Link Prediction in Weighted Symptom Networks

Buket Kaya and Mustafa Poyraz
Department of Electrical and Electronics Engineering, Fırat University, Elazığ, Turkey
bkaya@firat.edu.tr

Abstract—The saying “treat the disease, not the symptoms” is widespread, a cliche for eliminating or repairing the root of a problem rather than mitigating the negative effects. It is taken for granted that prevention is the best course of action. It is ironic, then, that many of today’s best “disease treatments” are actually symptom suppressors. This paper predicts the onset of future symptoms on the base of the current health status of patients. The problem of predicting the relations between symptoms (abnormal parameters in this paper) which can be shown as the reason of a disease in the future is a really difficult and, at the same time, an important task. For this purpose, the present paper first constructs a weighted symptom networks considering the relations between abnormal parameters. Then, it proposes a link prediction method to identify the connections between parameters, building the evolving structure of symptom network with respect to patients’ ages. To the best of our knowledge, this is the first attempt in predicting the connections between the results of laboratory tests. Experiments on a real network demonstrate that the proposed approach can reveal new abnormal parameter correlations accurately and perform well at capturing future disease risks.

I. INTRODUCTION

Social networks are structures composed by entities that can be connected by different forms of social relations. In this structure, connections and entities tend to appear and disappear along time, which turns them into highly dynamic and complex systems. Link prediction is the most fundamental problem that attempts to infer which new links are likely to occur in the near future based on the topological, node and edge properties in a given network [1]. That is, if we are presented with a snapshot of a network at the current time, the goal is predicting links that will occur in the next time step. As part of the recent surge of research on large, complex social networks and their properties, a considerable amount of attention has been devoted to the computational analysis of social network evolution. In social networks nodes represent people or other entities embedded in a social context, and whose edges represent interaction, collaboration, or influence between entities.

Link prediction problem has interpreted and defined in many ways. All of these methods are based on the measures indicating the proximity between nodes. These measures proposed in the literature are generally categorized into semantic or topological/structural measures [2]. In semantic measures, the content of the nodes is taken into account to measure proximity. For example, in a co-authorship network, the similarity between keywords generated from published papers can be used to predict future connections among the authors. As apart from the semantic measures, the topological strategy employs the network structure to compute the proximity values. Topological measure are more widely adopted because they are more general and do not require detailed information related to the content of node. Moreover, this content is not always present depending on the social network considered. Topological measures are categorized into neighborhood-based or path-based measures [3]. The neighborhood-based measures consider the immediate neighbors of the nodes. According to this measure, two nodes are more likely to form a link if their sets of neighbors have a large overlap [2]. The most widely used measures among them are Common Neighbors [4], Preferential Attachment [4,5], Adamic-Adar Index [6] and Jaccard’s Coefficient [7]. The path-based measures generate a proximity value considering the paths between the related nodes. The main idea is that two nodes are more likely to form a link if there are short paths between the related nodes [8].

Many studies have been done to overcome the link prediction problem [1, 9, 10]. However, in order to predict new connections at future time, most of the previous studies are based on the applications of proximity measures to non-connected pairs of nodes in the network at present time. The values obtained from proximity measures indicating the similarity between pairs of nodes can be used either by unsupervised [1,9,11], or supervised link prediction [10,12,13].

In this paper, we propose a predictor to determine the risk of individuals to develop disease, and to undertake the correct actions at the earliest signs of illness. To this purpose, we first construct a weighted symptom network which indicates the relationships between abnormal parameters of disease. In this network, each node represents the abnormal parameters of patients; the edges connect these parameters appearing in the same patient. Then, we present a link prediction method based on the evolving cases of the constructed symptom network. The method increases the importance of more recent cases observed for a pair of symptom. Finally, we test the proposed method on the symptom network constructed with laboratory results of patients more than 210,000. The experimental results demonstrate the accuracy of our method on unsupervised prediction and encourage us for further analysis.

The rest of the paper is organized as follows. Section 2 explains the building of symptom network. Section 3 presents our link prediction method, proposing the evolving structure of the symptom network and the cases
in this network. The conducted experiments and the results achieved for the symptom network are discussed in Section 4. Conclusions and future research directions are included in Section 5.

