A Comorbidity Network Approach to Predict Disease Risk

Francesco Folino, Clara Pizzuti, Maria Ventura

Institute for High Performance Computing and Networking (ICAR)
Italian National Research Council (CNR)
Via P. Bucci 41C
87036 Rende (CS), Italy
{f.folino,pizzuti}@icar.cnr.it

Abstract. A prediction model that exploits the past medical patient history to determine the risk of individuals to develop future diseases is proposed. The 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. Furthermore, a phenotypic comorbidity network is built and its structural properties are studied in order to better understand the connections between illnesses. Experimental results show that the proposed approach is a promising way for assessing disease risk.

1 Introduction

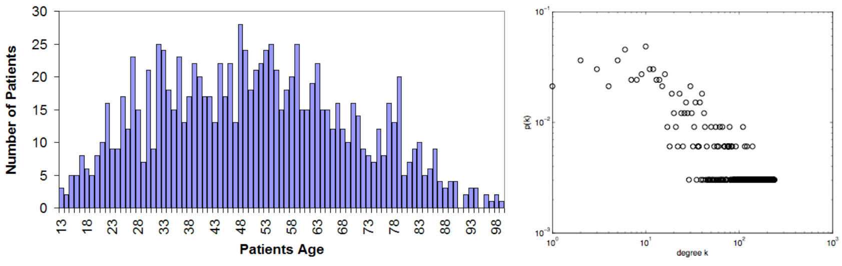
Health care is one of the most important research activity because of its implications in every day life of individuals. An emerging perspective in the last few years aims at identifying individuals most at risk for developing diseases plaguing present age. In fact, prevention or intervention at the 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. Hospitals and physicians collect thousands of patient clinical histories containing important information regarding illness correlations and development. This phenotypic information can be exploited to build a model that predicts disease risk by studying the comorbidity relationships between diseases whenever they contemporarily appear in the same individual. Advanced risk assessment tools are currently at disposal, mainly based on statistical techniques. Another approach for addressing the problem, which is gaining increasing interest, is the use of methodologies coming from the fields of knowledge discovery [5] and network analysis [6]. Some recent proposals in these contexts are those of [2–4].

In this paper we apply network and association analysis on a data set of patient medical records. Our aim is twofold: *(i)* study the relationships of comorbidity appearing in the data set, and *(ii)* generate a predictive model that uses the past patient medical history to determine the risk of individuals to develop future diseases. Analogously to [3] and [4], we construct a phenotypic comorbidity network and analyze its structural properties to better understand the connections between diseases. Then, differently from these approaches, we propose the utilization of association analysis [5] to generate a disease risk predictive model. 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. Experimental results show that approach is a promising method to predict individual risk disease by taking into account only the illnesses a patient had in the past.

The paper is organized as follows. The next section describes the data set used. Section 3 builds two phenotypic disease networks and analyzes its structural properties. In section 4 the predictive model is described. Finally, section 5 reports the evaluation of the proposed predictive approach on the patient data 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 age distribution for the study population is reported in Figure 1(a). The disease codes are those defined by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). 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.



(a) (b)

Fig. 1. Age distribution for the study population (a), degree distribution of the disease network computed using 3-digit codes (b).

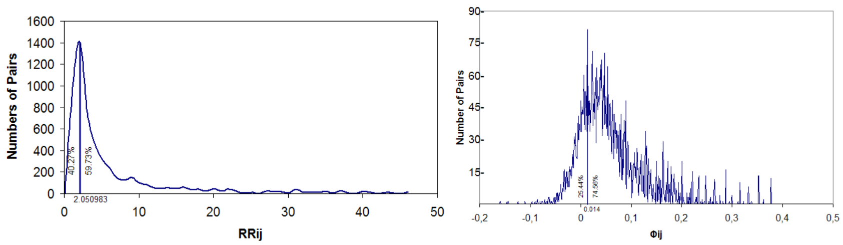
The data is completely anonymized, thus there is no way to identify the patients. In our database the number of diagnoses are 8768 spanning from 1990 to 2009. From an analysis of the patient records, we found that the raw data contained some uninteresting information. These patients have been discarded because not useful for the phenotypic network construction. 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. In order to obtain a more manageable network, the five digits ICD-9-CM codes have been collapsed to these first three digits, so the number of diseases was reduced to 330.

