A Comorbidity-based Recommendation Engine for Disease Prediction

Francesco Folino and Clara Pizzuti
Institute for High Performance Computing and Networking (ICAR)
Italian National Research Council (CNR)
Via Pietro Bucci, 41C
87036 Rende (CS), Italy
ffolino, pizzuti@icar.cnr.it

Abstract

A recommendation engine for disease prediction that combines clustering and association analysis techniques is proposed. The system produces local prediction models, specialized on subgroups of similar patients by using the past patient medical history, to determine the set of possible illnesses an individual could develop. Each model is generated by using the set of frequent diseases that contemporarily appear in the same patient. The illnesses a patient could likely be affected in the future are obtained by considering the items induced by high confidence rules generated by the frequent diseases. Experimental results show that the proposed approach is a feasible way to diagnose diseases.

1. Introduction

In the last few years we are witnessing to an increasing interest in the application of computational science methods to health care information and management systems. The utilization of information technologies that could significantly improve efficiency and effectiveness of health care strategies are very important because of the implications they could have in every day life of individuals.

An emerging viewpoint aims at identifying prospective health care models to determine the risk for individuals to develop specific diseases [17]. In fact, prevention or intervention at the disease's earliest onsets allow advantages for both the patient, in terms of life quality, and the medicare system, in terms of costs. However, recognizing the origin of an illness is not an easy task because it can be generated by multiple causes. Physicians prescribe laboratory tests only after the appearance of patient's complains, and use family and health history to assess the hypothesized problem. The approach is thus reactive, i.e. a medical treatment is undertaken only after the patient has already developed the disease, rather than proactive.

Hospitals and physicians, however, collect thousands of patient clinical histories that include valuable information regarding illness correlations and development. The patient medical records contain important enlightenment regarding the co-occurrences of diseases affecting the same individual. A comorbidity relationship between two illnesses exists whenever they appear simultaneously in a patient more than chance alone [11]. Although comorbidity is very common in the population and its extension increases with age, few investigations have been conducted on patient's comorbid conditions [5]. The comorbidity relationships between diseases, however, could be exploited to build a model that predicts the diseases a patient could have in the future.

Advanced risk assessment tools are currently at disposal, mainly based on statistical techniques [7, 8]. Another approach for addressing the problem, which is gaining increasing interest, is the use of methodologies coming from the fields of knowledge discovery [19].

Among the most recent proposals coming from this research field, Davis et al. [6, 3] have been the first that used patient clinical history for disease prediction. They built a collaborative assessment and recommendation engine, based on the ICD-9-CM codes, to predict future diseases. The engine relies on the collaborative filtering methodology [16] used for producing recommendations to people by collecting preferences from users having similar behaviors. A patient is characterized by a vector of diagnosed diseases and a prediction is made on the base of other similar patients. The similarity function adopted includes the inverse frequency of diseases to reduce the weights of very common sicknesses. In order to apply the collaborative filtering technique, the training set of patients is reduced by removing all those patients having one or no disease in common with the active patient.

Steinhaeuser and Chawla [18] used a hybrid technique based on collaborative filtering and nearest neighbor classification. The similarity between two patients is computed with the *Jaccard coefficient* [1, 12], which is the normaliza-

978-1-4244-9166-7/10/ $26.00

tion of common diseases that two patients have, with respect to their union. Given a patient, the k most similar patients are selected to make a prediction. They found that almost the 42% of diseases were predicted as expected. A disease network is also built and their structural properties studied.

