Case-based retrieval of similar diabetic patients

Damien Zufferey, Stefano Bromuri and Michael Schumacher
Institute of Information Systems, University of Applied Sciences Western Switzerland
Email: damien.zufferey@hevs.ch

***Abstract*—Patients suffering from diabetes often develop several comorbidities such as hypertension and dyslipidemia. The presence of the comorbidities leads to more complex patient profiles associated with specific patient treatments. In this paper we present a novel algorithm to help physicians, given a new case, in retrieving similar past patient cases. This novel algorithm is based on the bag-of-words (BoW) model to encode as features, the occurrence of each pre-computed cluster, for each patient, according to the approach of document classification. We then evaluate the algorithm on a real de-identified dataset of 3201 diabetic patients, demonstrating the advantage of our approach.**

Keywords*—*diabetes, comorbidities, health records, case-based retrieval, clustering.

I. Introduction

Patients having diabetes usually develop several comorbidities such as hypertension and dyslipidemia. The presence of these comorbidities often generates more complex patient profiles associated with specific patient treatments. Physicians working with these cases are well trained and propose treatment plans based on their past experiences. The problem with this procedure is that physicians do not have time to review several thousands of past cases. In addition, medical records are composed of different types of measured values such as laboratory tests or physical examinations. According to the multivariate aspect of the dataset, it is usually difficult for a human to measure similarity between patients.

This work presents an algorithm to help physicians, given a new case, in retrieving similar past patient cases. Having to deal with heterogeneous and partially incomplete medical data is a challenge from the perspective of information retrieval. One goal is to find a way to extract features, which characterize in a compact form each patient in the dataset. Another goal is to find metric in order to measure the similarity between patients and to rank them. According to this definition, the idea behind the bag-of-words (BoW) model and a clustering algorithm for constructing the codebook has been identified here as a manner to encode as features, the occurrence of each centroid (codeword), for each patient. BoW has been previously applied in document classification [1] and computer vision [2]. As clustering tools, we explore here the use of k-means and self-organizing map (SOM). The k-means algorithm is a centroid-based clustering approach to partition n observations into k clusters by minimizing the within-cluster sum of squares (WCSS). The k-means algorithm has been applied in various fields, such as geostatistics [3] and biomedical imaging [4]. SOM is a sort of artificial neural network (ANN) that is commonly used for multidimensional scaling. SOM is trained using an unsupervised learning approach in order to produce a discretized representation of the input space, called a map. In [5], SOM are described as a nonlinear generalization of Principal components analysis (PCA). Applications of SOM are found in geophysics [6] and climatology [7].

To test our approach we used a real de-identified set of health records from the Portavita company[1], where we devise an information retrieval algorithm that is capable to retrieve similar patient cases, given a new case. These datasets provide medical records, such as laboratory tests and physical examinations, from different profiles of diabetic type 2 patients. The novelty and significance of the proposed approach resides in the definition of a case-based retrieval (CBR) algorithm that allows to deal with heterogeneous medical data in an effective way, relating similar cases between them to help physicians in defining treatment plan and medication.

The rest of this paper is organized as follows: Section II presents a background on BoW, k-means and SOM; Section III presents our algorithm for relating similar patient cases; Section IV evaluates our approach; Section V puts our work in comparison with the state of the art; finally, Section VI concludes this paper and draws the lines for future works.

II. Background

A. Bag-of-words model

The BoW model is an approach to represent information in a simplified manner. This approach was originally used for text document classification [1], where the occurrence of each word in a document composes the feature vector. An adapted version of the BoW model, called bag of visual words, was recently applied in the field of computer vision for image classification [2]. First, a set of features, which are able to handle characteristics such as image intensity or image rotation, are extracted from images. Then, a codebook is computed from the extracted features. The creation of the codebook relies on vector quantization mechanisms such as clustering algorithms. Finally, the image is characterized by a histogram representing the frequency of each feature according to a codebook.