3 Phenotypic Disease Network

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 [3]. Our first goal was to make discernible the correlations among the diseases contained in our data set by building a network whose nodes are the diseases and a link between two nodes occurs when a comorbidity relation appears, i. e. when the couple of diseases affects at least one patient. The edges were labelled with the number of patients showing both the illnesses. An important property to study about networks is the degree distribution [1]. Figure 1(b) reports the degree distribution of our disease network. The figure points out that the network is a scale-free network, i.e. the degree distribution follows a power-law $p_k \approx k^{-\alpha}$, where $\alpha \approx 0.59$. Furthermore the clustering coefficient is 0.69 and the diameter is 4.

The number of edges computed between the nodes was 5736, a too high value to be visualized in a comprehensible manner. Since many edges had weight 1, we adopted the same statistical approaches proposed by Hidalgo et al. [3] to measure the strength of comorbidity relationships, and thus to discard those edges deemed less meaningful. The measures employed to quantify the strength between two sicknesses are the *Relative Risk* (*RR*) and the *ϕ-correlation*. The *RR* of observing a pair of diseases i and j appearing in the same patient is given by $RR_{ij} = \frac{CC_{ij}(N-CC_{ij})}{P_i P_j}$, where CC_{ij} is the number of patients affected by both diseases, N is the total number patients in the data sets, and P_i, P_j are the numbers of patients affected by diseases i and j, respectively. The *ϕ-correlation* is defined as $\phi_{ij} = \frac{CC_{ij}(N-CC_{ij})-P_i P_j}{\sqrt{(P_i P_j (N-P_i)(N-P_j))}}$.

The distribution of RR values for our data set is shown in Figure 2(a), and that of *ϕ-correlation* in Figure 2(b). As pointed out in [3], the Relative Risk overestimates relations involving rare diseases and underestimates relationships between very common sicknesses. On the other hand, *ϕ-correlation* underestimates comorbidity between rare and frequent diseases, and accurately discriminates associations between illnesses of similar appearances. Thus, we built a network by selecting only the statistically significant edges having $RR > 20$, and another network by discarding all the edges having $\phi \leq 0.06$. The two networks are depicted in Figure 3. The network on the left contains 618 edges, the other



(a) (b)

Fig. 2. Distribution of the Relative Risk between all disease pairs (a), and distribution of the ϕ-correlation between all disease pairs (b).

one has 2515 connections. The figure confirms the observation that the Relative Risk underestimates very common diseases. In fact, for example, illnesses like hypertension (code 401), diabetes mellitus (code 250), osteoarthrosis (code 715), or general sysmptons (780), do not appear in Figure 3 (left) because a high percentage of the population is affected by these problems. They are instead depicted in the figure on the right, together with all the others excluded. To better distinguish them, their size is bigger than those already present in the figure on the left.

4 Disease Risk Prediction

A general predictive model to assess disease risk can be realized by studying the patterns of co-occurrences across the medical patient records. Each patient can be associated with the list of diseases he has been affected during his life. Groups of illnesses occurring frequently in many patient records can be exploited to capture comorbidity relations and generate predictions about the diseases a patient can incur, given the past history of his health conditions. To this end, a valuable help can come from *association analysis*. Association analysis [5] is an important data mining methodology for discovering interesting hidden relationships in large data sets. It relies on the concept of *frequent itemset* to extract strong correlations among the items constituting the data set to study.