In this paper we propose a recommendation engine for disease prediction that combines clustering and association analysis techniques. The system, named *CORE* (*COmorbidity-based Recommendation Engine*), extends the approach proposed in [9] by introducing a clustering phase on the data set of patient records that allows the generation of local, more specialized and accurate prediction models, instead of a general, global model. *CORE* uses the past patient medical history for generating models able to determine the risk of individuals to develop future diseases. Analogously to Davis et al. [3], a patient is represented by the set of ICD-9-CM codes of diagnosed diseases, and a disease is predicted by comparing a patient with individuals having a similar clinical history. However, differently from their approach, we use association analysis [19] to generate a disease predictive model composed by more models, each specific to a particular patient profile. The model is built by using the set of frequent diseases that contemporarily appear in the same patient. The diseases the patient could likely be affected in the future are obtained by considering the items induced by high confidence rules generated by recurring disease patterns. The medical record of a patient is then compared with the patterns discovered by the model, and a set of illnesses is predicted. The approach is similar to that used in recommendation systems from web usage data, where given the pages visited by a user during a web session, a recommendation value for the next page the user will probably visit is computed on the base of behavioral profiles induced on groups of users sharing similar navigational habits [14, 15]. Experimental results show that the approach is a promising method to predict individual diseases by taking into account only the illnesses a patient had in the past, and that the specialization of the models on group of similar patients increases the prediction accuracy.

The paper is organized as follows. The next section briefly describes the data set used. In section 3 the *CORE* system is described. Section 4, finally, reports the evaluation of the proposed approach on a data set of patient medical records.

2. Data description

The data set consists of medical records of 1462 patients of a small town in the south of Italy. Each record contains a unique patient identifiers, date of birth, the gender, and the list of disease codes with the date of the visit in which that disease has been diagnosed. The disease codes are those defined by the International Classification of Diseases, Ninth

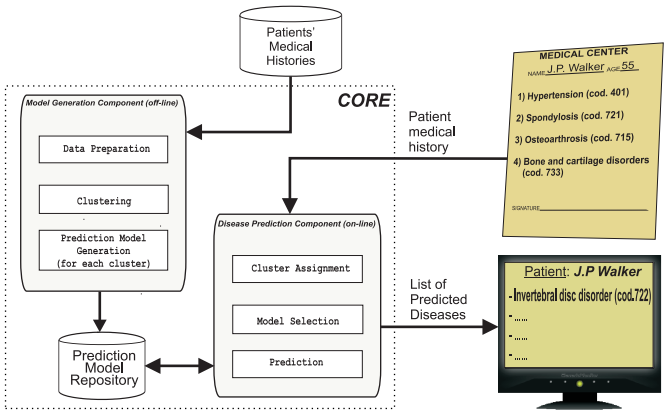


Figure 1: An overview of the *CORE* system.

Revision, Clinical Modification (ICD-9-CM). The International Classification of Diseases (ICD) and Related Health Problems supplies codes to classify diseases and a wide variety of signs. Every health condition is associated with a unique category and given a code, up to five digits long. The first three digits constitute the principal diagnosis, while the other two identify secondary diagnoses. The ICD is published by the World Health Organization and used worldwide for morbidity and mortality statistics, reimbursement systems and automated decision support in medicine. The data is completely anonymized, thus there is no way to identify the patients. In our database the number of diagnoses are 8768 spanning from 1990 to 2009. From an analysis of the patient records, we found that the raw data contained some disease not informative for our study. These diagnoses have thus been eliminated. Some patients had no or only one diagnosis. These patients have been discarded because not useful. After this preprocessing phase, the database reduced to 1105 patients and the number of diseases was 972. However, the number of diseases was still too high. As described above, the first three digits of a code denote the general diagnosis. Even if some details can be missed, these three digits are sufficiently informative to study the disease correlations. Thus, the five digits ICD-9-CM codes have been collapsed to these first three digits, in such a way the number of diseases was reduced to 330.

In the next section we first give an overview of the system architecture proposed to perform disease prediction, then a description of each embedded module is reported.