II. BUILDING WEIGHTED SYMPTOM NETWORKS

In this study, medical data comprises the results of the laboratory tests. In these data, there are many medical parameters (symptoms) tested for a patient. If all the values of these parameters are within reference interval, it can be said that the person coming to the hospital does not have a disease. However, if one of them or several parameters are not within reference interval, other words, it exceeds the maximum value of the interval or it is lower than the minimum value of the interval, then this parameter exhibits an abnormal case and it is probable that the person having this laboratory result has a disease.

In the symptom network, nodes are the abnormal parameters which are not within reference interval. If the value of the parameter \( X \) is larger than the maximum value of reference interval then, this node is represented by \( X.H \), where \( H \) stands for High. In contrast, if the value of the parameter \( X \) is lower than the minimum value of the reference interval then it is represented by \( X.L \), where \( L \) stands for Low. If the parameter \( X \) has a categorical or binary value, the unexpected value is represented by \( X.A \) where \( A \) stands for Abnormal. In the constructed symptom network, edges connect these pairs of abnormal parameters appearing in the same patient. Thus, a relationship between two abnormal parameters exists whenever they appear simultaneously in a patient. In this paper, we construct a symptom network and exploit the connections between abnormal parameters to build a prediction model. In this model, we propose the utilization of link prediction methods to generate the predictive model of abnormal parameters which causes to an illness.

The patient medical records contain an important view regarding the co-occurrences of abnormal parameters affecting the same patient. In order to understand better the correlations among the abnormal parameters contained in our dataset, a symptom network whose nodes are the abnormal parameters and a link between two nodes occurs every time those parameters are together appeared in the same patient, i.e., when the couple of parameters affects at least one patient, is built. The edges are labelled with the number of patients indicating both the parameters.

III. THE PROPOSED METHOD

As known, social networks are highly dynamic structures in which several connections and nodes tend to appear or disappear along time. The evolution of these networks brings valuable information about how connections tend to be formed, which will be then called as the link prediction task. In the present work, we propose a novel proximity measure that considers the evolving structure of a weighted network, such as symptom network, and evolving cases related to pairs of nodes in the network. The aim of the paper is to combine Homans’ idea [15] that the strength of a connection between two nodes is directly associated with how often they interact with one another and Newman’s method [16] that the bigger the number of common neighbors between two nodes, the higher is their probability to be connected in the future [14]. Since the weighted network is used in this paper, we modified Newman’s method in the form of “the larger the total weights of common neighbors between two nodes, the higher is their probability to be connected in the future”.

The main idea of the proposed approach is to increase or decrease the proximity score between two nodes depending on the evolving cases observed among the two nodes and their neighborhood. An evolving case, which will be explained in more detail in the next section, is defined through the increment or decrement of weights of links around nodes as the network evolves. In the proposed study, an evolving structure is first constructed by extracting sequential frames of the networks at the different age intervals. Then, proximity score is computed for each pair of nodes by aggregating the scores assigned for the evolving cases observed during the transitions of frames.

A. The Evolving Structure of Weighted Symptom Network

The evolving structure represents states of the network at different age intervals in the past. For this purpose, in this paper, we split the network into several age-sliced snapshots. Then, we built the frames by grouping sequential snapshots. For instance, the frame at \([18-25]\) age interval is a sub-graph containing all links observed among abnormal parameters of medical data of all people in between 18 and 25 years old. It should be noted that the size of each frame is the same as the length of the prediction window, pre-defined in the link prediction task.

B. The Evolving Cases

An evolving case is generally the action that leads a pair of nodes from a state (linked or unlinked) to another. An evolving case is defined as the increment or decrement of weight of a link between two nodes from a frame to its subsequent in a weighted network like symptom network. These cases can be categorized into one of three different types; consistent, strengthening and weakening.

Consistent Case:
A consistent case in the weighted symptom network occurs when the weight of the relation between two nodes does not increase more than a case changing rate pre-determined or does not decrease less than this rate while the network evolves, that is when two abnormal parameters share a link in a frame and the strength of this connection is preserved in the subsequent one. In such a case, in order to consider a consistent case during the transition from the \((k-1)\)-th to the \(k\)-th frame, \( C_k(\pi_i, \pi_j) \) score for each pair of nodes \((\pi_i, \pi_j)\) is defined as follows:
where \(w_{R-1}(x,y)\) and \(w_R(x,y)\) are the weights of connection between \(x\) and \(y\) nodes in frames \(F_{R-1}\) and \(F_R\), respectively. \(r\) is the case changing rate pre-determined, the constant \(c (0 < c \leq 1)\) indicates the positive score for consistent cases since the strength of a tie between two nodes is preserved.