Let DS be the set of medical patient records, $D = \{d_1, \ldots, d_n\}$ the set of illnesses appearing in DS, and $T = \{t_1, \ldots, t_m\}$ a set of m patient transactions, where each t_i is a subset of D, i.e. a set of diseases. Groups of diseases occurring frequently together in many transactions are referred to as *frequent itemsets*. The concept of frequency is formalized through the concept of *support*. Given a set $I = \{I_1, \ldots, I_k\}$ of frequent itemsets on T, the support of an itemset $I_i \in I$, $\sigma(I_i)$, is defined as $\sigma(I_i) = \frac{|\{t \in T | t_i \subseteq t\}|}{|T|}$, where $|\ .\ |$ denotes the number of elements in a set. The support, thus, determines how often a group of diseases appears together. It is a very important measure since very low support discriminates those groups of items occurring only by chance. Thus a frequent itemset, in order to be considered interesting, must have a support greater than a fixed threshold value, *minsup*. An association rule is an implication expression of

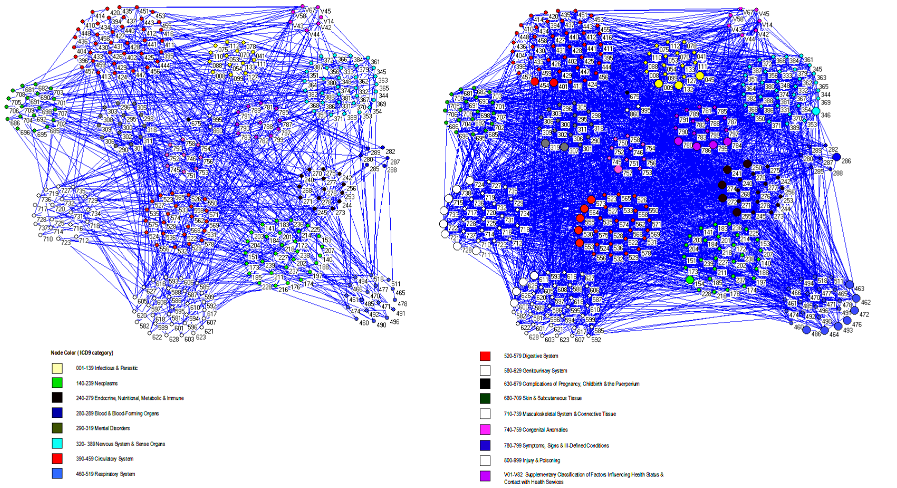

Node Color (ICD9 category)
001-139 Infectious & Parasitic
140-239 Neoplasms
240-279 Endocrine, Nutritional, Metabolic & Immune
280-289 Blood & Blood-Forming Organs
290-319 Mental Disorders
320- 389 Nervous System & Sense Organs
390-459 Circulatory System
460-519 Respiratory System
520-579 Digestive System
580-629 Genitourinary System
630-679 Complications of Pregnancy, Childbirth & the Puerperium
680-709 Skin & Subcutaneous Tissue
710-739 Musculoskeletal System & Connective Tissue
740-759 Congenital Anomalies
780-799 Symptoms, Signs & Ill-Defined Conditions
800-999 Injury & Poisoning
V01-V82 Supplementary Classification of Factors Influencing Health Status & Contact with Health Services

Fig. 3. Disease network with Relative Risk above 20 (top). Disease network with ϕ-correlation above 0.06 (bottom). Different colors denote the ICD-9-CM categories appearing in the dataset. Codes labeling bigger points are those that do not appear in the RR network.

the form $X \Rightarrow Y$, where X and Y are disjoint itemsets. The importance of an association rule is measured by its *support* and *confidence*. The support of a rule is computed as the support of $X \cup Y$ and tells how often a rule is applicable. The confidence is defined as $\frac{\sigma(X \cup Y)}{\sigma(X)}$, 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, and can be exploited to build a risk prediction model by matching the medical record of a patient against the patterns discovered by the model. In this scenario, the support determines how often a group of diseases appears together, while a rule like $X \Rightarrow \{d\}$, where $X \subseteq D$ is a subset of diseases and d is a single disease, having a high confidence allows to reliably infer that d will appear together with the items contained in X. The idea we pursue in this paper thus consists in using frequent itemsets of diseases for predicting a set of diseases a patient could likely be affected in the future, given the patient clinical history.