B. K-means

The k-means is an unsupervised algorithm to partition n observations into k clusters by attributing each observation to the cluster with the nearest mean.

C. Self-organizing map

The self-organizing map (SOM) [9] is a type of artificial neural network (ANN) that is trained using an unsupervised procedure. The goal is to find a mapping from high-

[1] http://www.portavita.eu

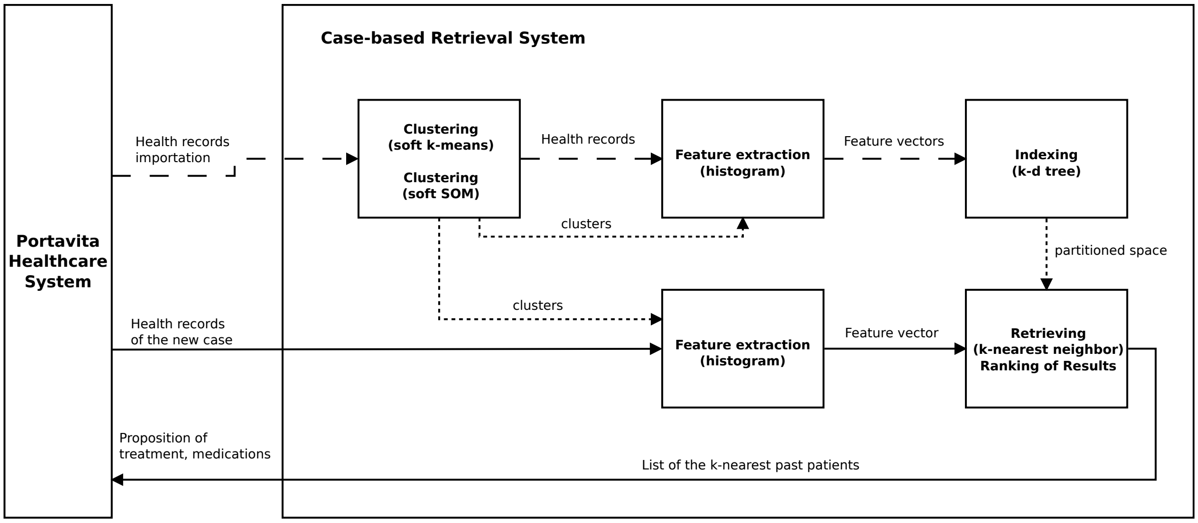

Fig. 1: The architecture of the case-based retrieval system.

dimensional input space to a discretized low-dimensional representation. This vector quantization mechanism offers some advantages over traditional clustering approaches like k-means, in particular the ability to model nonlinearity in the data.

III. Model

For the implementation, a variation of the bag-of-words model [2], which is popular in information retrieval, was used. We have chosen such an approach, because it deals well with time series that are not uniformly sampled and present different length as we currently have for our patients in the Portavita system. The architecture of our case-based retrieval (CBR) system is divided into 2 processes. The first process consists of creating a knowledge base of past experiences based on all health records available in the production system of Portavita. This is an off-line procedure, which has to be run periodically (e.g. once per week) in order to update the knowledge base with new validated cases. This process is represented with large-dashed lines in fig. 1, below each step of the process is explained.

Clustering

After having selected the set of fields to import from the production system, the clustering module is responsible to separate the data into several clusters. For that process, we use two different algorithms. The k-means clustering algorithm [8] for its ability to minimize distortion and the SOM algorithm [9] for its ability to model nonlinearity in the data.

