Similarity of patients in predictive models using medical data: case of auto-prescription drugs for diabetic patients

SOFIENE HABOUBI, ABIR BEN CHEIKH
Signals Images and Information Technologies Lab
National Engineering School of Tunis
University of Tunis El Manar
BP 37 Le Belvedere, 1002 Tunis
TUNISIA

Abstract: - Patient similarity analysis is a prerequisite for applying machine learning technology to medical data. The subject of this article is "Similarity of patients in predictive models using medical data". This article describes, how to create a system capable of analyzing patient data. We first implemented a methodology to extract useful information from raw data. We then determined, for the set of data extracted from a real database from a medical practice, the similarities that may exist between patients; based on several explanatory variables. We then derived a meaningful distance metric to measure the similarity between the patients represented by their key indicators. Thus, we have proven the importance of defining strong associations between determined attributes. We have developed a process to select the best attributes that will lead to the prediction. Finally, and after grouping the patients using the partition clustering approach (the k-medoid algorithm), we built a predictive linear regression model. For the learning phase, we combined different supervised and unsupervised techniques. We have chosen medical prescription as the area of application to predict the right medication for the patient. The results obtained show that the proposed approach can produce good predictions.

Key-Words: - Data Mining, Machine Learning, Medical Informatics, Similarity Measures, Multiple Linear Regression Model, k-Medoids Clustering.

1 Introduction

Health information has been increasingly computerized in electronic medical records (EMRs), which are systematic collections of electronic health information about individual patients or populations, including demographics, diagnostic history, medications, lab test results, vital signs, etc.

The efficient use of electronic health records is important for many medical applications such as clinical decision support, comparative research, and efficient disease modelling.

With the huge growth of electronic medical records, different sources of information are becoming available about patients.

The first objective is, how we can use this information data from EMR in a secure way, to improve patient outcomes without causing added effort on the part of doctors; To accomplish the goal of meaningful re-use of EMR data, the detection of similarities between patients becomes an important concept to facilitate analysis, reducing costs and improving healthcare systems.

The goal of patient similarity detection is to calculate the distance of similarity between two patients based on their EMR data.

With the correct patient similarity, numerous applications can be initiated:

- Recovery of cases of similar patients for a target patient.
- Comparison between patient from the same group (family, race, geographic location, …)

In common practice in medicine, the patient comes to the doctor, after the routine procedure and tests, the doctor checks the completed work-up, which is why a lot of data remains unexplored in the office, which is a big problem. in the domain of health.

Our objective is to provide an autonomous system for the research and analysis of data described in the medical records of each patient, to derive a measure of similarity between them and to make this machine capable of correctly predicting the right treatment, by applying

machine learning methods. Thus, throughout the brief, we are mainly interested in three questions:

1. How can we represent information about a patient?
2. How do we equip the machine with knowledge about a patient's condition?
3. How can this machine, like humans, process and use this information efficiently and produce with an acceptable degree of accuracy a prediction of a drug elicited by a situation?

2 Background

Patient similarity is the idea of tracking down the best and least powerful therapies dependent on the clinical records of similar individuals with practically identical medical conditions.

Patient Similarity Analysis can uphold the predictive model of health and treatment issues by getting to information about comparable patients in comparable circumstances and which medicines have frequently worked for them. Specialists by and large expect to coordinate with the best treatment based on their individual experience from past patients..

Data Analysis, including patient similarity calculations, can be applied to records from the Health Indicator Warehouse and other health information records. This analysis considers a more extensive areas and a bigger example. Thus, the bigger example improves the certainty of the analysis for more exact analyses and more proper treatment of medical conditions with less events of medication interaction and different complications.

Patient similarity can also be used for predictive patient analysis. Predictive analytics can help physicians predict potential health problems, and how long-term health effects of different medical conditions, rather than limiting themselves to the best treatments and likely short-term treatment outcomes.

Recently, several techniques have been published to predict certain condition outcomes using large data of patient cohorts. Two models are discovery of cardiovascular breakdown over a half year before the examination date of clinical conclusion [1] and derivation of patient anticipation dependent on patient similarities [2]. Subsequently, machine learning techniques [3] have been utilized to calculate patient similitude, for instance Support Vector Machines (SVM). Every one of these techniques depend on learning models requiring preparing sets of an adequate number of cases.