We use a sliding window of fixed size w over the medical records for capturing the patient's 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 prediction of possible forthcoming illnesses. Given a fixed window size w, we consider only the frequent itemsets of size $w+1$ that contain the w items appearing in the current medical patient record t_i. The prediction of the next disease is based on the confidence of the corresponding association rule whose consequent is exactly the disease to be predicted. Thus, if the rule has a confidence value greater than a fixed threshold, the disease on the right of the arrow is added to the set of predicted illnesses.

In order to explain the way our prediction approach works, let the transactions reported in the top table of Figure 4 be a set of some patient's medical records. By fixing the minimum support threshold σ to 0.8 (i.e., an itemset is frequent if it is present at least 4 times in the transaction set), the algorithm finds the patterns in the bottom table.

t_1	$\{401, 722, 723, 715\}$
t_2	$\{401, 722, 715, 462, 723\}$
t_3	$\{401, 722, 715, 462\}$
t_4	$\{722, 715, 401, 462\}$
t_5	$\{723, 401, 722, 715, 462\}$

Length 1	Length 2	Length 3	Length 4
{401} (5)	{401, 722} (4)	{401, 722, 462} (4)	{401, 722, 462, 715} (4)
{722} (5)	{401, 462} (4)	{401, 722, 715} (5)	
{462} (4)	{401, 715} (5)	{401, 462, 715} (4)	
{715} (5)	{722, 462} (4)	{722, 462, 715} (4)	
	{722, 715} (5)		
	{462, 715} (4)		

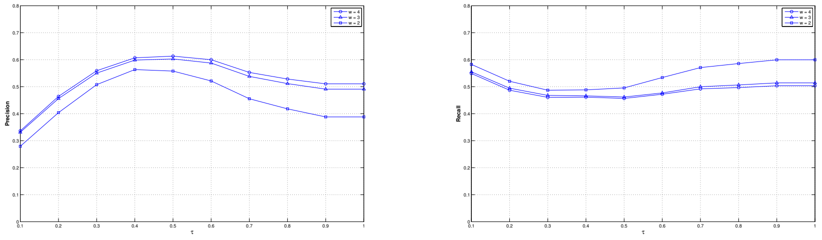
Fig. 4. Example of transactions involving some common diseases (a). Frequent itemsets mined by the algorithm(b).

Now, let $t = \{722, 715, 401, 733\}$ be a new medical record. If the window size w is set to 2, this means that only the two first diseases are used to generate the predictions, i.e., $\{722, 715\}$. By matching $\{722, 715\}$ against the 3-frequent itemsets, the items with code 401 and 462 are proposed as likely, next diseases. The scores of predicted illness 401 and 462 are 1 and 0.8. As previously described, these scores correspond to the confidences of the association rules $\{722, 715\} \Rightarrow \{401\}$ and $\{722, 715\} \Rightarrow \{462\}$, respectively.

5 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 we used for the experiments consists of 1105 transactions involving 330 distinct diseases. In order to perform a fair evaluation we applied the well-known *k-fold cross validation* method [5], with $k = 10$.

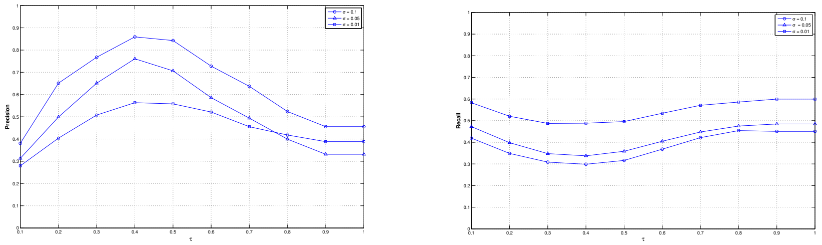
We tested our approach in the following way. Each transaction t in the evaluation set is divided in two groups of diseases. The first group of diseases, called $head_t$, are used for generating predictions, while the remaining, referred as $tail_t$, are used to evaluate the predictions generated. The length of $head_t$ is tightly related to the maximum window size allowable for the experiments. In our case, since the mining phase of frequent patterns produced itemsets of size at most 5, the length of $head_t$ has been fixed to 4. Thus, given a window size $w \leq\mid t \mid$, we select the first w diseases for generating the predictions and the remaining



(a) (b)

Fig. 5. Impact of w on precision and recall measures when $\sigma = 0.01$.