**Strengthening Case:**

Strengthening cases occur when the weight of link between two nodes on different frames is increased more than the case changing rate pre-determined. In other words, they happen when the weight of two abnormal parameters in a frame increases more than the case changing rate in the next frame. The strengthening case \(S_k(x,y)\) score associated to a pair \((x,y)\) and a frame \(F_k\) is computed as follows:

\[
S_k(x,y) = \begin{cases} 
0 & \text{if } w_k(x,y) \geq w_{k-1}(x,y) + r \cdot w_{k-1}(x,y) \\
\alpha & \text{otherwise}
\end{cases}
\]  

\[(2)\]

where \(\alpha (0 < \alpha \leq 1)\) indicates a reward for the strengthening cases, its value should be positive since the tie between two nodes is strengthened.

**Weakening Case:**

Weakening case is opposite to the strengthening one. In the kind of this case, the strength of existing link between two nodes decreases from a frame to its subsequent. The weakening case score \(D_k(x,y)\) is computed as follows:

\[
D_k(x,y) = \begin{cases} 
0 & \text{if } w_k(x,y) \leq w_{k-1}(x,y) - r \cdot w_{k-1}(x,y) \\
\beta & \text{otherwise}
\end{cases}
\]  

\[(3)\]

where \(\beta (-1 \leq \beta < 0)\) is a decrement value and should be negative since the strength of the connection between the nodes tends to decrease.

C. The Proximity Score in Weighted Symptom Networks

Many methods for link prediction compute scores to pairs of nodes by employing a chosen proximity measure. Thus, it is determined how similar those nodes are and, consequently, how likely a connection between them will be formed in a near future.

In this paper, we have combined the scores associated to primary cases, which are the evolving cases directly related to pairs of nodes under consideration, with the scores associated to secondary cases, which are the evolving cases observed in the nodes’ neighborhood. Moreover, we have raised the importance of more recent cases taking into account a monotonically increasing function in between the frames. For this purpose, we used a logarithmic function to weight recent cases more heavily than the old ones.

Finally, the proximity score concerning a given pair of nodes \((x, y)\) is determined as:

\[
\begin{align*}
\text{Score}(x, y) &= \sum_{k=1}^{n} I(k) [P(x, y, k) + \alpha S(x, y, k)] \\
S(x, y, k) &= C_k(x,y) + S_k(x,y) + D_k(x,y) \\
P(x, y, k) &= \sum_{i=1}^{n} \prod_{j \in I(k)} P_i(x_j, y, k)
\end{align*}
\]

\[(4)\]

\[(5)\]

\[(6)\]

\(P(x, y, k)\) computes the score of the case for the pair of nodes \((x, y)\) observed in the transition from the frame \(k-1\) to the frame \(k\). The case observed at this moment may be one of three evolving cases; consistent, strengthening or weakening. \(S(x, y, k)\) gives the aggregated score of secondary cases concerning to the pair \((x, y)\). \(I(k)\) represents the importance value of the frame. In this study, \(I(k)\) is \(\log(k)\). Here, \(J(x)\) is the set of neighbors of the node \(x\) in the constructed network. The parameter \(\alpha\) is an amortization factor that shows how strong secondary cases affect the tie between \(x\) and \(y\).

In our experiment, the proposed measures were compared with different measures previously adopted in the literature for link prediction. We will use the proximity scores in the weighted networks since the edges are labelled with the number of patients having abnormal parameters in this paper. For this purpose, five proximity scores modified for weighted networks were considered; Common Neighbors [4], Jaccard’s Coefficient [7], Preferential Attachment [5], Adamic-Adar Coefficient [6] and Resource Allocation Index [17].

IV. EXPERIMENTAL RESULTS

We conducted a set of experiments using real medical data in order to evaluate our proposed method. For this purpose, we first collected three different laboratory results of the 210,134 patients. These are complete blood count, clinical biochemistry blood test and urinalysis, which are the main sources for our study. Then, we concentrate on predicting links between abnormal parameters of these tests for various versions of the symptom network constructed. The proposed method was implemented in C++.

