Analysis of diabetic patients through their examination history

Dario Antonelli[a], Elena Baralis[b], Giulia Bruno[a], Tania Cerquitelli[b,*], Silvia Chiusano[b], Naeem Mahoto[b]

[a]*Department of Production Systems and Economics, Politecnico di Torino, Turin, Italy*
[b]*Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy*

Abstract

The analysis of medical data is a challenging task for health care systems since a huge amount of interesting knowledge can be automatically mined to effectively support both physicians and health care organizations. This paper proposes a data analysis framework based on a multiple-level clustering technique to identify the examination pathways commonly followed by patients with a given disease. This knowledge can support health care organizations in evaluating the medical treatments usually adopted, and thus the incurred costs. The proposed multiple-level strategy allows clustering patient examination datasets with a variable distribution. To measure the relevance of specific examinations for a given disease complication, patient examination data has been represented in the Vector Space Model using the TF-IDF method. As a case study, the proposed approach has been applied to the diabetic care scenario. The experimental validation, performed on a real collection of diabetic patients, demonstrates the effectiveness of the approach in identifying groups of patients with a similar examination history and increasing severity in diabetes complications.

Keywords: data mining, cluster analysis, patient examination history, diabetes

*Corresponding author

Email addresses: dario.antonelli@polito.it (Dario Antonelli), elena.baralis@polito.it (Elena Baralis), giulia.bruno@polito.it (Giulia Bruno), tania.cerquitelli@polito.it (Tania Cerquitelli), silvia.chiusano@polito.it (Silvia Chiusano), naeem.mahoto@polito.it (Naeem Mahoto)

Preprint submitted to

1. Introduction

Nowadays, large amount of medical data, storing the medical patient history, is collected by health care organizations. Data mining, which focuses on studying effective and efficient algorithms to transform large amounts of data into useful knowledge (Pang-Ning T. and Steinbach M. and Kumar V., 2006), may provide valuable insight into these huge data collections. For example, data mining techniques can be used to extract a variety of information on the patient history such as the medical protocols usually adopted for patients with a given disease. Healthcare organizations can exploit this knowledge to improve their current processes, assess new medical guidelines, or enrich the existing ones. Medical guidelines represent standard medical pathways specifying the actions necessary to treat, with optimal effectiveness and efficiency, patients with a given disease.

This study addresses the problem of analysing patients' examination data to identify the examination pathways commonly followed by patients. This issue is crucial for health care organizations, because it can significantly impact on the effectiveness of the medical treatments as well as on the costs incurred by the organizations.

The data analysis framework proposed in this paper exploits a multiple-level clustering approach to discover, in a data collection with a variable distribution, cohesive and well-separated groups of patients with a similar examination history. Cluster analysis is an exploratory data mining technique that partitions a data object collection into groups based on object properties, without the support of additional a priori knowledge (in contrast with classification algorithms using class label information) (Pang-Ning T. and Steinbach M. and Kumar V., 2006). To cluster data collections with a variable distribution, the proposed multiple-level clustering strategy iteratively focuses on disjoint dataset portions and locally identifies clusters. Among the state-of-the-art clustering techniques, the density-based DBSCAN algorithm (Ester et al., 1996) has been adopted due to its properties. DBSCAN allows the identification of arbitrarily shaped clusters, is less susceptible to noise and outliers, and does not require the specification of the number of expected clusters in the data. To highlight the relevance of specific examinations for a given clinical condition, in the proposed framework patient examination data has been represented in the Vector Space Model (VSM) (Salton G., 1971) using the TF-IDF method (Pang-Ning T. and Steinbach M. and Kumar V., 2006).

As a reference case study, the proposed framework has been applied to a real dataset of diabetic patients provided by the National Health Center (NHC) of the Asti province (Italy). The results showed that, starting from a large collection of raw examination data, the framework allows the identification of clusters containing patients with a similar examination history. More specifically, clusters contain patients with increasing disease severity, as patients are tested with more and more specific examinations to diagnose diabetes complications. The results were discussed with the support of clinical domain experts, showing a fairly good correlation among the examination pathways suggested by the clusters and the guidelines for diabetes disease (ICD-9-CM, 2011).

The paper is organized as follows. Section 2 describes the state-of-the-art clustering methods and explains the selection of the DBSCAN algorithm for this study. Section 3 analyses previous related work. Section 4 presents the proposed framework and describes its building blocks, while the results obtained for the real diabetic patient dataset are discussed in Section 5. Finally, Section 6 draws conclusions and future work.

2. Selection of the clustering algorithm

Cluster analysis partitions objects into groups (clusters) so that objects within the same group are more similar to each other than those objects assigned to different groups (Pang-Ning T. and Steinbach M. and Kumar V., 2006). Clustering algorithms require the definition of a metric to evaluate the similarity (or distance) between objects based on the features describing objects.

Clustering algorithms can be classified into the following four categories (Pang-Ning T. and Steinbach M. and Kumar V., 2006): (i) center, (ii) density, (iii) model, and (iv) hierarchical-based methods.

In *center-based methods* (e.g., K-means (Juang & Rabiner, 1990)), a cluster is a set of data objects in which each object is closer (more similar) to the prototype that defines the cluster than to the prototype of any other cluster. The prototype is the most representative point in the cluster. These methods find spherical-shaped clusters, unless clusters are well separated, and are sensitive to outliers.