Feature extraction

At this step, a feature vector, which characterizes each patient, is computed. The solution, which has been implemented, consists of an histogram of all laboratory tests and physical examinations related to a patient. The calculation is based on Euclidean distance between the center of each cluster and the health records values. The feature vectors will contain normalized coefficients explaining how a vector of values representing a patient is assigned to each cluster. In particular, we use a soft assignment strategy where a patient's vector can be assigned to more than one cluster according to the proximity of the vector to the cluster centroid. See figure 2. More formally, a feature vector $\mathbf{f}$, for a patient, is computed as following:

Given L the set of laboratory tests belonging to the patient,
P the set of physical examinations belonging to the patient,
C the set of centroids for laboratory tests and
D the set of centroids for physical examinations,

$\mathbf{f} = \mathbf{g} \bullet \mathbf{h}$, where $\bullet$ means concatenation of vectors,

$$\mathbf{g}\colon \forall\, 1 \leq i \leq |C|,\ g_i = \frac{\sum_{\mathbf{x_j} \in L} dist(\mathbf{x_j}, \mathbf{C_i})}{\sum_{\forall i} \sum_{\mathbf{x_j} \in L} dist(\mathbf{x_j}, \mathbf{C_i})},$$

where $dist$ is the euclidean distance between two vectors,
$\mathbf{g}$ is the feature vector for the lab tests.

$$\mathbf{h}\colon \forall\, 1 \leq i \leq |D|,\ h_i = \frac{\sum_{\mathbf{x_j} \in P} dist(\mathbf{x_j}, \mathbf{D_i})}{\sum_{\forall i} \sum_{\mathbf{x_j} \in P} dist(\mathbf{x_j}, \mathbf{D_i})},$$

where $dist$ is the euclidean distance between two vectors,
$\mathbf{h}$ is the feature vector for the physical examinations.

As there are two sets of clusters (k-means and SOM), the two computed feature vectors are concatenated into one.

Indexing

At this step, a k-dimensional tree (k-d tree) [10] is built with all extracted feature vectors. The k-d tree is an acceleration structure which partitions a k-dimensional space using a binary tree. The goal is to allow fast multidimensional search (such as nearest neighbor searches).

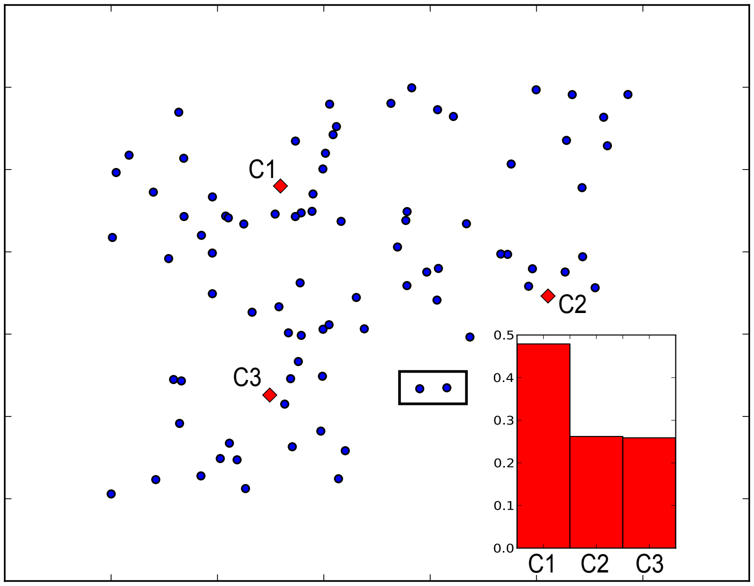

Fig. 2: Feature vector (histogram) for a patient.

Feature Extraction and Retrieval

The second process consists of retrieving similar cases with respect to a new case, based on his health records. This process is represented with continuous lines in fig. 1, below each step of the process is explained. For the health records of the new case, the same process, as explained before, is applied in order to build the feature vector which characterizes the patient. This component has the responsibility to search for the new case the k-nearest neighbors previously indexed in the k-d tree. The search algorithm is based on euclidean distance and the k feature vectors with the smallest distance are selected.

IV. Evaluation

The amount of patients that we considered for our analysis is 3201. We selected this patients by taking into consideration the following constraints:

- the patient must have at least one laboratory test;
- the patient must have at least one physical examination;
- the patient must have a diagnosis for complications (none, hypertension, dyslipidemia, heart problems).

