Risk Prediction on the Patient’s Outcomes

Kyaw May Oo, Ni Lar Thein

University of Computer Studies, Yangon, Myanmar

kmayoo19@gmail.com
nilarthein@mptmail.net.mm
http://www.ucsy.edu.mm

Abstract: Identifying the dire outcomes and patients profiling in terms of disease related data and patient characteristics can significantly influence the decision making about the patient with disease control. To estimate the patient outcomes with respect to the patient’s symptoms, we propose the exploitation of frequent patterns as an underlying technique for this purpose. We use the modification of SFM (Swarm to Frequent Notesets Mining) algorithm to mine the frequent patterns. We demonstrate that our approach offers comparative performance in terms of accuracy and execution time. Our proposed approach 1) can avoid the weighting up process which makes the most computational cost in classical classification approach; 2) can provide more user understandability than the statistical model prediction approach; and 3) is more computational simpler than both approaches.

Key words: Classification, Data Mining, Frequent pattern, Prediction, SFM.

INTRODUCTION

Hospital patient related databases are a rich source of information about the health status of patients. We investigate the potential of this data to be used to predict the risk of a dire outcome for patients. In practical terms, an accurate risk prediction system could give clinicians an early indication of danger, thereby allowing enough time for medical intervention or closer monitoring of the patient. While the medical aspect of this research is important, the central aim of this paper is to present a practical approach and to investigate the exploitation of frequent pattern as an underlying technique for risk prediction purpose.

Identifying the dire outcomes and patients profiling in terms of disease related data and patient characteristics provide new insights into the complexity and causes of adverse. In general, methods developed for identifying disease concentrations often apply to adverse which is pinpoint concentrations of disease that migrate over time. It can provide to raise awareness of prevention and early detection.

There are two key ways in which one may offer a risk prediction for a patient suffering from diseases. One can find arguments in favor of adverse outcome and another against adverse outcome and try to balance the former against the latter to reach a conclusion. In this approach the reasons for the decision are clear, but the weighting up of arguments is difficult. Another one can construct a statistical model to estimate the probability of adverse given the patient’s symptoms. This latter approach offers reliable conclusions but no obvious qualitative chain of inference to the conclusion.

In this research frequent pattern mining is applied for understanding the characteristics of the adverse outcomes associated to patient profiling. For more clearly we use only the frequent patterns, but do not use the association rules. In particular, the modification of SFM technique of frequent pattern mining is used as an explorative technique to generate adverse outcomes patterns. More specifically, adverse outcomes circumstances that frequently occur together in respective patients will be identified. Furthermore, these patterns are compared with benign characteristics occurring in those patients. This allows the investigation of the differences between patient patterns for adverse and benign.

The remainder of this paper is organized as follows. First we briefly describe some widely used methods for prediction techniques are provided. This will be followed by a formal introduction to the pattern mining with the concept of frequent pattern mining algorithms as well as the traditional and our new frequent pattern mining problem definition in Section 2. In Section 3, we present the SFM algorithm and its modification. In Section 4, the empirical study is explained with a description of the dataset and the results of this study are presented. The paper will be
completed with a summary of the conclusions and directions for future research.

1. Relation to other work

We briefly describe some widely used techniques for risk prediction from patient profiling to disease control.

These risk prediction characteristics are currently modeled using statistical techniques to provide an estimate of the probability of benign and adverse. In the case of breast cancer disease, two commonly used systems are the Nottingham Prognostic Index (NPI) [5], which uses data from large UK studies, and results derived from the American Surveillance, Epidemiology and End Results (SEER) database [12], which is used by systems such as Adjuvant Online [11]. Both techniques rely on multivariate analyses of large volumes of data (based on over 3 million people for SEER) to calculate prognostic formulae.

These tools, and others like them, are effective at providing estimates of risk of adverse such as death and local recurrence. However, they have two major weaknesses. Whilst effective, they lack explanatory power in a human-readable form. Therefore, extra knowledge that has not been captured by the statistical analysis cannot be easily incorporated. Secondly, knowledge that post-dates the formation of the formulae is very difficult to incorporate. Therefore, while they excel at providing an accurate assessment of population-based risk, they have weaknesses in the individualization of that risk.

Humans are often poor at manipulating explicit probabilities [6] [4]; however, clinicians have the ability to process additional knowledge that statistically-based systems often either ignore or treat on a perfunctory level. We would like to support clinical decision making by providing explicit probabilistic estimates of risk whilst also allowing the reasoning behind this risk prediction to be clear. In order to do this, we need to provide the common patterns in a form, with the significance of the correlation that can be used by clinicians so that they may integrate both types of knowledge.