In *density-based methods* (e.g., DBSCAN (Ester et al., 1996)), a cluster is a dense area of data objects surrounded by an area of low density. These

approaches are less sensitive to the presence of outliers than center-based techniques and can identify non-spherical shaped clusters.

Model-based methods (e.g. EM (G. McLachlan and T. Krishnan, 1997), COBWEB (Fisher, 1987a,b)) hypothesize a mathematical model for each cluster, and then determine the best fit between the model and the object collection. These methods can deal with outliers and noise. However, similarly to K-means, EM requires the specification of the number of expected clusters.

Hierarchical-based methods exploit an agglomerative or divisive approach to generate a hierarchical collection of clusters. The (most common) agglomerative approach (Pang-Ning T. and Steinbach M. and Kumar V., 2006) initially assigns each data object to a singleton cluster. The two closest clusters are then iteratively merged using a cluster proximity measure (e.g., single-link, complete-link, or group average average) (Pang-Ning T. and Steinbach M. and Kumar V., 2006). These methods are often used when the underlying application requires the creation of a taxonomy, which is not the case for our application scenario.

Density-based methods showed remarkable properties for clustering patients based on their examination history. More specifically, the very effective density-based algorithm DBSCAN (Ester et al., 1996) has been selected in this study. Differently from other algorithms (e.g., K-means), DBSCAN is less sensitive to outliers and can find arbitrarily shaped clusters. Outliers, when not identified and isolated as in DBSCAN, are clustered together with the other data objects, thus decreasing cluster cohesion. In addition, DBSCAN does not require an a priori specification of the number of clusters in the data, as opposed to K-means and EM.

Medical datasets can include outliers as specific examination pathways for some disease conditions and clusters can be non-spherical shaped. Besides, since our aim is discovering the examination pathways usually adopted for a given disease through an explorative data analysis, the expected number of clusters can be hardly guessed a priori. For these reasons, the DBSCAN algorithm has been adopted for the cluster analysis in this study. The main characteristics of DBSCAN are reported in the following Section 2.1.

To discover clusters in datasets with a variable distribution, state-of-the-art clustering algorithms can be applied in a multiple-level fashion to focus on different dataset portions and locally identify cluster (e.g., the bisecting K-means algorithm (Pang-Ning T. and Steinbach M. and Kumar V., 2006)).

The multiple-level strategy adopted in this study exploits the DBSCAN algorithm to select the dataset part analyzed at each iteration and locally cluster it.

2.1. The DBSCAN algorithm

The DBSCAN algorithm (Ester et al., 1996) relies on two input parameters, named Eps and $MinPts$, to define a density threshold in the data space. A dense region in the data space is a n-dimensional sphere with radius Eps and containing at least $MinPts$ objects.

The DBSCAN algorithm iterates over the data objects in the collection by analyzing their neighborhood. It classifies objects as being (i) in the interior of a dense region (a core point), (ii) on the edge of a dense region (a border point), or (iii) in a sparsly occupied region (a noise or outlier point). Any two core points that are close enough (within a distance Eps of one another) are put in the same cluster. Any border point close enough to a core point is put in the same cluster as the core point. Outlier points (i.e., points far from any core point) are isolated.

DBSCAN can discover arbitrarily shaped clusters and identify outliers as objects in a low density area in the data space. The effectiveness of DBSCAN is affected by the selection of the Eps and $MinPts$ values. Section 4.2 discusses how this issue has been addressed in this study.

3. Related work

Several works exploited data mining techniques to analyze medical data by addressing different pathologies and facets of various diseases.

Clustering algorithms have been widely exploited to analyze medical data for patients affected by different diseases (Ahmed & Funk, 2011; Buczak et al., 2009; Choong et al., 2000; Mulroy et al., 2003; Van Rooden et al., 2010; Santamaria et al., 2003). (Van Rooden et al., 2010) reviewed the cluster methods used to identify Parkinson's disease subtypes. It showed that the K-means algorithm (Juang & Rabiner, 1990) was mostly adopted, but also highlited its two major limitations, i.e., the sensitiveness to outliers and the need of defining the expected number of clusters. To overcome these issues, the DBSCAN algorithm has been used in this study, being able to internally evaluate the optimal number of clusters and automatically identify outliers.

In (Buczak et al., 2009), patients were represented by tracking the number of occurrences for each clinical event (e.g., hospital visit, lab order). Patients

were then grouped using a hierarchical agglomerative clustering method with Ward's linkage as proximity measure (Pang-Ning T. and Steinbach M. and Kumar V., 2006). Differently from (Buczak et al., 2009), in this study we exploit the TF-IDF scheme to weight the relevance of specific examinations for each diabetes condition. In addition, a multiple-level cluster approach is used to identify groups of patients in datasets with a variable data distribution.

Concerning the analysis of medical data for diabetic patients, several works, mainly exploiting classification techniques, have been proposed (Karegowda et al., 2012; Meng et al., 2012; Mohamudally & Khan, 2011; Zhong et al., 2012). Classification is a supervised data mining approach that assigns new unlabeled data to a class label by means of a model built from data with known class label (Pang-Ning T. and Steinbach M. and Kumar V., 2006).