Salton was the first to propose the Vector Space Model (VSM) in 1968 [4], initially it is implemented in SMART [5]. An information retrieval framework that calculates the distance of similarity between structural features of reports. The matrix consists of 0/1 values: 1 : if the “word” present int the report, else 0. In 1986, Salton presented a considerable improvement in performance by using "Term Frequency Weight" [6] rather than of binary values.

Since SMART, VSM has been widely used in data information retrieval [7, 8], in classification [9, 10] and clustering [11].

The Vector Space Model was also used to calculate similarity in the biomedical field. With the goal of recognizing possibility connection between diseases based on genetic relationships. A transformation of the VSM is proposed by Sarkar et al. [12], that links knowledge about genes and inferred diseases through three knowledge bases: GenBank, Online Mendelian Inheritance in Man and Medline. This work shows, the relatedness between diseases, via this gene relationships, was resolved utilizing “Cosine Similarity Metric”.

Lee et al. [13] use MSM to apply data likeness based cosine similarity metric to an ICU information dataset to distinguish patients who are generally like each list patient and foresee their results. They applied a VSM to quantitative structured information and showed that their methodology outflanked standard seriousness scores regularly utilized in concentrated care units.

In 2013, in a published research paper, a Vector Space Model approach was applied to distinguish careful site contaminations after neurosurgical strategies in full content reports [14]. The strategy applied to patient account records accomplishes with a recall score (92%) and an accuracy of 40%.

Every one of these published works have recommended that the VSM approach might be compelling in similitude between patients data records.

3. Background

The proposed similarity approach consists of a learning part and a prediction part. These two parts are necessarily preceded by a data pre-processing step:

- The pre-processing phase consists in transforming the data into vectors of characteristics.

- The learning phase consists of taking the vectors as input and building our learning model as an output.
- The prediction phase consists of using a characteristic vector to predict the appropriate drug code for a new patient according to the learned model.

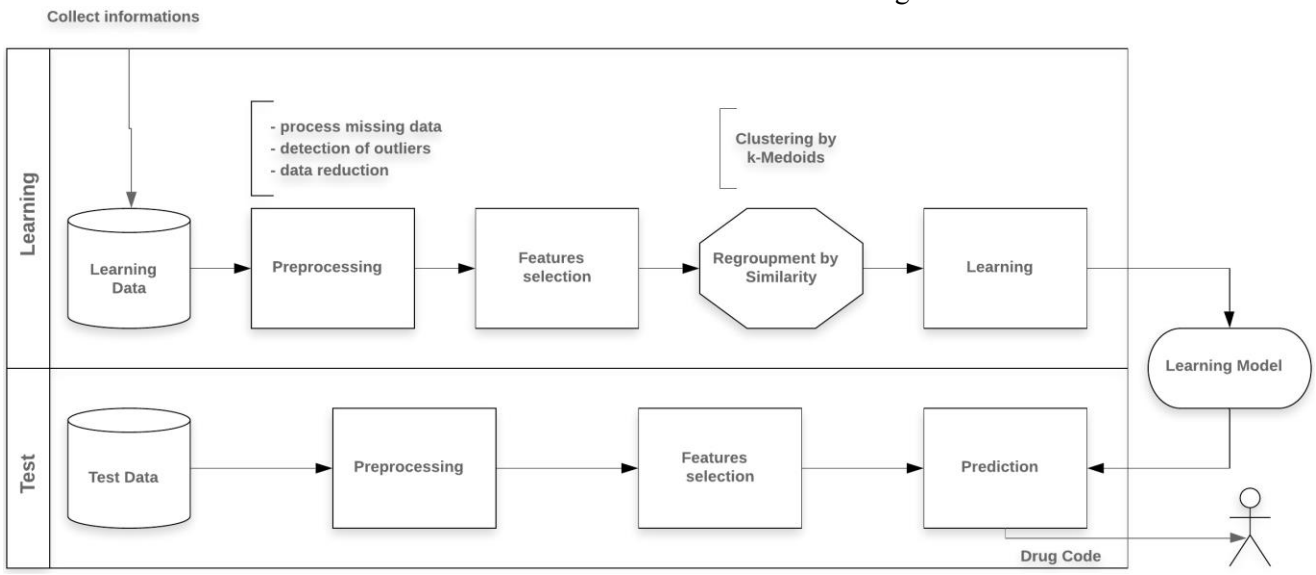

Figure 1. Similarity approach

Data preprocessing is one of the most complex steps in data mining, which deals with preparing and transforming data and at the same time seeking to make knowledge discovery more efficient. In addition, some modelling techniques are very sensitive to predictors, such as linear regression.
Our raw medical data are sensitive to missing values, noisy data, incomplete data, inconsistent data and outliers [10]. Indeed, it is therefore important that they are processed before being extracted.