In particular we have chosen hypertension, dyslipidemia and heart problems because they are the most frequent complications for patients affected by diabetes type 1 or 2. Despite these constraints, the dataset presents missing values in the physical examinations and laboratory tests. To mitigate this problem, we decided to substitute missing values by the corresponding mean value of the field. More complex strategies to replace missing values are possible, for example we could have applied a probabilistic generative approach in order to generate missing values. For the moment, we decided to go with the simplest approach, but for the next phases of the project we will consider more sophisticated solutions for this problem. Furthermore, as already mentioned before, our dataset presents laboratory tests and physical examinations that are not uniformly sampled. This does not constitute a problem as we are using the bag-of-words approach, which can deal with this problem implicitly.

From the perspective of the features that we have chosen for our CBR approach, for the physical examinations, the fields that were chosen are: BMI, weight, heart rate, height, waist circumference, diastolic blood pressure, systolic blood pressure; and for the lab tests we have chosen the following values: HDL (cholesterol), LDL (cholesterol), hba1c, albumin/creatinine ratio, glucose at fasting, sodium, potassium, hemoglobin, gammaGT, triglyceride.

For the purposes of the evaluation, the dataset has been divided randomly in two parts. The first half, the training set, has been used only for the clustering step of the process, see figure 1. The second half, the test set, has been used for the feature extraction, the indexing and the retrieval. In our evaluation we select patients affected by either hypertension, dyslipidemia or heart problems and put them in relation with the training set to retrieve patients with similar problems. From the standpoint of proportions of patients affected by a complication, hypertension counts 57% of positive cases in the dataset, dyslipidemia counts 40% of positive cases in the dataset, and heart problems count 27% of positive cases in the dataset.

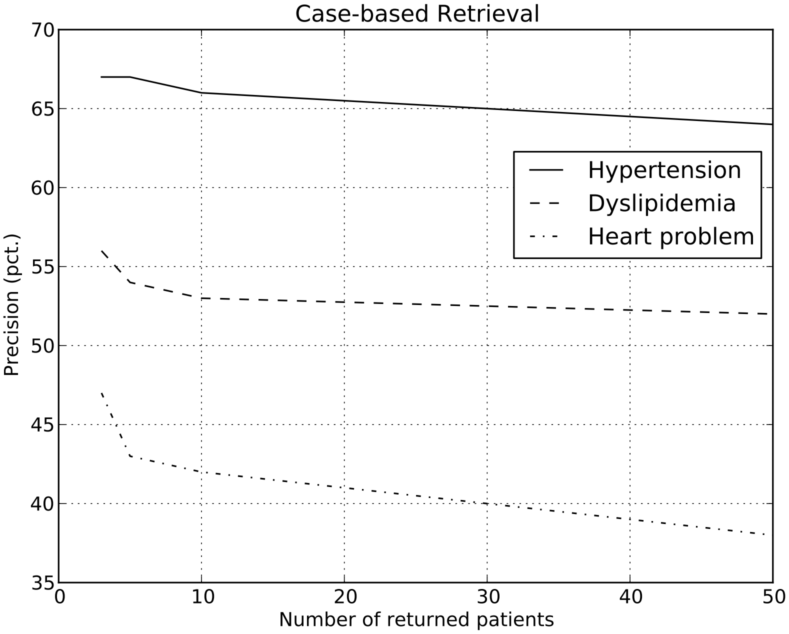

Fig. 3: Precision to Retrieve Patients Affected by Hypertension, Dyslipidemia and Heart problem.