In (Karegowda et al., 2012), diabetic patients were categorized with a K-nearest neighbor (KNN) classifier (Pang-Ning T. and Steinbach M. and Kumar V., 2006). To ehnance classification accuracy, in a pre-processing step a genetic algorithm and a feature selection technique (i.e., the correlation method (Pang-Ning T. and Steinbach M. and Kumar V., 2006)) are used to identify the relevant features for classification. In (Meng et al., 2012) the authors compared three prediction models (i.e., logistic regression, decision tree, and artificial neural networks) for diabetic patients classification. The comparison was based on common risk factors collected from both diabetic and pre-diabetic patients with a standard questionnaire. This work showed that the decision tree model (C5.0 (Pang-Ning T. and Steinbach M. and Kumar V., 2006)) yielded the best accuracy, followed by logistic regression, and artificial neural networks. The work in (Zhong et al., 2012) proposed a multi-level support vector machine approach to classify and predict clinical charge profiles as well as the length of hospital stay for patients affected by heart, diabetes, and cancer diseases. In (Mohamudally & Khan, 2011), different data mining algorithms (the K-means algorithm, C4.5 decision tree classifier, artificial neural networks, and 2D graphs for data visualization) are integrated to predict, classify, and visualize a medical diabetic dataset.

A parallel effort has been devoted to exploit clustering techniques for diabetic patients by addressing different issues as food analysis (Phanich et al., 2010), gait patterns (Sawacha et al., 2010), and relationships among diabetes and risk factors (Chaturvedi, 2003). Food clustering analysis for diabetic patients has been proposed in (Phanich et al., 2010). Using Self-Organizing Map (SOM) and the K-means algorithm, this work provided a Food Recommendation System suggesting proper substituted foods. The

work in (Sawacha et al., 2010) proposed a cluster analysis of biomechanical data to group patients with similar diabetic gait patterns. In (Chaturvedi, 2003) the spatial clusters of diabetes prevalence in Texas has been analyzed. The relationship of risk factors (i.e., age and obesity) associated with diabetes have also been analyzed. Differently from these works, this study exploits clustering techniques to identify groups of patients with similar examinations history.

4. Methodology

The proposed framework to analyse the patient examination history contains four main steps: (i) data collection, (ii) data transformation, (iii) cluster analysis, and (iv) cluster evaluation. The building blocks of the framework are shown in Figure 4 and detailed in the following subsections.

The patient examination log data is first collected and then transformed using the Vector Space Model (VSM) representation (Salton G., 1971). Examination frequencies are weighted through the Term Frequency (TF) - Inverse Document Frequency (IDF) scheme (Pang-Ning T. and Steinbach M. and Kumar V., 2006).

The multiple-level clustering approach is applied to identify, in a dataset with a variable distribution, groups of patients with a similar examination history. The DBSCAN algorithm has been exploited for the cluster analysis.

Finally, clustering results are evaluated through a quality index balancing both intra-cluster homogeneity and inter-cluster separation. Silhouette (Rousseeuw, 1987) has been considered as reference index. Cluster sets are also analysed together with a domain expert to assess their significance. To analyse the examination pathways represented by the cluster set, each cluster has been characterized with the examinations appearing in it.

4.1. Representation of the patient examination history

The data recording the patient examination history is represented using the Vector Space Model (VSM) (Salton G., 1971). Each patient is a vector in the examination space. Each vector element corresponds to a different examination and is associated with a weight describing the examination relevance for the patient. More specifically, it reports the weighted number of times the examination was repeated by the patient. The Term Frequency (TF) - Inverse Document Frequency (IDF) scheme (Pang-Ning T. and Steinbach M. and Kumar V., 2006) has been adopted to weight examination frequency.

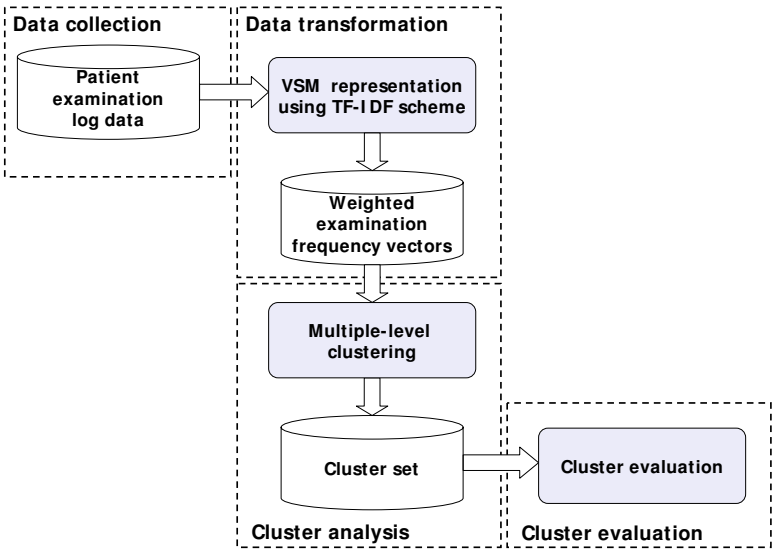

Figure 1: The proposed framework for the cluster analysis of patient examination data

Both the VSM representation and the TF-IDF scheme have been applied in previous works to represent text documents.