3 A Framework for Disease Prediction

We first give an overview of the system, then a description of each module is reported. The *CORE* prediction system, as depicted in Figure 1, consists of two main components: an off-line component for the *model generation*, and an on-line component for the *disease prediction*. The

model generation component involves a preprocessing step to transform the raw data to a transactional set of patient records constituted by a sequence of ICD-9-CM codes. i.e. the list of diseases a patient had. Then, clustering is performed to group patients on the base of the diseases they share, and a representative is generated for each cluster found. The representative computation is a very important step since it is used by the prediction module to decide the right model to apply for foreseeing next diseases of the current patient. After that, frequent patterns for each cluster are computed and a model is generated for each group. Pertaining the on-line component, it first assigns a patient to a cluster by matching him against each cluster representative, then the prediction model associated with this cluster is selected, and finally the next expected diseases are released by applying this model.

It is worth to notice that the system in Figure 1 is a general predictive architecture, parametric with respect to both the clustering algorithm and the prediction model used. Therefore, by suitably customizing the algorithms for the kind of data at disposal, this architecture could be profitably exploited also in different scenarios. In the following a detailed description of each single component in the architecture is provided.

3.1. Model Generation

Data Preparation. Let m be the number of patients contained in the original data set of patient histories. The data preparation module transforms the original data set into a new data set $T = \{t_1, \ldots, t_m\}$, where each t_i is a patient medical record of variable size constituted by a sequence of ICD-9-CM disease codes. Thus T summarizes the medical histories relative to all the m patients.

Clustering. The main motivation for grouping the set T of patients sharing most of their disease history, it that it is an effective way of improving the accuracy of the prediction model, as experimental results will show. Performing clustering is not an easy task at all, since its performances are tightly related to the kind of method used. For the purposes of this paper, we decided to exploit a straightforward variant of the traditional *k-means* clustering [10, 13] able to deal with categorical tuples of variable size, like those present in the dataset T.

For a given parameter k, this algorithm partitions T into k clusters $C = \{C_1, \ldots, C_k\}$ in a way that high intra-cluster similarity and low inter-cluster similarity are guaranteed. C is a partitioning of T, i.e., $\bigcap_{i=1..k} C_i = \emptyset$ and $\bigcup_{i=1..k} C_i = T$. Each record $t_i \in T$ is assigned to a cluster C_j according to its distance $d(t_i, r_j)$ from a vector r_j that represents the cluster at hand, and is called the *representative* of the cluster. Formally, the clustering algorithm finds a partition C such that:

1. for each C_i the representative r_i is computed
2. $t_i \in C_j$ iff $d(t_i, r_j) < d(t_i, r_l)$ for $1 \leq l \leq k$, $j \neq l$
3. C minimizes the cost function $Q_k = \sum_{i=1}^{k} \sum_{t_j \in C_i} d(t_j, r_i)$

Essentially, the algorithm works as follows. Firstly, k records are selected from T randomly. They represent the initial cluster centers, and each other $t_i \in T$ is assigned to a cluster on the base of condition 2). Then, the algorithm updates the representative of each cluster and re-assigns each record consequently. The iterations terminate when the representatives do not change any more, i.e., the condition 3) holds.

It is worth to note that the schema above is parametric w.r.t. the definitions of distance d and representative r. Since in our scenario we deal with categorical data, we used a kind of distance that proved to work very well in this setting: the *Jaccard* distance. This measure is derived by the *Jaccard coefficient* [1, 12] which is based on the idea that the similarity between two itemsets is directly proportional to the number of their common items and inversely proportional to the number of different ones. Therefore, given two records t_i and $t_j \in T$, the Jaccard distance can be defined as:

$$d(t_i, t_j) = 1 - \frac{|t_i \cap t_j|}{|t_i \cup t_j|}$$

The next step pertains a suitable definition for the cluster representative. Intuitively, the representative should model the content of the cluster in order to make trivial the interpretation of the cluster itself. Among various possibilities, an easy and effective way for building the representative consists in using the frequent items belonging to the cluster. The frequency degree can be controlled by introducing a user-defined threshold value γ representing the minimum percentage of occurrences an item must have for being inserted into the cluster representative. More formally, given $T_{C_i} = \{t_1, \ldots, t_q\}$ the set of records belonging to the cluster C_i, $D_{C_i} = \bigcup_i t_i = \{d_1, \ldots, d_p\}$ the set of items of C_i, i.e. the disease codes, and $\gamma \in [0, 1]$, then the representative r_{C_i} for C_i can be computed as follows:

$$r_{C_i} = \{d \in D_{C_i} | f(d, T_{C_i})/q \geq \gamma\} \quad (1)$$

where $f(d, T_{C_i}) = |\{t_i \in T_{C_i} | d \in t_i\}|$ is the number of medical records of cluster C_i in which d appears.

