq*-gram distance metrics (a metric from each group of metrics). Fig. 6(c) and Fig. 6(d) show the clusters of patients using the same metrics for the second drug.

Since the *k*-means algorithm is known to be biased by the starting positions, it needs to be re-run more than once. As a result, we may get more than one outcome. The figures of the clusters represented in this work are those resulting from the most frequent clustering (the majority vote, in our experiments 10 times running). In order to distinguish between the clusters, we report Table VI and Table VII that contain the number of patients in each cluster for the first and second drug, respectively. *Cluster*1 in the tables contains the number of all the patients who are similar to each other, whereas *cluster*2 contains all the patients who have large distances between each

TABLE III
SMALLEST NORMALISED DISTANCES FOR PATIENTS TAKING DESLORATADINE DRUG

	The Normalised Distances						
patient's ID	Euclidean	Minkowski	Manhattan	Hamming	Edit Distance	1, 3-Grams	2, 3-Grams
a670605Up, a670602uS	0	0	0	0	0.25	0	0
a6732002X, a673200WF	0	0	0	0	0.25	0.888889	1
a6732002X, a673201@y	0	0	0	0	0.25	0.888889	1
a673200tm, a673201j7	0	0	0	0	0.25	0.888889	1
a673201@y, 673200WF	0	0	0	0	0.25	0.888889	1
a673201Wt,a6732025y	0	0	0	0	0.25	0.888889	1
a673201wI, a678701pI	0	0	0	0	0.6666	0.888889	1
a67340327, a681001KN	0	0	0	0	0.25	0	0
a677505bO, a677505pe	0	0	0	0	0.25	0.888889	1
a683104@Y, 677505bO	0	0	0	0	0.6666	0.888889	1
a683104@Y, a677505pe	0	0	0	0	0.6666	0.888889	1
a673201wI, a777805mH	0	0	0	0	0.25	0.888889	1
a673402zw, a683105Bk	0	0	0	0	0.25	0.888889	1
a678701pI, a777805mH	0	0	0	0	0.25	0.888889	1
a791600uB,a777806FG	0	0	0	0	0.25	0.888889	1
a7916065T, a777800Gj	0	0	0	0	0.25	0.888889	1

TABLE IV
THE SHARED MEDICAL EVENTS FOR PATIENTS IN TABLE III

Patient's ID	The medical events For the first patient	The medical events for the second patient	The description of the event
a670605Up, a670602uS	1B8	1B8	Itchy eye symptom
a6732002X, a673200WF	17Z	17Z	Respiratory symptom NOS
a6732002X, a673201@y	17Z	17Z	Respiratory symptom NOS
a673200tm, a673201j7	ZL5	ZL5	Referral to orthopaedic surgeon
a673201@y, 673200WF	17Z	17Z	Respiratory symptom NOS
a673201Wt,a6732025y	740	740	Submucous diathermy to turbinate of nose
a673201wI, a678701pI	H05	H05	Upper respiratory tract infection NOS
a67340327, a681001KN	171	171	Cough
a677505bO, a677505pe	8B3	8B3	Medication requested
a683104@Y, 677505bO	8B3	8B3	Medication requested
a683104@Y, a677505pe	8B3	8B3	Medication requested
a673201wI, a777805mH	H05	H05	Upper respiratory tract infection NOS
a673402zw, a683105Bk	H17, 8B3	H17, 8B3	Hay fever or pollens, Medication requested
a678701pI, a777805mH	H05	H05	Upper respiratory tract infection NOS
a791600uB,a777806FG	H17	H17	Hay fever or pollens
a7916065T, a777800Gj	A78	A78	Verrucae warts or Molluscum contagiosum