The adopted data representation allows highlighting the relevance of specific examinations for a given patient condition. The TF-IDF value increases proportionally to the number of times an examination appears in the patient history, but is offset by the frequency of the examination in the patient collection, which helps to control the fact that some examinations are generally more common than others. Unweighted examination frequencies do not properly characterize the patient condition, since standard routine tests usually appear with high frequency, while more specific tests may appear with lower frequency.

More formally, let $\mathcal{D}$ be a collection of patient records and $\Sigma = \{e_1, \ldots, e_k\}$ the set of examinations done by at least one patient in $\mathcal{D}$. Each patient p_i in $\mathcal{D}$ is represented by a weighted examination frequency vector v_{p_i} of $|\Sigma|$ cells. Each element $v_{p_i}[j]$ of vector v_{p_i} reports the weighted frequency w_{p_i,e_j} of examination e_j for patient p_i, i.e.,

$$v_{p_i} = [w_{p_i,e_1}, \ldots, w_{p_i,e_{|\Sigma|}}]. \quad (1)$$

The TF-IDF weight w_{p_i,e_j} for the pair (p_i, e_j) is computed as the product of two terms, called Term Frequency (TF_{p_i,e_j}) and Inverse Document Frequency (IDF_{e_j}),

$$w_{p_i,e_j} = TF_{p_i,e_j} * IDF_{e_j}. \quad (2)$$

The Term Frequency TF_{p_i,e_j} for the pair (p_i, e_j) represents the relative frequency of examination e_j for patient p_i. It is given by

$$TF_{p_i,e_j} = f_{p_i,e_j} / \sum_{1 \leq k \leq |\Sigma|} f_{p_i,e_k}, \quad (3)$$

where f_{p_i,e_j} is the number of times patient p_i underwent examination e_j and $\sum_{1 \leq k \leq |\Sigma|} f_{p_i,e_k}$ is the total number of examinations done by patient p_i.

The Inverse Document Frequency IDF_{e_j} for examination e_j represents the frequency of e_j in the patient collection. It is computed as

$$IDF_{e_j} = Log[|\mathcal{D}|/|p_k \in \mathcal{D} : f_{p_k,e_j} \neq 0|], \quad (4)$$

where $|\mathcal{D}|$ is the number of patients in the collection $\mathcal{D}$ and $|p_k \in \mathcal{D} : f_{p_k,e_j} \neq 0|$ is the number of patients in $\mathcal{D}$ who underwent (at least once) examination e_j. Mathematically, the base of the log function does not matter and constitutes a constant multiplicative factor towards the overall result.

The TF-IDF weight w_{p_i,e_j} for the pair (p_i, e_j) is high when examination e_j appears with high frequency in patient p_i and low frequency in patients in the collection $\mathcal{D}$. When examination e_j appears in more patients, the ratio inside the IDF's log function approaches 1, and the IDF_{e_j} value and TF-IDF weight w_{p_i,e_j} become close to 0. Hence, the approach tends to filter out common examinations.

4.2. The multiple-level DBSCAN approach for cluster analysis

Density-based algorithms can effectively discover clusters of arbitary shape and filter out outliers, thus increasing cluster homogeneity. Clusters are identified as dense areas of data objects surrounded by an area of low density.

In the DBSCAN algorithm, density is evaluated based on the user-specified parameters *Eps* and $MinPts$. One single execution of DBSCAN discovers dense groups of patients according to one specific setting for these parameters. Patients in lower density areas are labeled as outliers and not assigned

to any cluster. Hence, different parameter settings are needed to discover clusters in datasets with a variable data distribution as the one considered in this study. Groups of patients with close examination histories may have both different cardinalities and densities, i.e., groups may contain examination histories with different degrees of similarity.

The proposed *multiple-level clustering* approach allows clustering datasets with a variable distribution by iteratively applying the DBSCAN algorithm on different (disjoint) dataset portions. The whole original dataset is clustered at the first level. Then, at each subsequent level, patients labeled as outliers in the previous level are re-clustered. The DBSCAN parameters Eps and $MinPts$ are properly set at each level.

In the following, Section 4.2.1 presents the measure used to evaluate the similarity between patient examination histories, while Section 4.2.2 describes how the DBSCAN parameters and the number of clustering levels have been selected.

4.2.1. Similarity between patient examination histories

The cosine similarity measure has been adopted to evaluate the similarity between the weighted examination frequency vectors representing the patient examination histories. This measure has been often used to compare documents in text mining (Steinbach et al., 2000).

Let p_i and p_j be two arbitrary patients in the collection $\mathcal{D}$. Let v_i and v_j be the corresponding weighted examination frequency vectors as in Equation 1. The cosine similarity between v_i and v_j is computed as

$$cos(v_i, v_j) = \frac{v_i \bullet v_j}{\|v_i\| \, \|v_j\|} = \frac{\sum_{1 \leq k \leq |\Sigma|} v_i[k] v_j[k]}{\sqrt{\sum_{1 \leq k \leq |\Sigma|} v_i[k]^2} \sqrt{\sum_{1 \leq k \leq |\Sigma|} v_j[k]^2}}, \qquad (5)$$

where $cos(v_i, v_j)$ is in the range [0,1]. $cos(v_i, v_j)$ equal to 1 describes the exact similarity of examination histories for patients p_i and p_j, while $cos(v_i, v_j)$ equal to 0 points out that patients have complementary histories (i.e., the sets of their examinations are disjoint).