In order to evaluate the performance of our method, the laboratory results obtained from Fırat University Hospital in Turkey were used in our experiments. Patients’ medical records were taken suitably to the ethics rules of the hospital. As mentioned earlier, while nodes represent the abnormal parameters, edges occur when the couple of these abnormal parameters affect at least one patient. The total number of nodes is 386 since if a parameter having continuous value is within reference interval it is represented by two nodes as lower than or higher than normal. The number of edges is 78276. In this study, we performed experiments using three versions of the network.
(1) an unweighted version in which a link is presented between a pair of abnormal parameters if they appeared together at the least one patient.

(2) a weighted version of the network in which each link between a pair of abnormal parameters is weighted by the total number of patients showing both abnormal parameters. Here, the case changing rate \(r\) is 50% and \(I(k)\) is 1.

(3) another weighted version of the network where the case changing rate \(r\) is the same with the previous one but \(I(k) = \log(k)\).

The weights in the relationships indicate how strong is the link between two parameters instead of only considering the binary relation between them. The purpose of doing experiments with the different versions of the network is to investigate whether the use of weights in relationship and different importance values of the frames provide better prediction results. In our experiments, the frames were determined with respect to ages of patients as apart from the studies in the literature. Therefore, the aim of the analysis is to investigate how the symptom network evolves according to ages of patients.

The evolving structure of symptom network was initially built. In this construction, the frames at \([20,30), [30,40), [40,50)\) age intervals were determined for the laboratory results and the snapshot at \([50,60)\) age interval was used as the prediction frame. In this paper, we have employed neighborhood-based measure because information about the connections around nodes is important for us. The values \(c, s, d\) which are given for the consistent, strengthening and weakening cases, respectively, are estimated by evaluating the link prediction performance on a validation set. The validation set is the last frame before the prediction frame. Since the aim of the link prediction task is to investigate a new link or weighted link in a network, the non-connected pairs to be considered are chosen from the validation set. However, the non-connected pairs that were directly or indirectly not affected by some case during the network evolution were eliminated. Here, it should be also noted that a pair of nodes is assumed as negative at all the predictors in the this paper if the weight of the edge connecting the nodes decreases more than the case changing rate from the validation set to the prediction frame and assumed as positive otherwise. We also used the validation set to empirically evaluate and select the most appropriate values of \(c, s, d\) and \(\alpha\) parameters in the link prediction task. We showed with some tests that the best performance was achieved at \(c = 0.4, s = 0.9, d = -0.3\) and \(\alpha = 0.07\).

The first experiment deals with finding the precision values of three different versions of the symptom network in order to compare our case-based scores to the traditional proximity scores modified for weighted networks. Figure 1 reports the results of this experiment. As can be seen from Figure 1, the best prediction performances are achieved by our link predictor over weighted version 2 of the network. In this experiment, the unweighted version of the network obtains worse results.

Similarly, the second experiment finds the recall values of every three versions of the network. Results in Figure 2 show that while the weighted version 2 of the network exhibits worse performance at the traditional proximity measures, it obtains a better result at our method. The last experiment compares F-measure values for every three versions of the network. Our case-based scores given in Figure 3 obtained the best results from all the methods considered. In our experiments, we observed a performance gain when the case-based score using \(I(k) = \log(k)\) was compared to the case-based score using \(I(k) = I\) in the network. This situation reveals an important fact that recent cases carry more information about the occurrence of links.

![Figure 1. Precision values of different predictors in symptom network](image)
V. CONCLUSIONS

In this paper, we proposed a link prediction method for finding the most appropriate connections between abnormal parameters in symptom network. The method employs unsupervised machine learning strategy for link prediction. As the contributions of the paper, we first built a novel weighted symptom network which was combined from complete blood count, clinical biochemistry blood test and urinalysis. Next, we propose an efficient link prediction method for weighted social networks, such as symptom network, by defining the evolving cases. In that network, the frames were constructed with respect to patients’ ages.

The experiments conducted on real data sets illustrate that the proposed approach produces meaningful results and has reasonable efficiency. Our case-based link prediction score outperforms all the methods compared in this paper in terms of prediction, recall and F-measure. The proposed method could be also directly applied to other diverse types of weighted social networks, or it can be extended to address problems not yet considered.

In the future, we will revise our case-based score to make this score more realistic, improve our algorithm such that one could have better performance in more complex networks having bipartite nodes.

REFERENCES