Thus, in this paper we present a frequent pattern mining approach, which can reap the rewards of each of above two approaches while avoiding their disadvantages, to prediction the adverse outcome. The focus of this paper is how to apply frequent pattern mining approach, called SFM, for discovering different types of patterns in risk prediction problem from patient profiling. This approach can find frequent patterns in favor of adverse and patterns against adverse with their probability of adverse.

2. Pattern Mining

Frequent pattern mining is fundamental problems of data mining tasks and a very young research field born in 1993 [2]. It aims to find frequently occurring subsets in a sequence of sets. After the problem birth, a large number of frequent pattern mining algorithms have been published in the literature. Agrawal, Imielinski and Swami introduced the first algorithm, called AIS, as well as the problem itself [2]. Shortly after that Agrawal and Srikant made improvement over AIS by exploiting the monotonicity property and renamed Apriori, a widest well-known algorithm [3]. After that, many Apriori modification algorithms were proposed, for instance DHP [8], DIC [14], Partition [13], the Sampling algorithm [16], Eclat [17], and CBW [15].

2.1. Definitions

We now introduce some of the basic terminology of frequent patterns. A transaction database is a sequence of transactions \( T = \{t_1, t_2, ..., t_n\} \), where each transaction is an item set \( t_m \subseteq T \). An item set with \( k \) elements is called a \( k \)-item set. The support of an item set \( X \) in \( T \) denoted as sup(X), is the number of those transactions that contain \( X \), i.e. sup(X) = \(|\{t_m : X \subseteq t_m\}|\). An item set is frequent if its support is greater than or equal to a support threshold, originally denoted by min sup.

2.1.1. Traditional Frequent Pattern Mining Problem Definition

The traditional frequent pattern (itemset) mining problem definition is: given a set of transactions, a user-specified minimum support, find all frequent patterns that are above the user-specified minimum support in a given transaction database. This is well suited for the traditional market basket analysis problem. We believe that the problem definition is should be modified to mine the frequent patterns for our target application as the reasons explained below.

2.1.2. Our new Frequent Pattern Mining Problem Definition

Frequent pattern mining algorithm produces an exponential amount of patterns which lead to decrease the performance. Moreover to relevant our target prediction application we put the constraints, which make the narrow search space, in problem definition. This suggests that the new problem definition of frequent pattern mining for our target application is: given a set of transactions, a user-specified minimum support, a target criteria (i.e. recurrence dire outcome for breast cancer case), desired frequent pattern size, find all frequent patterns with desired size that are above the user-specified minimum support and satisfy the target criteria in a given transaction database.

3. Method

In this section, we present a brief introduction to SFM algorithm for the frequent pattern mining problem. We also present the modification over SFM with minor adaptation to be more relevant for our target prediction application.

3.1. Our SFM Algorithm

The SFM algorithm is based on the behavior of real ant colonies, a field of Swarm Intelligent, as well
as some data mining concepts. The SFM algorithm is shown in figure 1. For further details, see [7]. In this system, the following three phases are composed.

**Begin**

\[ \varepsilon = \text{min-pheromone}; \]

\[ k = 0; \]

\[\forall \text{ predefined nodeset } S \]

\{ \text{Ant}_{k+1} \text{ travel predefined nodesets; }\]

\{ \text{Update pheromone amount of trial followed } \text{by } \text{Ant}_{k+1}; \}

\}\n
\[\forall \text{ node } S \]

\{ \text{Define one-step-neighbors; }\]

\{ \text{Create Ant-routing-table; }\]

\{ \text{if } (|\text{neighbors}|>1) \{\text{ns}\} = \text{node}; \}

\}\n
\[\forall \text{ ns } S \]

\{ \text{Nodeset} = \text{Nodeset Generation}(n_s); \]

\[\forall \text{ Nodeset } S \]

\{ \text{Calculate pheromone amount of Nodeset}; \]

\{ \text{Answer } = \{\text{Nodeset } | \text{Nodeset.phero }= \varepsilon\}; \}

**End**

**Figure 1. SFM Algorithm**

**3.1.1. Defining the Neighbors**

In this first phase, we assume that all the transaction of dataset D be the previous existing path (i.e. already constructed nodesets) of the traveling space S traveled by real ants. So to determine the deposited pheromone amount, let us assume that a colony of ants be traveled over S. Within these traveled, every ant deposited pheromone on each travel.

Thus after the colony of ants has been traveled the whole paths in S, we can define the one-step neighbors for each node by using the clues of the previous ants. In this case, we define the neighbor of a node, which is a node that has the desired utilization (i.e. desired pheromone amount) from that node. We call the one-step neighbors table for each node as an ant-routing-table. By means of these ant-routing-tables, we construct the all possible nodesets in the coming phase.