4.2.2. Number of clustering levels and DBSCAN parameters

The dataset density is preliminarly analysed using the k-dist graph (Pang-Ning T. and Steinbach M. and Kumar V., 2006) to select both the number of

iterations for the multiple-level clustering approach and the Eps and $MinPts$ values for each iteration.

For each patient in the collection, the k-dist graph plots the distance to its k^{th} nearest neighbor according to their examination history. On the x-axis patients are sorted by the distance to the k^{th} nearest neighbor, while on the y-axis distances to the k^{th} nearest neighbor are reported. The k value corresponds to the $MinPts$ parameter. The y-axis represents possible values of the Eps parameter. By cutting the graph at a given value on the y-axis, the corresponding p_x value on the x-axis partitions the patient collection into the following two subsets. Patients placed on the left hand side of p_x are labeled by DBSCAN as core points, and those on the rigth side of p_x as outlier or border points.

Sharp changes in the k-dist graph identify dataset portions with a different density (Pang-Ning T. and Steinbach M. and Kumar V., 2006). The multiple-level strategy analyzes these dataset portions in different iterations. The Eps value to cluster each dataset part is selected in correspondence with the sharp change appearing in the graph. Section 5.2 discusses the selection of DBSCAN parameters and number of iteration levels for the diabetes dataset considered in this study.

4.3. Cluster evaluation

The discovered cluster set is evaluated using the Silhouette index (Rousseeuw, 1987). Silhouette allows evaluating the appropriateness of the assignment of a data object to a cluster rather than to another by measuring both intra-cluster cohesion and inter-cluster separation.

The silhouette value for a given patient p_i in a cluster C is computed as

$$s(p_i) = \frac{b(p_i) - a(p_i)}{max\{a(p_i), b(p_i)\}}, s(p_i) \in [-1, 1], \tag{6}$$

where $a(p_i)$ is the average distance of patient p_i from all other patients in the cluster C, and $b(p_i)$ is the smallest of average distances from its neighbour clusters. The silhouette value for a cluster C is the average silhouette value on all its patients. Negative silhouette values represent wrong patient placements, while positive silhouette values a better patient assignments. Clusters with silhouette values in the range [0.51,0.70] and [0.71,1] respectively show that a reasonable and a strong structure have been found (Kaufman, L. and

Rousseeuw, P. J., 1990). The cosine similarity metric has been used for silhouette evaluation, since this measure was used to evaluate patient similarity in the cluster analysis (see Section 4.2.1).

4.4. Data mining tool

The DBSCAN algorithm available in the RapidMiner system has been used for the cluster analysis within the proposed framework. The RapidMiner toolkit (Rapid Miner Project, 2013) is an open-source system consisting of a number of data mining algorithms to automatically analyze a large data collection and extract useful knowledge.

The procedures for data transformation and cluster evaluation have been developed in the Phyton programming language (Python Software Foundation, 2013). These procedures transform the patient examination log data into the VSM representation using the TF-IDF scheme and compute the silhouette values for the cluster set provided by the cluster analysis.

5. Results and discussion

This section presents and discusses the results obtained when analysing a real collection of examination log data for diabetic patients with the proposed framework. In the following, Section 5.1 describes the dataset considered in this study, while Section 5.2 specifies the framework configuration for the cluster analysis. The cluster results are presented and discussed in Section 5.3. Finally, Section 5.4 evaluates the performance of the multiple-level clustering approach in terms of execution time.

5.1. Diabetic patient dataset

The dataset considered in this study was collected by the Local Health Center (LHC) of the Asti province in Italy. The LHC database stores all the accesses to the health care system in the Asti province in the year 2007. From this database the examination log data of all patients with overt diabetes were extracted. Raw data contain 95,788 records with examinations performed by 6,380 patients. They contain both (i) routine and (ii) more specific examinations to analyze diabetes complications on various degrees of severity. The dataset includes both male and female patients in a wide age range (between 4 and 95 years). The diagnostic and therapeutic procedures are defined using the ICD 9-CM (International Classification of Diseases, 9th revision, Clinical Modification) (ICD-9-CM, 2011).

5.2. Selection of parameters for clustering diabetic patients

To select the number of iterations for the multiple-level clustering strategy and the DBSCAN parameter for each level, we relied on the k-dist graph as discussed in Section 4.2.2. In selecting these parameters, we addressed the following issues. To discover representative examination pathways for the diabetes, we aim at avoiding clusters including few patients. In addition, to consider all different patient examination histories, we aim at limiting the number of patients labeled as outliers and thus unclustered.

We plotted the k-dist graph, and analysed the cluster results, by varying the $MinPts$ (i.e., k) value. Low $MinPts$ values were not suitable for our purpose, because DBSCAN identifies small groups of patients. At the same time, large $MinPts$ values may increase the number of outlier points included in the clusters. The experimental results showed that $MinPts$ in the range [25,35] provided similar results, that were compliant with the above issues. We selected $MinPts$=30. From the k-dist graph for k=30, we observed three main dataset portions with a different density for Eps in the range [0.2-0.3], [0.4-0.5], and [0.6-0.7], respectively. Consequently, we adopted a three-level clustering approach, with each level focusing on one among the three dataset parts. The selected Eps values were 0.3, 0.4, and 0.6 for the first, second, and third clustering level, respectively.