$|t| - w$ for testing their prediction. Fixed $head_t$ and a confidence threshold τ, we produce the prediction set $P(head_t, \tau)$ containing all the predictions whose score is greater than τ. Then 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*. 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. In order to obtain an overall evaluation score for each measure (fixed a confidence threshold τ) we computed the mean over all transactions in the test set. In the experiment presented we measured both precision and recall by varying 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.



(a) (b)

Fig. 6. Impact of σ on precision and recall measures when $w = 2$.

Figure 5 shows the impact of w on precision and recall of the predictions. We obtained these results by fixing the overall support for the mining of frequent patterns to 0.01. As expected, the results in Figure 5(a) clearly reveal that the precision increases as a larger portions of patient's medical history, i.e. an increasing number of diseases, are used to compute predictions. Conversely, the recall is negatively biased by larger window sizes, but this effect slightly fades for higher values of τ (see Figure 5(b)).

Figure 6 displays the behavior of precision and recall metrics when the support threshold varies. We used $w = 2$ since it is the maximum allowable window size when the support reaches the value 0.1. Increasing the support threshold has two main positive effects: *(i)* improving the precision of predictions, and *(ii)* ensuring the scalability of the association rule mining algorithm, since a lower number of frequent itemsets are computed. However, as a side effect, a higher support results in a potential loss of some important, yet infrequent, diseases in the prediction set. In the medical context, this kind of illnesses could be particularly important and more informative for producing a correct diagnosis. Figure 6(a) clearly points out better performances of precision when the support threshold increases. Indeed, it is easy to notice that, for $w = 2$ and $\tau = 0.4$, we obtain a precision of 0.5635 if $\sigma = 0.01$, whereas the precision reaches the value 0.7608 and 0.8593 for $\sigma = 0.05$ and $\sigma = 0.1$, respectively. An inverse trend can be noted in Figure 6(b) for the recall which, even for $w = 2$ and $\tau = 4$, progressively decreases from the value 0.4903, when $\sigma = 0.01$, to 0.2985, when $\sigma = 0.1$, respectively.

6 Conclusions

We constructed a phenotypic comorbidity network and studied its structural properties to better understand the connections between diseases. Then we presented a methodology based on associative rules to generate a predictive model that uses the past medical history of patients to determine the risk of individuals to develop future diseases. 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, in particular with a collaborative filtering technique based on the k-nearest-neighbor, like that employed by [4].

Acknowledgements. This work has been partially supported by the project *Infrastruttura tecnologica del fascicolo sanitario elettronico*, funded by Technological Innovation Department, Pres. Cons. Min.

References

1. Reka Albert and Albert-László Barabási. Staistical mechanics of complex networks. *Reviews of modern physics*, 74:47–97, 2002.
2. Darcy A. Davis, Nitesh V. Chawla abd Nicholas A. Christakis, and Albert-László Barabási. Time to CARE: a collaborative engine for practical disease prediction. *Data Mining and Knowledge Discovery Journal*, 20:388–415, 2010.
3. César A. Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas A. Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4), 2009.
4. Karsten Steinhaeuser and Nitesh V. Chawla. A network-based approach to understanding and predicting diseases. In *in Social Computing and Behavioral Modeling, Springer*, 2009.
5. Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining.* Pearson International Edition, 2006.
6. Stanley Wasserman and Katherine Faust. *Social Network Analysis. Methods and Applications.* Cambridge University Press, 1994.