**3.1.2. Constructing the nodesets**

In this phase, to construct the nodesets, we also use another colony of ants. Each ant can only start a node that has at least two neighbors. Firstly each ant starts initial nodeset including the start node and one of its neighbors and then adds one node at a time to its current partial nodeset. The choice of the node to be added to the current partial nodeset is in lexicographic order of the neighbors of start node, and adds the choice node to the current partial nodeset if the choice node is neighbor of all nodes in current partial nodeset.

An ant keeps adding nodes one-at-a-time to its current partial nodeset until the ant is unable to continue. This situation can rise if there is not possible to choose and/or add to the current partial nodeset. When these stopping criteria are satisfied, the ant has built a nodeset.

This process is repeated for all start nodes (i.e. all initial transitions. After all transitions has been considered to construct nodeset, the construction process stops and at that time the system has been discovered several nodesets.

**3.1.3. Defining the Frequent Nodesets**

In the final phase, we have to define the frequent nodesets/ patterns among the generated nodesets/ patterns. To determine the exact pheromone amount of all generated nodesets, a second scan of the dataset is required. The actual pheromone amount of all nodesets is computed during a second scan through the database. In this case, we used the prefix-tree to store the nodesets generated in phase 2.

**3.2. Prediction using SFM adaptation**

We now describe how our SFM adaptation algorithm to mine frequent patterns which may be used for our target application. The same mining process may be used to mine the frequent pattern. To narrow the search space and to fix our target criteria, we made the modification on the original SFM algorithm in first two phases. The adaptation of SFM algorithm is shown in figure 2.

As mentioned above, in the first phase we put the destination limit in the pheromone deposition. We consider only the pheromone deposition amount along the ants’ travel with the desired destination d (i.e. the desired dire outcome, e.g. recurrence outcome in breast cancer case), not consider the whole pheromone amount of ants’ travel. In this way, we can reduce the cardinality of the one-step neighbors for each node. This is no effect on frequent nodesets/patterns for our target applications due to 1) all frequent nodes and all one-step neighbors are relevant with our criteria such as both target prediction goal and pattern size, and 2) the rest of others are not associated with our criteria (i.e. non-desired outcome).

In nodeset construction phase, we made size limitation on nodeset. So it needs to construct the nodeset with desired size and is not necessary to construct the others with other size. It can reduce both time and search space complexity in not only nodeset construction but also frequent nodeset defining. In next section, we present how our proposed approach can efficiently apply and provide to risk prediction
application in medical domain.

Begin
ε = min-pheromone;
k = 0;
δ = nodeset-size;
∀ predefined nodeset ∈ S
{ Ant_{k+1} travel predefined nodesets;
  Update pheromone amount of trial followed by Ant_{k+1};
}
for (desired-destination-node d)
{dn} = Define one-step-neighbors of d;
∀ (node ∈ S ∧ node ∈ {dn})
{ Define one-step-neighbors;
  Create Ant-routing-table;
  if (#neighbors>1) {ns}=node;
}
∀ n_s ∈ {ns}
{Nodeset} = Nodeset_Generation(n_s, δ);
∀ Nodeset ∈ {Nodeset}
{ Calculate pheromone amount of Nodeset;
  Answer = {Nodeset | Nodeset.phero >= ε};
}
End

Figure 2. Adaptation of SFM algorithm

4. Empirical Study and Results

SFM is executed in C++ compiler on a 2.66GHz PentiumIV with 256MB of memory, running WindowsXP.

4.1. Dataset

Experiments were done using some public-domain datasets, obtained from Machine Learning Repository [1]. The datasets used in this work are Hepatitis database, which has 155 cases with 19 predicting attributes (six of them were continuous, and so were discretized); Dermatology database, has 358 cases with 34 predicting attributes (only one is continuous age, and so was discretized); Ljubljana breast cancer dataset, which has 282 cases with nine predicting attributes (all categorical); and Wisconsin breast cancer database, which has 683 cases with nine predicting attributes (all predicting attributes are continuous in the range of 1 to 10 and were discretized).

For our target prediction application, we map the original classes in database into only two classes called adverse and benign (i.e. normal and present / recurrence and non-recurrence / die and alive). Then we convert the chosen class-label into transaction database. For instance, Ljubljana breast cancer dataset with nine predicting attributes (all categorical) is converted into a transaction database with 53 numerical attributes.