5.3. Analysis of the clustering results

Starting from a large collection of raw examination data, the proposed framework allows the discovery of a set of clusters containing patients with a similar examination history. The multiple-level DBSCAN approach, iterated for three levels, computed clusters that progressively contain patients with increasing severity in diabetes, because patients are tested using more and more specialized examinations. More specifically, first-level clusters contain patients mainly undergoing routine tests to monitor diabetes conditions, or some basic tests to diagnose disease complications. Second-level clusters collect patients that are tested using an increasing number of examinations to diagnose some diabetes complications. Examinations become progressively more numerous and specific in third-level clusters, indicating patients that can have diabetes complications of increasing severity. Since at each level clusters contain more specific examinations, a lower number of patients is contained in each cluster and the cluster size tends to reduce progressively. Clusters show good cohesion and separation as they are characterized by high

silhouette values. The results, discussed with the support of a clinical domain expert, show a fairly good correlation among the examination pathways suggested by the clusters and the guidelines for diabetes disease (ICD-9-CM, 2011).

Cluster properties are discussed in detail in the following subsections. Tables 1, 2, and 3 report, for each first- and second-level cluster, the examination frequencies computed as the percentage of patients in the cluster tested by each examination. Clusters are named as C_{i_j} in the tables, where j denotes the level of the multiple-level DBSCAN approach providing the cluster and i locally identifies the cluster at each level j.

5.3.1. First-level cluster set

First-level clusters are reported in Tables 1 and 2. Clusters can be partitioned into the following two main groups: clusters containing patients (i) with standard examinations to monitor diabetes conditions (clusters C_{1_1}-C_{5_1}, in Table 1), (ii) coupled with basic examinations to diagnose disease complications (clusters C_{6_1}-C_{11_1} in Table 2).

The two largest clusters (C_{1_1} and C_{2_1}) contain patients who mostly performed standard routine tests. Besides routine examinations, all patients in cluster C_{3_1} had a specialistic visit and were tested with usual basic examinations to diagnose the most frequent diabetes complications, as risks for cardiovascular disease and eye problems. All patients in clusters C_{4_1} and C_{5_1} only had a checkup visit, together with the glucose level test in cluster C_{5_1}. These two clusters may include patients usually tested in private structures and periodically reporting test results to NHC.

Patients in clusters C_{6_1}-C_{11_1} were additionally tested to diagnose diabetes complications in the (a) eye (C_{6_1}), (b) cardiovascular system (C_{7_1}), (c) both eye and cardiovascular system (C_{8_1}), (d) carotid (C_{9_1}), and (e) limb (C_{10_1}). Finally, (f) cluster C_{11_1} includes tests for the liver, kidneys, and in particular cardiovascular system. Differently from second- and third-level clusters, diabetes complications were monitored in clusters C_{6_1}-C_{11_1} using a few (quite standard) tests which showed a limited degree of severity. Only the cardiovascular system has been thoroughly tested in cluster C_{11_1}. Standard routine examinations still appear in clusters C_{6_1}-C_{11_1} even though (usually) with a lower frequency than in clusters C_{1_1}-C_{5_1}.

Patients with an examination history significantly dissimilar from all the others are labeled as outliers and not included in any cluster. The DBSCAN

Table 1: Examination frequencies (%) in first-level clusters containing patients undergoing routine tests (Eps=0.3, $MinPts$=30)

Category	*Examination*	C_{1_1}	C_{2_1}	C_{3_1}	C_{4_1}	C_{5_1}
Routine	Checkup visit	78	96	58	100	100
	Glucose level	78	98	63	-	100
	Urine test	72	97	58	-	-
	Venous blood	96	75	35	-	-
	Capillary blood	72	97	58	-	-
	Haemoglobin	100	-	-	-	-
	Specialistic visit	-	13	100	-	-
Cardiovascular	Electrocardiogram	-	-	100	-	-
Eye	Fundus Oculi	-	-	28	-	-
Number of patients		223	1,764	43	110	41
Silhouette		0.67	0.55	0.85	0.99	1.0

algorithm was re-applied on these patients only (3,509 patients), with different parameter values. The results are discussed in Section 5.3.2.

5.3.2. Second-level cluster set

Second-level clusters contain patients with more diversified examination histories. More specifically, the following two main categories of clusters can be identified: (i) clusters containing patients tested with specific examinations to diagnose a given diabetes complication (clusters C_{1_2}-C_{2_2}); (ii) clusters with patients who underwent various examinations to diagnose different diabetes complications (clusters C_{3_2}-C_{5_2}). These two categories respectively indicate patients that can be seriously affected by a particular disease complication or by more than one disease complication at the same time. Second-level clusters are reported in Table 3.

Clusters C_{1_2} and C_{2_2} include specific examination to diagnose eye complications. More specifically, all patients in cluster C_{1_2} underwent a battery of tests to assess vision and ability to focus on objects (called "Eye examination" in Table 3). Instead, all patients in cluster C_{2_2} had Retinal photocoagulation, a laser operation done in cases of long-term eye complications, such as proliferative retinopathy.