Fig. 3 shows the results for the patients affected by hypertension, dyslipidemia or heart problem. Such curves were computed by averaging the results 100 times. We calculated the precision following the standard formula of information retrieval reported below:

$$\frac{|RelevantDocuments \cap RetrievedDocuments|}{|RetrievedDocuments|}$$

The formula states that the number of positive cases is divided by the number of returned patients. The precision of the retrieval algorithm has been evaluated in 4 different cases: we retrieved 3, 5, 10 and 50 patients to plot the curve. As physicians do not have much time to review similar cases, precisions at 5 and 10 should be more relevant than the precision at 50.

For hypertension, we see that at 3, 5 and 10 the amount of retrieved patients which are related to our query is around 65% - 70%. This results tells us that it is quite easy to retrieve patients affected by hypertension, and that our algorithm can discriminate people affected by hypertension, as the random precision would be on average 57%.

Concerning dyslipidemia, the precision at 3, 5 and 10 in this case is lower, around 50% - 55%, with respect to the hypertension case, probably also because the proportion of patients affected by dyslipidemia is only 40% in the dataset. Still, if we consider the random case, the precision would only be 40%.

Finally, the results for a query where we try to relate a patient affected by heart conditions with other patients with the same complication. In this case the proportion of patients affected by heart problems in the dataset is only 27%. In this case, the precision at 3 is around 47%, but then it drops to around 43% for the precision at 5 and 10. This happens due to the fact that there are less patients in the database with this condition, but also because heart conditions involve a large spectrum of issues, for which we do not propose a distinction for the moment. Overall, the results of the algorithm are encouraging as doctors can effectively find relevant cases at 3, 5 and 10, for all of the three complications considered.

V. Related works

The problem of retrieving similar patients from a database of physiological data is not a new problem. Sun et al. in [11] propose a dissimilarity measure that takes into consideration the experience of the specialist as well as the physiological data of the patient. The difference between us and Sun et al. is that we do not define any particular distance for dissimilarity, we rather use Euclidean distance. A difference with [11] is that we use kd-trees to accelerate our KNN approach. The approach of Sun et al. does not take into consideration the dynamics of the patient health records, while in our CBR system, thanks to the visual words approach, we can consider also the patient evolution in time.

The Inreca project [12] focuses on case-based reasoning for medical purposes. Such a project makes use of kd-trees to store the different cases in order to speed up the retrieval of the patients, as we do, but one difference between the two approaches is that Inreca does not apply a visual words approach like we do. The advantage of our approach is that we can take into consideration the dynamics of the patients involved in the CBR, and that we can have different length of treatment, but still being able to perform a comparison amongst the patients. Furthermore, our approach allows us to query the CBR system with an arbitrary number of patient records, which is a problem that is not considered in Inreca.

In [13], Lin and Li present an approach to analyze time series by using a bag of words approach like we do in this paper. The difference between the work in this paper and [13] is that, rather than applying a K-means algorithm and then the bag of words approach with soft-assignment, Lin and Li apply Symbolic Aggregate Approximation (SAX) [13]. In our case, the time series of the Portavita dataset are extremely discrete, consequently a signal processing approach like SAX is not suitable for our purposes. Further investigation is though needed to see if the k-means based approach can be substituted with SAX.

In [14], a method based on modified multivariate bag-of-words and SAX is presented to classify physiological data. The main difference with the standard approach to the bag-of-words problem is that in [14] multivariate time series are considered. With respect to [14], our time series are multi-variate, but they are very discrete in time, so an approach using SAX would not produce enough symbols to discriminate amongst the different patients.

In [15] a dynamic time warping (DTW) distance is presented to retrieve cases of similar patients affected by hepatitis B or C. Such an approach is based on comparing the shape of the curves described by the evolution of the physiological values of the patients. In [15], DTW achieves high accuracy rates (88% if 20% of the patients are retrieved). The DTW approach achieves better results than the bag-of-words but it assumes that the patients have all the same temporal granularity in collecting their lab tests which is almost never the case in real datasets, furthermore the approach presented in [15] considers only a dataset of 102 patients, while we consider a dataset of about 3000 diabetic patients.