4.2. Performance Measurements and Results

We evaluated the performance of the SFM in two aspects: execution times from frequent pattern mining point of view and accuracy rates from prediction point of view.

4.2.1. Execution Times

We evaluated the comparative performance of the SFM, and Apriori [3] using three fold cross-validations. Figure 3 reports the total execution times obtained running Apriori and our SFM on the same datasets described in above for various threshold values. Intuitively it can be seen that SFM can perform faster than the Apriori on the given datasets for various different threshold values.

We also compared the execution times for different number of cases with constant threshold 0.1 to prove how much efficiently operate with SFM. As shown in figure 4, SFM can perform better than the Apriori as scale-up testing. Therefore, SFM outperforms than Apriori, especially for a given support threshold as increasing the number of cases to mine.
4.2.2. Accuracy Rate

The cross-validation accuracy rate was also used as a common used performance measure among the several criteria that could be used to evaluate the predictive accuracy of discovered patterns. Accuracy rate is defined as the quotient between the number of correctly identify the adverse outcomes among those identify by the system. We evaluated comparative performance of the adaptation of SFM, Ant-Miner [9] and C4.5 [10] using three fold cross-validations.

The results obtained by the adaptation of SFM, Ant-Miner and C4.5 in all used datasets are summarized in the following table 1, and figure 5. These tables show the information about accuracy rate, and size of patterns used.

Table1. Results with the adaptation of SFM

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy rate</th>
<th>Size of patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hepatitis</td>
<td>88.20%</td>
<td>11</td>
</tr>
<tr>
<td>Dermatology</td>
<td>87.76%</td>
<td>15</td>
</tr>
<tr>
<td>Ljubljana</td>
<td>73.34%</td>
<td>9</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>96.03%</td>
<td>9</td>
</tr>
</tbody>
</table>

All the accuracy rates with respect to the given size of pattern are stated in table 1. The obtained results were compared with Ant-Miner and other machine learning methods C4.5 found in the literature, for the same datasets. Figure 5 shows the accuracy rates comparison among these algorithms using a 3-fold cross-validation.

Figure 5. Accuracy Results

By comparing these, our results are more accurate than the classical classification C4.5, and Ant-miner has the more accuracy than our approach in most cases. In general, our frequent pattern mining approach can be compare against with ordinary classification approach in accuracy. As mention in section 1, there is one main strength that we can avoid the weighting up process. Due to this, our approach may efficiently handle the large amount of data than the classical classification approach.

Figure 6. Accuracy rate vs. pattern size for adverse with SFM

The accuracies of the result patterns with respect to the size of the pattern are illustrated in figure 6. From that figure, we can examine that the accuracy of the pattern is very closely related to the size of the pattern used as well as the data nature itself.

As the patterns from frequent pattern mining approach are very user friendly, our approach can provide user understanding. Some of the patterns with size 4 for breast cancer patients in recurrence-events are as follows:

- Menopause = premeno ∧ Inv_nodes = 0-2 ∧ Node_Caps = no ∧ Breast = left
- Inv_nodes = 0-2 ∧ Node Caps = no ∧ Breast = left ∧ Irradiat = no
- Menopause = premeno ∧ Irradiat = no ∧ Inv_nodes = 0-2 ∧ Node_Caps = no

As mention above pattern, we can examine that the exploitation of frequent pattern mining approach is better descriptor of this application domain with reliable than the statistical model approach which offers reliable conclusions but lacks explanatory power in a human readable form.

As a result, risk predictions of the patients’ outcome by the frequent pattern mining approach achieve comparative accuracy performance over a given datasets with Ant-Miner and standard classification algorithm C4.5. Our approach can also avoid the weighting up process that is the one main computational operation in classification approach. Moreover, due to the user understandable the exploitation of frequent pattern mining approach are better descriptors with reliable than the statistical model for this application domain.

5. Conclusion and Future Work

Obviously, the potential of modern risk prediction techniques is yet to be fully exploited in the medical domain. We hope this paper will motivate clinicians to consider what could be done to obtain data that will permit medical risk to be forecasted more efficiently and accurately. We have presented a practical approach to predict the risk on the basis of a substantial amount of patient data. The reader should bear in mind that the primary aim of this paper is to
describe our approach to evaluate the patient risk prediction. And also realize that more informative clinical data needs to be collected for better risk predictions to be made. In this case our approach is very suitable because of the scalable with the immense amount of data.

Further experimentation over larger datasets in this domain e.g. gene expression of patient profiling and other domains e.g. agriculture is planned. And also, an extension of the SFM that efficiently exploits the pushing constraints (e.g. monotone constraints) in frequent pattern mining is underway.

REFERENCES