All patients in clusters C_{3_2}-C_{5_2} may suffer complications on the cardiovascular, liver, and kidneys systems, but with different degrees of severity.

Table 2: Examination frequencies (%) in first-level clusters including basic tests to diagnose diabetes complications (Eps=0.3, $MinPts$=30)

Category	*Examination*	C_{6_1}	C_{7_1}	C_{8_1}	C_{9_1}	C_{10_1}	C_{11_1}
Routine	Checkup visit	77	78	66	62	68	97
	Glucose level	74	74	64	62	59	97
	Urine test	74	74	64	57	56	92
	Venous blood	57	60	53	48	44	97
	Capillary blood	74	73	63	55	56	92
	Haemoglobin	-	-	-	14	12	100
	Specialistic visit	-	-	-	-	-	-
	Complete blood count	-	-	-	-	3	3
Cardiovascular	Electrocardiogram	-	100	100	-	-	42
	Cholesterol	-	-	-	-	-	100
	HDL Cholesterol	-	-	-	-	-	100
	Triglycerides	-	-	-	-	-	100
Eye	Fundus Oculi	100	-	100	-	-	39
Liver	ALT	-	-	-	-	-	100
	AST	-	-	-	-	-	100
Kidney	Creatinine	-	-	-	-	-	3
	Creatinine clearance	-	-	-	-	-	100
	Culture urine	-	-	-	-	-	100
	Microscopic urine analysis	-	-	-	-	-	100
	Uric Acid	-	-	-	-	-	100
Carotid	ECO doppler carotid	-	-	-	100	-	-
Limb	ECO doppler limb	-	-	-	-	100	-
Number of patients		294	144	140	42	34	36
Silhouette		0.65	0.74	0.79	0.95	0.97	0.90

Patients in cluster C_{5_2} are at risk of more severe liver complications than those in clusters C_{3_2} and C_{4_2}, since their examination history includes more tests in this category. Analogously, cluster C_{3_2} is characterized by more severe renal complications than clusters C_{4_2} and C_{5_2}. Cardiovascular complications have similar severity in clusters C_{3_2}-C_{5_2}, because examinations in this category appear with similar frequency.

At this stage, 2,939 patients are classified as outliers and not assigned to any cluster. We further applied the DBSCAN algorithm on them by modifying the parameter setting. The results are analyzed in Section 5.3.3.

5.3.3. Third-level cluster set

The results collected at this stage (with DBSCAN parameter Eps=0.6 and $MinPts$=30) show a similar trend to second-level clusters. Third-level clusters contain patients that may suffer more complications at the same time (e.g., complications on carotid, liver, and cardiovascular systems) and are tested with more specific examinations (e.g., transcutaneous oxygen and carbon dioxide monitor or upper abdominal ultrasound). By stopping the multiple-level DBSCAN approach at this level, only 1,239 patients labeled as outliers remain unclustered, with respect to the initial collection of 6,380 patients considered at the first level (i.e., about 19% of patients). By further applying the DBSCAN algorithm on this outlier set, fragmented groups of patients can be identified. These clusters can represent patients affected by more rare diabetes complications, and thus with examinations different from those done by most patients.

5.4. Execution time

The run time of DBSCAN at the first, second, and third level is 9 min 10 sec, 2 min 25 sec, and 1 min 45 sec, respectively. The run time progressively reduces because less patients are considered at each subsequent level.

6. Conclusion

To get the most out of large and complex medical databases, innovative data analysis techniques are needed to extract useful knowledge in a timely fashion. In this paper a data analysis framework based on a multiple-level clustering approach has been proposed to identify groups of patients with a similar examination history in a dataset with a variable data distribution. The TF-IDF method has been exploited to represent patient examination

Table 3: Examination frequencies (%) in second-level clusters, including examinations to diagnose more severe diabetes complications (Eps=0.4, $MinPts$=30)

Category	*Examination*	C_{1_2}	C_{2_2}	C_{3_2}	C_{4_2}	C_{5_2}
Routine	Checkup visit	65	90	22	98	71
	Glucose level	52	92	22	100	95
	Urine test	37	90	19	98	68
	Venous blood	35	84	100	100	98
	Capillary blood	37	85	17	98	63
	Haemoglobin	13	34	100	98	83
	Specialistic visit	-	13	-	-	-
	Complete blood count	3	15	45	7	93
Cardiovascular	Electrocardiogram	17	21	70	47	20
	Cholesterol	7	25	100	97	93
	HDL Cholesterol	4	26	100	99	92
	Triglycerides	7	26	100	97	90
Eye	Fundus Oculi	50	38	53	49	27
	Angioscopy	-	30	-	-	-
	Eye examination	100	5	-	-	-
	Renital photocoagulation	-	100	-	-	-
Liver	ALT	-	21	10	98	98
	AST	-	21	8	97	98
	Bilirubin	-	2	-	-	100
	Gamma GT	-	-	100	-	95
Kidney	Creatinine	4	20	99	11	78
	Creatinine clearance	2	10	-	99	-
	Culture urine	2	11	67	97	44
	Microscopic urine analysis	4	16	2	82	73
	Uric acid	-	13	3	97	68
	Microalbuminuria	-	7	100	60	37
Carotid	ECO doppler carotid	-	5	-	-	2
Limb	ECO doppler limb	-	-	-	-	-
Number of patients		46	61	139	283	41
Silhouette		0.92	0.88	0.72	0.69	0.87

data, thus highlighting the relevance of the different examinations. The proposed methodology has been validated in the diabetic care scenario on a real dataset of diabetic patients. The analysis identified cohesive and well-separated groups of patients with standard or more specific examinations for diabetes, showing a good correlation with medical guidelines for diabetes (ICD-9-CM, 2011). Future developments of the proposed approach will explore the correlation between patient examination data and additional aspects of the medical treatments such as pharmaceutical drug therapies. Furthermore, we plan to apply the proposed approach to different medical contexts (e.g., cardiac patients).