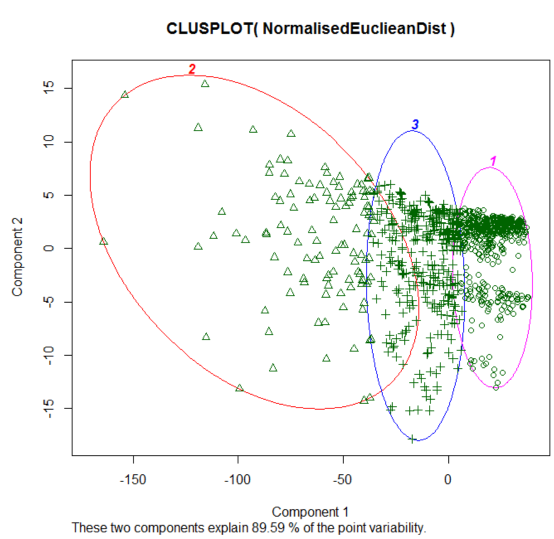
TABLE V
SMALLEST NORMALISED DISTANCES FOR PATIENTS TAKING DOXEPIN DRUG

	The Normalised Distances						
patient's ID	Euclidean	Minkowski	Manhattan	Hamming	Edit Distance	1, 3-Grams	2, 3-Grams
a793901c8,a9910027z	0	0	0	0	0.4761	0.971	0.967
b977401S1,a999104cU	0	0	0	0	0.4	0.923	1
g989501KB,a999104cU	0	0	0	0	0.375	1	1
g989501KB,b990804AL	0	0	0	0	0.5	1	1
b990804AL, a999104cU	0	0	0	0	0.375	1	1

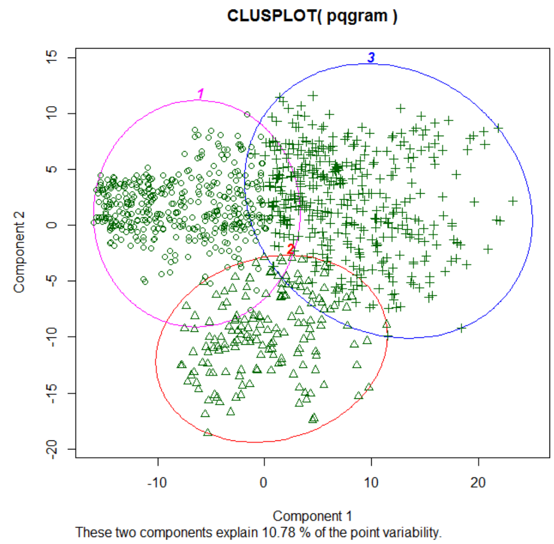
other. The remaining patients are grouped in *cluster*3.

IV. DISCUSSION

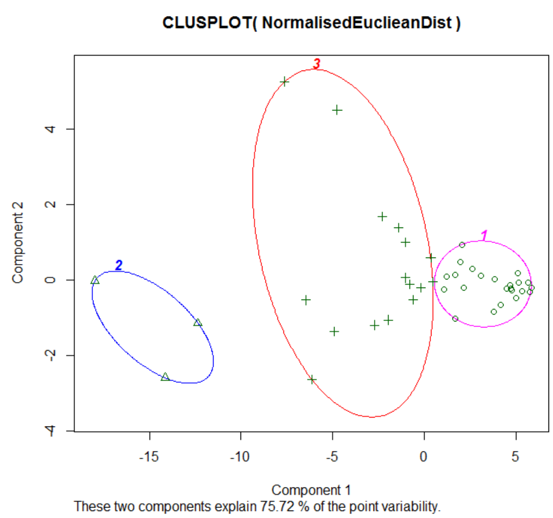
In this paper, six different distance metrics are applied to the THIN database for DESLORATADINE and DOXEPIN drugs. Our main objective is to find which metrics are useful for measuring distances in THIN data, with an emphasis on the *pq*-gram metric which is designed for hierarchical data like the read codes in THIN. We have implemented the distance metrics using two different types of data structures and compared their results. The two data structures are the tree-like structure of the group of *pq*-gram and Edit Distance metrics as shown in Fig. 4 and the frequency table or matrix for the group of geometric and Hamming metrics as shown in Table II. The distance metrics have been applied to the data and generally, the results revealed that these metrics produced good similarity distances between patients' data. Regarding the *pq*-gram, the distances depend mainly on the number of intersected *pq*-grams between two trees as well as the values of the parameters p and q. Choosing the correct values of p and q is a matter of tradeoffs. In [14], Srivastava et al. analysed the sensitivity of *pq*-gram distances with the values of p and