3.1 Data Cleaning

Missing information is normal in real data, and it profoundly effects on the result of the last analysis, which can make the outcome unreadable.
There are various sorts of missing information. We ought to have a decent comprehension of why the information is absent. In the event that the information is missing aimlessly or if the vanishing is identified with a specific indicator, yet the indicator has no relation to the result, at that point the information data can in any case represent the population [11].
Numerous methods have been proposed to manage missing information data [12-14] and can generally be separated into two categories.
The first and easiest would be to delete missing data directly. In the event that the missing information is arbitrarily disseminated, or if the vanishing is identified with an indicator that doesn't relate with the reaction and the dataset is adequately huge, eliminating the missing information data has little impact on analysis performance.
The second category is to fill in the missing information based on the other information.

3.2 Outlier detection

The outlier is characterized as a removed perception point from standard information. The presence of outlier in our prescient model utilizing clinical information can break the limit of a model.
C. Grouping of patients by similarity based on the partitional clustering approach
One technique for extracting information from unlabelled data that can be very useful in our case is: the k-medoid algorithm [15].
The K-means classification algorithm is sensitive to outliers [15], because an average is strongly impacted by the outlier and therefore the center of cluster is determinate incorrectly. In this case, the k-medoid algorithm can correctly represent the center of our cluster, using a data with aberrant and unusual values.
So instead of using an imaginary midpoint as the center of a cluster that is not part of the cluster itself, K-medoids uses a representative real point as the cluster reference points to represent it. With a

minimal using distances from other vectors, Medoid represent the most central point in the cluster.
The basic strategy of the K-medoids clustering algorithms is to find k clusters in n objects by: Choose arbitrarily k-Medoids. Assign each remaining object to the nearest medoid. Randomly choose a non-medoid: for each non-medoid calculate the similarity with other medoid. for each non-medoid will be assigned to most similar medoid, then calculate the new clusters.

3.3 Learning

Regression analysis is a measurement method for contemplating the reliance of a reaction variable on at least one indicator, including prediction future estimations of a response, finding the main indicators, and assessing the effect of changing an indicator or treatment on the estimation of the response [9].
Descriptive statistics were used to characterize the basic features of the data established in the actual work.
Correlation analysis is used to study the relationships between the medication prescribed for the patient and the explanatory variables that go into choosing the most suitable medication.

3.4 Prediction

The notion of learning introduces the fact of learning a set of relationships between the criteria characterizing the element to be classified and its target class.
In order to determine the prediction of the drug code for a patient we used the operator "classification by regression" which is a nested operator, that is to say that it has a sub-process, or all the regression models. are combined into a classification model, we built a classification model using the regression learner provided in its sub-process.

4 Automatic prescription of drugs for diabetes

4.1 Database

During this project, we had the opportunity to exchange and work with doctors specializing in the disease of diabetes, who made available to us a Microsoft Office Access database containing all the information relating to a patient (personal data , diagnostics, treatments, biochemical tests, ...)

Table 1. Structure of database

Table	Attributes	Description
Identity	File number	Primary key, patient record number
	Last name	Patient's name
	First name	Patient's first name
	Age	The patient's age
	Address	His address
	Civil status	The patient's family status
	Profession	The patient's job
	Phone	Patient's phone number
Followed	Followed N	Identify the patient during a visit
	Date conslt	The date of consultation
	TA (Min) D	Minimum blood pressure (right arm)
	TA (Max) D	Maximum blood pressure (right arm)
	TA (Min) G	Minimum blood pressure (left arm)
	TA (Max) G	Maximum blood pressure (left arm)
	Weight	Weight of patient
	Diagnostic	Hepatitis or not
Lab Exam	Exam N	Exam number
	Name Exam	The name of the exam
	UC	Examination result
Prescription	Ref	Drug reference
	Drug code	Drug name
	Dose / Day	The dosage of the appropriate drug per day
	Frequency	The frequency of intake

First, we only implemented the indicators corresponding to the data from the "Prescription" table, and we will gradually add the indicators from other tables.
After communicating with our experts, the tables we have selected are presented below in table 1.
We have tried to reflect both the indicators directly related to our model and which has an important role in the mechanism of predicting the right drug adapted to each patient.
To reduce the complexity of modeling our process, it is important to reduce the information flow of medical record data. This process filtered the data from the medical records, only for important functions to the doctors.