7. Acknowledgments

The authors wish to thank Dr. Baudolino Mussa and Dr. Dario Bellomo for their advice and fruitfull discussions.

References

Ahmed, M. U., & Funk, P. (2011). Mining rare cases in post-operative pain by means of outlier detection. In *IEEE Symposium on Signal Processing and Information Technology (ISSPIT)* (pp. 35–41).

Buczak, A. L., Moniz, L. J., Feighner, B. H., & Lombardo, J. S. (2009). Mining electronic medical records for patient care patterns. In *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (pp. 146–153).

Chaturvedi, K. (2003). Geographic concentrations of diabetes prevalence clusters in texas and their relationship to age and obesity. *http://www.ucgis.org/summer03/studentpapers/kshitijchaturvedi.pdf. Retrieved*, *9*, 2010.

Choong, P. F. M., Langford, A. K., Dowsey, M. M., & Santamaria, N. M. (2000). Clinical pathway for fractured neck of femur: a prospective, controlled study. *Med J Aust.*, *172*, 423–426.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining (KDD)* (pp. 226–231).

Fisher, D. H. (1987a). Improving inference through conceptual clustering. In *National conference on Artificial intelligence (AAAI)* (pp. 461–465).

Fisher, D. H. (1987b). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, *2*, 139–172.

G. McLachlan and T. Krishnan (1997). *The EM algorithm and extensions*. Wiley series in probability and statistics. John Wiley and Sons.

ICD-9-CM, I. (2011). International Classification of Diseases, 9th revision, Clinical Modification. Available: http://icd9cm.chrisendres.com. Last access on March 2011, .

Juang, B.-H., & Rabiner, L. (1990). The segmental k-means algorithm for estimating parameters of hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *38*, 1639–1641.

Karegowda, A., Jayaram, M., & Manjunath, A. (2012). Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients. *International Journal of Engineering and Advanced Technology (IJEAT)*, (pp. 147 – 151).

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley.

Meng, X.-H., Huang, Y.-X., Rao, D.-P., Zhang, Q., & Liu, Q. (2012). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung Journal of Medical Sciences*, (pp. 93–99).

Mohamudally, N., & Khan, D. M. (2011). Application of a unified medical data miner (umdm) for prediction, classification, interpretation and visualization on medical datasets: the diabetes dataset case. In *International conference on Advances in data mining: applications and theoretical aspects* (ICDM) (pp. 78–95). Berlin, Heidelberg: Springer-Verlag.

Mulroy, S., Gronley, J., Weiss, W., Newsam, C., & Perry, J. (2003). Use of cluster analysis for gait pattern classification of patients in the early and late recovery phases following stroke. *Gait Posture*, *18*, 114–25.

Pang-Ning T. and Steinbach M. and Kumar V. (2006). *Introduction to Data Mining*. Addison-Wesley.

Phanich, M., Pholkul, P., & Phimoltares, S. (2010). Food recommendation system using clustering analysis for diabetic patients. In *IEEE International Conference on Information Science and Applications (ICISA)* (pp. 1–8).

Python Software Foundation, P. S. (2013). Python Programming Language Official Website. Available: http://www.python.org/ Last access on January 2013, .

Rapid Miner Project, R. M. (2013). The Rapid Miner Project for Machine Learning. Available: http://rapid-i.com/ Last access on January 2013, .

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, (pp. 53–65).

Salton G. (1971). *The SMART retrieval system: experiments in automatic document processing*. Prentice-Hall.

Santamaria, N., Houghton, L., & Kimmel, A., L. Graham (2003). Clinical pathways for fractured neck of femur: A cohort study of health related quality of life, patient satisfaction and clinical outcome. *Australian Journal of Advanced Nursing*, *20*, 24–29.

Sawacha, Z., Guarneri, G., Avogaro, A., & Cobelli, C. (2010). A new classification of diabetic gait pattern based on cluster analysis of biomechanical data. *Journal of Diabetes Science and Technology*, *4*, 1127–38.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *KDD Workshop on Text Mining*.

Van Rooden, S. M., Heiser, W. J., Kok, J. N., Verbaan, D., van Hilten, J. J., & Marinus, J. (2010). The identification of parkinson's disease subtypes using cluster analysis: A systematic review. *Movement Disorders*, *25*, 969–978.

Zhong, W., Chow, R., & He, J. (2012). Clinical charge profiles prediction for patients diagnosed with chronic diseases using multi-level support vector machine. *Expert Systems with Applications*, *39*, 1474 – 1483.