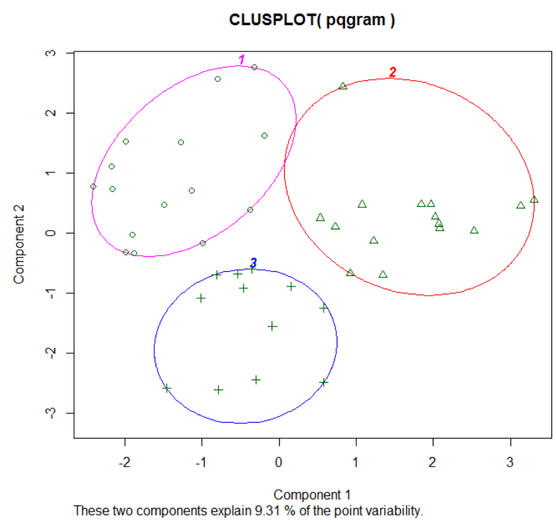
(a) The clusters of patients using Euclidean metric, DECLORATEDINE drug



(b) The clusters of patients using *pq*-gram metric, DECLORATEDINE drug



(c) The clusters of patients using Euclidean metric, DOXEPIN drug



(d) The clusters of patients using *pq*-gram metric, DOXEPIN drug

Fig. 6. The clusters of patients using Euclidean and *pq*-gram metrics

q and concluded that increasing p relative to q implies that more importance is being given to the ancestors than to the children of the trees, i.e. two nodes are considered to be the same only when they share p common ancestors.

Thus, in our case the smaller the value of p relative to q, the more probability of finding the intersected *pq*-grams between two trees and the more importance is given to the data rather than the structure of the trees. Based on that, the results in the seventh column in Table III and Table V on the preceding page are better compared to the results of the eighth column of the same tables. In general, the *pq*-gram metric is not the best metric compared to the other metrics as it depends on many parameters (p, q and the tree structure), but it could highlight some similar patients and measure the similarity between their data as shown in Table III (e.g. patients a670605Up, a670602uS and patients a67340327, a681001KN). On the other hand, Table V contains some non-similar distances produced by the *pq*-gram and Edit Distance metrics, for example the two patients (g989501KB and a999104cU) have the normalised distance equal to 1 which means there is no similarity between both patients' data. The reason behind that could be the lack of data for the DOXEPIN drug. That is to say, the more data available the more probability of having similar data for patients in the THIN database.

After finding all the distances using the chosen metrics, we verified our results by considering all the population of patients for each drug and by checking weather these distance

TABLE VI
THE NUMBER OF PATIENTS IN EACH CLUSTER FOR DESLORATADINE DRUG

	Cluster 1 (similar)	Cluster 2 (Non-similar)	Cluster 3 others
Euclidean Metric	513	114	361
Minkowski, p=3	578	89	321
Manhattan Metric	602	89	304
Hamming Metric	579	75	334
PQ-Gram Metric	409	164	415
Edit Distance Metric	284	332	372

TABLE VII
THE NUMBER OF PATIENTS IN EACH CLUSTER FOR DOXEPIN DRUG

	Cluster 1 (similar)	Cluster 2 (Non-similar)	Cluster 3 others
Euclidean Metric	23	3	16
Minkowski, p=3	81	7	17
Manhattan Metric	23	3	16
Hamming Metric	23	3	16
PQ-Gram Metric	15	15	12
Edit Distance Metric	16	15	11

metrics discriminate sufficiently using clustering the patient population. Fig. 6(a), Fig. 6(b), Fig. 6(c) and Fig. 6(d) show the results of clustering using the k-means algorithm. The latter is the simplest clustering method and requires the number of clusters to be known in advance. In this work, we chose the number of clusters to be equal to 3. However, more proper data analysis is required for future work and more than three clusters might be considered. The clusters have been plotted using the *clusplot* function from R software which is representing all the observations by points in the plots using the principal component analysis [15]. PCA is used in the data set for the purpose of visualisation and no feature selection has been carried out. The clusters are labeled using numbers (1, 2 and 3) as shown in Fig. 6 and the geometric and Hamming metrics discriminate successfully on the population for both drugs. We chose only two figures for each drug, one for each group of metrics. Table VI and Table VII show the number of patients in each cluster. The patients in Table III are grouped in *cluster*1 for all the metrics used, while the patients in Table V are grouped in *cluster*1 for the geometric and Hamming distance metrics only. In contrast, the distances for the same patients using *pq*-gram and Edit Distance metrics have a very poor similarity. Thus *cluster*1 for the both metrics contains some similar distances other than those in Table V. The reason behind that probably is the lack of data related to the second drug. In general, *cluster*1 in Table VI and Table VII contains the similar patients who have all or some medical events related to the drug in common, while *cluster*2 contains the non-similar ones. All the other patients who are not in *cluster*1 or *cluster*2 are grouped in *cluster*3 as shown in Fig. 6(a), Fig. 6(b), Fig. 6(c) and Fig. 6(d).