4.2 Preprocessing

Outliers generally affect information data scaling and regression fits. To detect outliers, we implemented the linear regression method, as part of a problem of predicting drugs suitable for the right patient.

4.3 Data reduction and feature extraction

In our data, we found that the values of minimum and maximum blood pressure of the left and right arms, are strongly correlated in most patients. In order to deal with these correlated predictors, we are based on an approach, which consists of removing a minimum number of predictors with highest

pairwise correlations and ensuring that all pairwise correlations are under a specific level[9].
The basic idea is to remove strongly correlated predictors iteratively, with the following steps:

- Calculate the correlation matrix of the predictors.
- Determine which pair of predictors has the greatest absolute pairwise correlation.
- Compute the average between the first predictor and the other predictors.
- Eliminate the predictors if have an average correlation greater than other predictor.
- Repeat all steps until correlations are under the specific level.

4.4 Learning

The objective of this achievement is the similarity calculation to predict the class in a relevant category. Indeed, we want to build a model based on learning values (name of categories) and use it to classify new data.
The Methodology applied in our model is:

- Use the results of the preprocessing step as input for our system.
- Fit our model with the training part of our dataset.
- Testing our model with the second part of our dataset.
- Calculate and analyze the performance of our model .

In our model we chose to use the multivariate linear regression algorithm to make predictions, a target variable "drug code" predicted using several predictive variables (age, balance sheet result, TA (min), TA (max), diagnostic, …). In our model, drug code is the result of the correlation of several predictor variables.

4.5 Experimentation and analysis of results

To assess the relevance of our model, we chose to use the multivariate linear regression algorithm, a target variable "drug code" predicted using several predictor variables.
The experiment is performed using a training data set consisting of 1000 patients with 13 different attributes.
The dataset is divided into two parts: 70% of the data is used for training and 30% for testing.
Our main objective is to predict and explain the occurrence of the drug code (variable to be predicted) from the characteristics of the people (Age, Exam values, TA (min), TA (max), Diagnosis, Etc.) who present the explanatory variables.

Table 2. Evaluation Metrics

Evaluation parameter	Result
Squared Correlation	0.879
Adjusted Coefficient of Determination	0.773
Mean Squared Error (MSE)	0.342
Root Mean Squared Error (RMSE)	0.364

The performance of linear model depends on the margin with which it separates the data making it the most efficient machine learning algorithm for prediction based on several factors. For a good predictive model, the RMSE values should be low (<0.5).
The value of Mean root squared error RMSE obtained as presented in the table 2 is 0.364 (<0.5) reflects the strong ability of the model to make predictions accurately, so the model has a value of Correlation Coefficient R2 high 0.879 (≥0.7).
Indeed, we can say that the quality of fit between the model built and the observed data is good.
There are different metrics to evaluate this prediction model. Most are based on the confusion matrix that intersects the actual class of individuals in the learning game with the class predicted by the model
In order to estimate the performance of our classification model, we focus on the evaluation of learning based on three magnitudes, these are "Accuracy", "Precision" and "Recall".

- Accuracy: is the sum of the correct results (all classes combined) divided by the sum of all the results for all the classes. [11]
- Precision: is the number of correct results, from one of the classes, divided by the sum of the results (correct and incorrect) from that same class.
- Recall: this is the number of patients correctly associated with this drug divided by the total number of patients who should actually be prescribed by this drug.

From the table results (table 3), we notice that the accuracy values remain within 50% to 100%.
The precision in this case is generally high and has even reached 100% for drug codes such as “ARGINO”, “INSULTARD”, “LIPANOR” and “TRIATEC”.
In this case we can say that these drugs correspond to the right patients. The algorithm was unable to assign any drugs to poorly represented drug codes such as "SECTRAL".

Table 3. Sample extracted from confusion matrix

	AMAREL	**ARGINO**	**SECTRAL**	**INSULTAD**	**LYSANXIA**	**LIPANOR**	**TRIATEC**	**…**	**Class Precision**
AMAREL	132	0	0	0	0	0	0	…	**96**
ARGINO	0	148	0	0	0	0	0	…	**100**
SECTRAL	0	0	0	0	0	0	0	…	**0**
INSULTAD	0	0	0	60	0	0	0	…	**100**
LYSANXIA	20	0	0	0	20	0	0	…	**33**
LIPANOR	0	0	0	0	0	40	0	…	**100**
TRIATEC	0	0	0	0	0	0	50	…	**100**
Class recall	**84**	**92**	**0**	**50**	**33**	**80**	**100**	…	

5 Conclusion

In this paper, we described the step of partitioning our patients into groups of similar characteristics, we used a powerful approach in the field of partitional clustering which is the k-medoid algorithm.