Clearly, the clustering algorithm assumes that the number of clusters k has to be fixed at the beginning. Thus, another open issue is how to set k in order to obtain the best partitioning. Ideally, the best partitioning is achieved for the value k^* in correspondence of which the cost function Q_k has its global minimum. However, finding k^* could be unfeasible in practice. Therefore, we pragmatically

recurred to a sub-optimal solution: we iterated the clustering algorithm by ranging k in $[1,|T|]$ until the first, local minimum for Q_k is reached.

Prediction Model Generation. A disease prediction model *DPM* for the dataset T can be defined as a couple $DPM = \langle C, M \rangle$, where $C = \{C_1, \ldots C_k\}$ is a clustering of T and $M = \{M_1, \ldots M_k\}$, where each M_i is the prediction model built on top of C_i.

In order to build a prediction model for each cluster C_i, we follow the same approach introduced in [9], that exploits *association analysis* [19] for inducing a pattern-based prediction schema able to generate predictions about the diseases a patient can incur in the future, given the past history of his health conditions. The main difference with the previous approach is that, in this paper, we deal with "local" models instead of a unique, global model built on the whole dataset. Intuitively, and as validated by experimental results, local models tend to produce better prediction accuracy because the predictions are generated by considering the most similar individuals of the patient under examination.

Employing association analysis for prediction purposes is not new in the data mining literature. It relies on the concept of *frequent itemsets* to extract strong correlations among the items constituting the data set to study. Originally, association analysis has been applied to market basket data, where each item represents the purchase done by a customer. However, it can be easily transposed into the medical context by associating an item with a disease, and by considering an itemset as the set of diseases a patient had along his life until the present. For extracting patterns, we apply the well-known *Apriori* algorithm [2] that efficiently searches for frequent itemsets by cutting the exponential search space of candidate itemsets. The concept of frequency is formalized through the concept of *support*.

Given a set $I_{C_i} = \{i_1, \ldots, i_l\}$ of frequent itemsets induced on a cluster $C_i = \{t_1, \ldots, t_p\}$, the support of an itemset $i_j \in I_{C_i}$, $\sigma(i_j)$, is defined as:

$$\sigma(i_j) = \frac{\mid \{t_i \mid i_j \subseteq t_i, t_i \in C_i\} \mid}{\mid C_i \mid}$$

where $\mid . \mid$ denotes the number of elements in a certain set. The support, thus, determines how often a group of diseases appear together. It is a very important measure because very low support discriminates those groups of items occurring only by chance. Thus a frequent itemset, to be considered interesting, must have a support greater than a fixed threshold value *minsup*.

An association rule is an implication expression of the form $X \Rightarrow Y$, where X and Y are disjoint itemsets. The importance of an association rule is measured by both its $support$ and $confidence$ values. The support of a rule is computed as the support of the $X \cup Y$ and tells how often a rule is applicable. The confidence is defined as $\sigma(X \cup Y)/\sigma(X)$, and determines how frequently items in Y appear in transactions that contain X. Frequent itemsets having a support value above a minimum threshold are used to extract high confidence rules that can be exploited to build a prediction model by matching the medical record of a patient against the patterns discovered by the model.

In our scenario, the support determines how often a group of diseases appears together, while a rule like $X \Rightarrow \{d\}$ (where X is a set of frequent diseases and d is new disease) having a high confidence, allows to reliably infer that d will appear along with the diseases contained in X.