V. SUMMARY AND CONCLUSION

Two groups of distance metrics have been considered for two kinds of data structures from the THIN longitudinal health-care database, and then compared. The comparison is done by firstly looking at whether each metric can measure any distances and if all the metrics find the same similar patients with the same distances, and secondly by clustering the whole population of patients to find if the metrics sufficiently discriminate those patients. The results show that the two groups of metrics worked successfully in finding similar distances for similar patients and group all them in one cluster when clustering using the k-means algorithm.

In conclusion, the *pq*-gram metric might not be the best metric for THIN data, but it can measure similar distances and group them in one cluster. That is to say, it highlighted some known medical events related to the drugs been taken, for example the cough and itchy eye symptoms related to DESLORATADINE drug. As each group of metrics depends on different data structures and in order to choose the appropriate distance measure for the THIN data, we may need an appropriate structure of the data: for example, a mixed data structure from both the hierarchical and non-hierarchical data. By making the tree structure for all the levels of read codes, the distances can be calculated for read codes only. As a result of that, the *pq*-gram could find the related medical codes to each other in a better way.

REFERENCES

[1] N. Augsten, M. Böhlen, and J. Gamper, "The pq-gram distance between ordered labeled trees," *ACM Transactions on Database Systems (TODS)*, vol. 35, no. 1, p. 4, 2010.

[2] K. Kailing, H.-P. Kriegel, and S. Schönauer, "Content-based image retrieval using multiple representations," in *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, 2004, pp. 982–988.

[3] K. Zhang and D. Shasha, "Simple fast algorithms for the editing distance between trees and related problems," *SIAM journal on computing*, vol. 18, no. 6, pp. 1245–1262, 1989.

[4] R. Cordeiro de Amorim and B. Mirkin, "Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering," *Pattern Recognition*, vol. 45, no. 3, pp. 1061–1075, 2012.

[5] R. Shahid, S. Bertazzon, M. L. Knudtson, and W. A. Ghali, "Comparison of distance measures in spatial analytical modeling for health service planning," *BMC health services research*, vol. 9, no. 1, p. 200, 2009.

[6] J. Reps, J. M. Garibaldi, U. Aickelin, D. Soria, J. E. Gibson, and R. B. Hubbard, "Discovering sequential patterns in a uk general practice database," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2012, pp. 960–963.

[7] J. Reps, J. Feyereisl, J. M. Garibaldi, U. Aickelin, J. E. Gibson, and R. B. Hubbard, “Investigating the detection of adverse drug events in a uk general practice electronic health-care database,” *UKCI, the 11th Annual Workshop on Computational Intelligence, Manchester*, 2011.

[8] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[9] J. F. Committee and R. P. S. of Great Britain, *British national formulary (bnf)*. Pharmaceutical Press, 2012, vol. 64.

[10] J. C. Oxtoby, “Metric and topological spaces,” in *Measure and Category*. Springer, 1971, pp. 39–41.

[11] T. Körner, “Metric and topological spaces,” 2010.

[12] J. C. Gower, “Euclidean distance geometry,” *Mathematical Scientist*, vol. 7, no. 1, pp. 1–14, 1982.

[13] S. Hosangadi, “Distance measures for sequences,” *arXiv preprint arXiv:1208.5713*, 2012.

[14] N. Srivastava, V. Mishra, and A. Bhattacharya, “Analyzing the sensitivity of pq-gram distance with p and q,” *ACM*, 2010.

[15] M. Maechler, *Cluster Analysis Extended Rousseeuw et al.*, R CRAN, 2013.