After having created families of similar behaviour from the characteristics extracted from our learning base, we presented the linear regression method within the framework of a drug code prediction problem adapted to the right patient by explanatory variables selected by our expert.

The results obtained show that the proposed approach can produce good predictions, with an accuracy value of 82.68%.

Nevertheless, the results of this work constitute the bases of a work to be continued and improved for a much more in-depth study. First, we want to extend our experiments to several themes to show the real dependence between the descriptors and the theme in which they are used, and to apply the approach to other types of noisy data in order to confirm the relevance of our approach.

Indeed, it would be interesting to look into the study of the correlations between hypertension disease and risk factors for diabetes using data mining techniques.

"Diabetes Increases Risk for Hypertension, and Vice Versa [16]." We want to focus on changing the lifestyle which plays a vital role in combating these two diseases and adopting joint intervention of the two diseases into the daily routine, so that these diseases can be brought under control.

The intervention includes the risk factors that need to be treated such as diet, weight, smoking cessation and physical activity so that these diseases can be controlled.

References:

[1] J. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches," *Medical care*, vol. 48, pp. S106-S113, June 2010.

[2] F. Wang, J. Hu, and J. Sun, "Medical prognosis based on patient similarity and expert feedback," in *Proc. 21st Int. Conf. Pattern Recognition (ICPR2012)*, 2012, pp. 1799-1802.

[3] J. Sun, F. Wang, J. Hu, and S. Edabollahi, "Supervised patient similarity measure of heterogeneous patient records," *{SIGKDD} Explorations*, vol. 14, no. 1, pp. 16-24, 2012.

[4] G. Salton and M. Lesk, "Computer Evaluation of Indexing and Text Processing," *J. {ACM},* vol. 15, no. 1, pp. 8-36, 1968.

[5] R. M. Hayes, "*The SMART retrieval system; experiments in automatic document processing*," vol. 9, p. 199, 1973.

[6] G. Jahoda, "*Automatic information organization and retrieval*," vol. 5, p. 230, 1970.

[7] C. D. Manning, P. Raghavan, and Schtze, Hinrich, *Introduction to information retrieval*. Cambridge University Press, 2008.

[8] M. Kwak, G. Leroy, J. D. Martinez, and J. Harwell, "Development and evaluation of a biomedical search engine using a predicate-based vector space model," *Journal of biomedical*

[9] K. Golub, "Automated subject classification of textual web documents," *J. Documentation*, vol. 62, no. 3, pp. 350-371, 2006.

[10] L. Xie, G. Li, M. Xiao, and L. Peng, "Novel classification method for remote sensing images based on information entropy discretization algorithm and vector space model," *Comput. Geosci*., vol. 89, pp. 252-259, 2016.

[11] L. Jing, M. K. Ng, and J. Z. Huang, "Knowledge-based vector space model for text clustering," *Knowl. Inf. Sys*t., vol. 25, no. 1, pp. 35-55, 2010.
[12] I. N. Sarkar, "A vector space model approach to identify genetically related diseases," *J. Am. Medical Informatics Assoc*., vol. 19, no. 2, pp. 249-254, 2012.
[13] J. Lee, D. M. Maslove, and J. A. Dubin, "Personalized mortality prediction driven by electronic medical data and a patient similarity metric," *PloS one*, vol. 10, no. 5, p. e0127428, 2015.
[14] B. Campillo-Gimenez, N. Garcelon, P. Jarno, J. M. Chapplain, and M. Cuggia, "Full-text automated detection of surgical site infections *secondary to neurosurgery* in Rennes, France," 2012.
[15] D. Cao and B. Yang, "*An improved k-medoids clustering algorithm*," ed, 2010.
[16] V. Tsimihodimos, C. Gonzalez-Villalpando, J. B. Meigs, and E. Ferrannini, "Hypertension and Diabetes Mellitus: Coprediction and Time Trajectories," *Hypertension* (Dallas, Tex. : 1979), vol. 71, pp. 422-428, March 2018.