3.2. Disease Prediction

Once the M_i models are built for each discovered cluster C_i, performing the prediction is rather straightforward. The next likely diseases are computed by the *disease prediction* component (see the *CORE* architecture in Figure 1) at the time a new patient arrives. The prediction phase encompasses three main tasks:

- *Cluster Assignment*, where the patient is recognized as member of a cluster by matching him against each cluster representative;
- *Model Selection*, where the model M_i (relative to the corresponding cluster) is selected;
- *Prediction*, where the proper prediction is performed by exploiting M_i.

Actually, the *Prediction* step works in this way. We set a sliding window of fixed size w over the medical records for capturing the patient history depth used for the prediction. A sliding window of size w means that only the last (in time order) w diseases appearing in the record influence the computation of possible forthcoming illnesses. Thus, fixed w, we consider the frequent itemsets of size $w+1$ induced on C_i that contain the w items appearing in the current medical patient record $t_i \in C_i$. The prediction of the next disease is based on the confidence of the corresponding association rule whose antecedent are the w frequent items of t_i, and the consequent is exactly the disease to be predicted. If this rule has a confidence value greater than a fixed threshold, its consequent is added to the set of predicted illnesses.

For the sake of clarity, let us perform a prediction on a medical patient record $t_i^w \in C_i$ of size w. We match t_i^w against all the frequent itemsets $I_{C_i}^{w+1}$ of size $w+1$ induced on C_i. Each itemset $i_i^{w+1} \in I_{C_i}^{w+1}$ containing t_i^w contributes to the set of the candidate diseases with a prediction d_i. It is easy to note that $i_i^{w+1} = t_i^w \cup \{d_i\}$. Finally, if the confidence

of the rule $t_i^w \Rightarrow \{d_i\}$ (i.e., $\sigma(t_i^w \cup \{d_i\})/\sigma(t_i^w)$) is greater than a fixed threshold τ, the disease d_i is considered reliable, and it is added to the set of predicted diseases.

Example. In order to explain the way our prediction approach works in practice, let us consider the set T of patient medical records reported in Figure 2.

t_1	401 715 722 723
t_2	401 721 715 722 723
t_3	401 721 715 722
t_4	241 255 595 780
t_5	241 255 272 595 780

Figure 2: Set T of patient records involving some common diseases.

Let us suppose $k = 2$ be the number of clusters that minimizes the cost function Q_k (see the discussion on clustering in Section 3.1), and $\gamma = 0.5$ be the minimum percentage of occurrences a disease must have for being inserted into the cluster representative (see Equation 1). On the base of the above parameters, it is easily verifiable that the clustering algorithm (Section 3.1) finds the clusters C_1 and C_2, as reported in Figures 3(a) and 3(b), respectively. Furthermore, the clusters are equipped with their representatives: $r_{C_1} = \{401, 721, 715, 722, 723\}$ and $r_{C_2} = \{241, 255, 272, 595, 780\}$. After the clusters have been built, a disease prediction model is carried out for each cluster found.

t_1	401 715 722 723
t_2	401 721 715 722 723
t_3	401 721 715 722

(a)

t_4	241 255 595 780
t_5	241 255 272 595 780

(b)

Figure 3: Cluster C_1 (a) and Cluster C_2 (b).

Now, let $t = \{401, 721, 715, 733\}$ be a new patient disease record. Since the distance $d(t, r_{C_1}) = 1 - 3/6 = 0.5$ is lower than $d(t, r_{C_2}) = 1 - 0/9 = 1$, t is recognized belonging to C_1, thus the model built upon C_1 is exploited to perform the predictions. By fixing $\sigma = 0.8$, the model shown in Figure 4 is obtained.

I^1	I^2	I^3	I^4
721 (2)	721, 715 (2)	721, 715, 722 (2)	401, 721, 715, 722 (2)
715 (3)	721, 722 (2)	715, 722, 723 (2)	401, 715, 722, 723 (2)
723 (2)	715, 723 (2)	401, 721, 715 (2)	
722 (3)	715, 722 (3)	401, 721, 722 (2)	
401 (3)	722, 723 (2)	401, 715, 723 (2)	
	401, 721 (2)	401, 715, 722 (3)	
	401, 715 (3)	401, 722, 723 (2)	
	401, 723 (2)		
	401, 722 (3)		

Figure 4: Disease risk prediction model built upon cluster C_1.