VI. Conclusion

In this paper we presented a novel algorithm based on the BoW model in order to rank similar diabetic patients. The obtained results are encouraging as doctors can effectively find relevant cases for all of the three complications considered. The use of the BoW model allows us to deal in an elegant manner with the presence of not uniformly sampled time series. Future work comprises the introduction of case-based retrieval that discriminate amongst patients with diabetes type 1 or 2 and patients that are affected by one or more co-morbidities of diabetes, such as micro-vascular or macro-vascular complications of diabetes. Finally, we are also considering the introduction of a relevance feedback feature in order to give the possibility for the doctor to indicate the pertinence of the returned results, and to use this information about the pertinence for the future queries.

Acknowledgment

This work was partially supported by the FP7 287841 COMMODITY12 project.

Thanks to the dutch company Portavita B.V. for the possibility to work on a real de-identified dataset of diabetic cases.

References

[1] Y. Ko, "A study of term weighting schemes using class information for text classification," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '12. New York, NY, USA: ACM, 2012, pp. 1029–1030. [Online]. Available: http://doi.acm.org/10.1145/2348283.2348453

[2] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 591 –606, april 2009.

[3] M. Honarkhah and J. Caers, "Stochastic simulation of patterns using distance-based pattern modeling," *Mathematical Geosciences*, vol. 42, pp. 487–517, 2010. [Online]. Available: http://dx.doi.org/10.1007/s11004-010-9276-7

[4] H. Ng, S. Ong, K. Foong, P. Goh, and W. Nowinski, "Medical image segmentation using k-means clustering and improved watershed algorithm," in *Image Analysis and Interpretation, 2006 IEEE Southwest Symposium on*, 0-0 2006, pp. 61 –65.

[5] H. Yin, "Learning nonlinear principal manifolds by self-organising maps," in *Principal Manifolds for Data Visualization and Dimension Reduction*, ser. Lecture Notes in Computational Science and Enginee, A. Gorban, B. Kégl, D. Wunsch, and A. Zinovyev, Eds. Springer Berlin Heidelberg, 2008, vol. 58, pp. 68–95. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-73750-6_3

[6] Y. Liu and R. Weisberg, "Patterns of ocean current variability on the west florida shelf using the self-organizing map," *Journal of Geophysical Research-Oceans*, vol. 110, no. C6, 2005.

[7] B. Hewitson and R. Crane, "Self-organizing maps: applications to synoptic climatology," *Climate Research*, vol. 22, no. 1, pp. 13–26, 2002.

[8] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Inf. Theor.*, vol. 28, no. 2, pp. 129–137, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1109/TIT.1982.1056489

[9] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59–69, 1982. [Online]. Available: http://dx.doi.org/10.1007/BF00337288

[10] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Transactions on Mathematics Software*, vol. 3, no. 3, pp. 209–226, September 1977.

[11] J. Sun, D. Sow, J. Hu, and S. Ebadollahi, "Localized supervised metric learning on temporal physiological data," in *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ser. ICPR '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 4149–4152. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2010.1009

[12] R. Bergmann, "Highlights of the european inreca projects," in *Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development*, ser. ICCBR '01. London, UK, UK: Springer-Verlag, 2001, pp. 1–15. [Online]. Available: http://dl.acm.org/citation.cfm?id=646268.683867

[13] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, ser. DMKD '03. New York, NY, USA: ACM, 2003, pp. 2–11. [Online]. Available: http://doi.acm.org/10.1145/882082.882086

[14] P. Ordonez, T. Armstrong, T. Oates, and J. Fackler, "Using modified multivariate bag-of-words models to classify physiological data," *2010 IEEE International Conference on Data Mining Workshops*, vol. 0, pp. 534–539, 2011.

[15] S. Tsevas and D. Iakovidis, "Fusion of multimodal temporal clinical data for the retrieval of similar patient cases," in *Biomedical Engineering, 2011 10th International Workshop on*, oct. 2011, pp. 1 –4.