If the window size w is set to 3, this means that only the first three diseases of t are used to generate the predictions, i.e., $t^3 = \{401, 721, 715\}$. By matching t^3 against the 4-frequent itemsets I^4, the disease with code 722 is candidate for being the likely, next disease the patient t may incur in. As previously stated, the disease 722 changes its status from candidate to predicted only if the confidence of the association rule r: $\{401, 721, 715\} \Rightarrow \{722\}$ is greater than the minimum confidence threshold τ. If we set $\tau = 0.8$, since the confidence $\sigma(\{401, 721, 715, 722\})/\sigma(\{401, 721, 722\}) = 1$, the disease 722 is definitively added to the set of predicted illnesses. Therefore, by means of the rule r, we foresee that a patient presenting hypertension (401), spondylosis (721), and osteoarthrosis (715), he is very likely to develop also intervertebral disc disorders (722). □

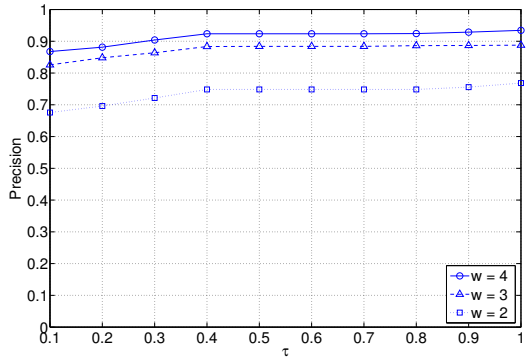
In the next section we show that *CORE* is effective in predicting diseases.

4. Experimental Results

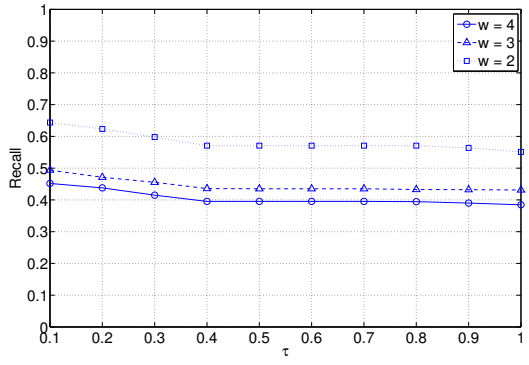
In this section we first define the measures used to test the effectiveness of our approach. Next, we present the results and evaluate them on the base of the introduced metrics. As discussed in Section 2, the dataset T we used for the experiments consists of 1105 patient records involving 330 distinct diseases. In order to perform a fair evaluation we applied the well-known 10-*fold cross validation* method [4], i.e., the original dataset is split in 10 equal-sized partitions. During each of the 10 runs, one of the partitions is chosen for testing, while the rest of them are used for training the prediction model. The cumulative error is found by summing up the errors for all the 10 runs. The strategy we followed for testing our approach is detailed in the following.

First of all, the records in the training set T_{train} are partitioned in k clusters, and for each group, a distinct prediction model M_i is built upon. Relatively to the dataset at hand, we empirically found that $k = 10$ and $\gamma = 0.5$ is the setting that ensures the best possible partitioning for the dataset at hand. A record t in the test set T_{test} is first assigned to one of the k cluster, then it is divided in two subsets of diseases. The first subset, called $head_t$, is used for generating predictions, while the remaining one, referred as $tail_t$, is used to evaluate the prediction. Actually, the length of $head_t$ is tightly related to the maximum window size w allowable for each cluster, and, intuitively, must be lower than the maximal length of frequent itemsets mined in each cluster. For instance, in this very specific case, since we verified that the prediction models M built on clusters (also for low values of support σ) produce frequent patterns of size at most 5, the maximum length of $head_t$ can't exceed 4. More in general, given a window size w, we select the first w diseases as $head_t$ and the remaining $|t| - w$ as $tail_t$. If the record t belongs to the cluster C_i, the relative prediction model M_i matches $head_t$ against all frequent patterns $I_{C_i}^{w+1}$ for generating the candidate predictions.

Fixed the minimum confidence threshold τ, $P(head_t, \tau)$ is the set containing all the candidate predictions whose confidence is greater than τ. Subsequently, the set $P(head_t, \tau)$ is compared with $tail_t$. The comparison of these sets is done by using two different metrics, namely *precision* and *recall* [19]. Precision and recall are two widely used statistical measures in the data mining field. In particular, precision is seen as a measure of exactness, whereas recall is a measure of completeness.



(a)



(b)

Figure 5: Impact of w on precision and recall measures when $\sigma = 0.1$.

By customizing these definitions to our scenario, we exploited precision for assessing how accurate the provided predictions are (i.e., the proportion of relevant predictions to the total number of predictions) and recall for testing if we predicted all the diseases the patients are likely to be affected in the future (i.e, the proportion of relevant predictions to all diseases that should be predicted). Formally, the precision of $P(head_t, \tau)$ is defined as:

$$precision(P(head_t, \tau)) = \frac{|P(head_t, \tau) \cap tail_t|}{|P(head_t, \tau)|}$$

and the recall of $P(head_t, \tau)$ as:

$$recall(P(head_t, \tau)) = \frac{|P(head_t, \tau) \cap tail_t|}{|tail_t|}$$

The cumulative precision (recall) scores drawn in Figure 5 are computed as the mean of the precision (recall) values achieved by each single record $t \in T_{test}$ over the size of T_{test}. More in detail, we measured both precision and recall by varying the threshold τ from 0.1 to 1. Moreover, in order to evaluate the impact of window size w on the quality of predictions, we ranged w from 2 to 4, by considering the predictions done on just one disease unreliable. The results has been obtained by fixing the overall support σ for the frequent patterns to 0.1. Notice that a low support value is necessary for ensuring an adequate length for the mined patterns also in the case of poor cluster homogeneity. As expected, the results in Figure 5(a) clearly reveal that the precision increases as a larger portions of patient medical history, i.e. an increasing number of diseases are used to compute predictions. Conversely, the recall is negatively biased by larger window sizes, as pointed out by Figure 5(b).

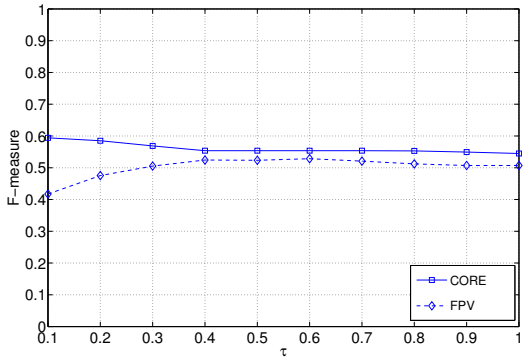


Figure 6: F-measure when $w = 4$, $\tau \in [0.1, 1]$, $\sigma = 0.1$ and $\sigma = 0.01$ for *CORE* and FPV, respectively.

After that, for the sake of comparison, we want to show that the overall prediction performances of *CORE* are better than those obtained by the approach in [9] (henceforth referred as FPV), where a unique, global prediction model M built upon the overall training set T_{train}, is employed. In order to perform a fair comparison, we recur to a well-known metric, the *F-measure* [19], which is the harmonic mean between precision and recall, and it is often used to examine the tradeoff between them:

$$F - measure = \frac{2 * precision * recall}{recall + precision}$$

For this experiment we fixed, for both *CORE* and FPV, $w = 4$ and τ varying from 0.1 to 1. As regards the support value σ, it can be noted that each local model contain, on average, $|T_{train}|/k$ records, where $|T_{train}|$ is the size of the training set and k is the number of clusters in which T_{train} has been partitioned. Thus, since the approaches deal with

different sizes of T_{train}, in order to have a comparable number of frequent itemsets mined by both, we suitably set σ to 0.1 for *CORE* and 0.01 for FPV. Figure 6 clearly shows the better overall performances of *CORE* w.r.t. FPV. This definitively proves that the specialization of the prediction models by means of clustering is meaningful.

5 Conclusions

We presented a recommendation engine based on the combination of clustering and association rules to generate a predictive disease model. The system uses the past medical history of patients to determine the diseases an individual could incur in the future. Experimental results showed that the technique can be a viable approach to disease prediction. Future works aims to compare our method with other proposals in literature, and to perform a more extensive evaluation on large medical history data sets.

Acknowledgements. This work has been partially supported by the project *Infrastruttura tecnologica del fascicolo sanitario elettronico*, funded by Technological Innovation Department, Presidenza del Consiglio dei Ministri, Italy.

References

[1] R. Mooney A. Strehk, J. Ghosh. Impact of similarity measures on web-page clustering. In *Proc of AAAI workshop on AI for Web Search*, pages 58–64, 2000.

[2] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proc. of ACM SIGMOD Conf. on Management of Data (SIGMOD'93)*, pages 207–216, 1993.

[3] D. A. Davis, N. V. Chawla, N. A. Christakis, and A. L. Barabási. Time to CARE: a collaborative engine for practical disease prediction. *Data Mining and Knowledge Discovery*, 20:388–415, 2010.

[4] P. A. Devijver and J. Kittler. *Pattern Recognition: A statistical Approach*. Prentice-Hall, London, 1982.

[5] B. Starfield et al. Comodbidity: Implications for the importance of primary care in 'case' managment. *Annals of Family Medicine*, 1(1):8–14, 2003.

[6] D. A. Davis et al. Predicting individual disease risk based on medical history. In *Proc. of the ACM Int. Conf. on Information and Knowledge Management (CIKM'08)*, pages 769–778, 2008.

[7] I. Lowensteyn et al. Can computerized risk profiles help patients improve their coronary risk? the results of the coronary health assessment study. *Preventive Medicine*, 27(5):730–737, 1998.

[8] P. W. F. Wilson et al. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97:1837–1847, 1998.

[9] F. Folino, C. Pizzuti, and M. Ventura. A comorbidity network approach to predict disease risk. In *Proc. of the Int. Conf. on Information Technology in Bio and Medcial Informatics (ITBAM'10)*, pages 102–109, 2010.

[10] F. Giannotti, C. Gozzi, and G. Manco. Clustering transactional data. In *Proc. of Principles of Data Mining and Knowledge Discovery (PKDD'02)*, pages 175–187, 2002.

[11] C. A. Hidalgo, N. Blumm, A. L. Barabási, and N. A. Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4), 2009.

[12] P. Jaccard. The distribution of the flora of the alpine zone. *New Phytologist*, 11:37–50, 1912.

[13] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the 5th Berkeley Symposium (vol. 1)*, pages 281–297, 1967.

[14] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In *Proc. of ACM Workshop on Web Information and Data Managment (WIDM '01)*, pages 9–15, 2001.

[15] J. E. Pitkow and P. Pirolli. Mining longest repeating subsequences to predict world wide web surfing. In *USENIX Symposium on Internet Technologies and Systems*, 1999.

[16] U. Shardanand and P. Maes. Social information filtering: algorithms for automating word of mouth. In *Proc. of ACM Conf. on Human Factors in Computing Systems (CHI'95)*, pages 210–217, 1995.

[17] R. Snyderman. Prospective medicine: The next health care transformation. *Academic Medicine*, 78(11):1079–1084, 2003.

[18] K. Steinhaeuser and N. V. Chawla. A network-based approach to understanding and predicting diseases. In *Social Computing and Behavioral Modeling*, 2009.

[19] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson International Edition, 2006.