Mining Unexpected Patterns by Decision Trees with Interestingness Measures

Rui-Dong Chiang  
Department of Computer Science and Information Engineering  
Tamkang University  
Taiwan, R.O.C.  
081863@mail.tku.edu.tw

Ming-Yang Chang  
Department of Obstetrics and Gynecology  
Chang Gung Memorial Hospital  
Taiwan, R.O.C.  
mychang@adm.cgmh.org.tw

Huan-Chao Keh  
Department of Computer Science and Information Engineering  
Tamkang University  
Taiwan, R.O.C.  
087173@mail.tku.edu.tw

Chien-Hui Chan  
Department of Computer Science and Information Engineering  
Tamkang University  
Taiwan, R.O.C.  
emmac@gmail.com

Abstract—We believe that unexpected, interesting patterns may provide researchers with different visions for future research. In this study, we propose an unexpected pattern mining conceptual model that uses decision trees to compare the recovery rates of two different treatments and to find patterns that contrast with the prior knowledge of domain users. In the proposed model, we define interestingness measures to determine whether the patterns found are interesting to the domain. By applying the concept of domain-driven data mining, we repeatedly utilize decision trees and interestingness measures in a closed-loop, in-depth mining process to find unexpected and interesting patterns. We use retrospective data from transvaginal ultrasound-guided aspirations to show that the proposed model can successfully compare different treatments using a decision tree, which is a new usage of decision trees.

Keywords—mining unexpected patterns; interestingness measures; treatment comparison; domain driven datamining

I. INTRODUCTION

In our experience, we acknowledge that medical experts usually feel more comfortable using statistical methods to conduct their studies. To measure the difference between treatment regimens, researchers usually divide patients into two groups [1-3]. For example, in [3], all patients were divided into two groups: “ethanol irrigation” and “ethanol retention,” respectively. They use Student’s t-tests to compare the means of the two treatments. Generally, a p-value less than 0.05 is regarded as statistically significant, which means that two sets of data are essentially different. Whereas the t-test is only suitable for comparing two treatments means, an analysis of variance can be used both to compare several means and in more complex situations [4]. However, it is still difficult to accurately describe the conditions of each group of patients using statistical methods in statements such as: when a patient’s “body weight < 69.8kg” and “age < 43,” treatment A is better than treatment B. Even though we can use regression analysis to estimate the relationships between variables [5], we still cannot easily determine the appropriate cutoff points for continuous variables.

Inappropriately partitioned data may result in finding no statistical significant difference between groups, particularly when the medical conditions of two groups of patients are very similar. Therefore, although we use statistical methods to find the cutoff point for continuous variables, it is necessary to repeatedly examine each cutoff point. Through a decision tree algorithm, we can easily divide patients into different groups and generate the different medical conditions of each group.

Traditional data mining emphasizes data-centered mining for interesting patterns. During the data mining process, individual user requirements and domain-related knowledge are less considered. Cao, et al. [6] advocated that the current algorithms, patterns, and produced models lack workability, actionability, and operable capability. Therefore, they proposed domain-driven data mining (D3M) to solve these problems. It has several key components: the constraints of the knowledge delivery environment, in-depth pattern mining, enhancing knowledge actionability, and closed-loop and iterative refinement [7]. In-depth pattern mining can discover interesting and actionable knowledge from a domain-specific viewpoint, and uncover deep data intelligence and interior business rules that traditional data mining cannot discover [6]. A closed-loop process means that the outputs of data mining can be fed back to change relevant factors or parameters in particular stages [8]. A pattern is actionable in a domain if it can be used to make decisions about future actions in the domain [9, 10]. Since clinical studies operate in complex environments, utilizing prior domain knowledge, constraints, and expert knowledge can enhance the capabilities and performance of data mining.

Surprisingness was first brought up by Silberschatz and Tuzhilin [11]. Surprising (or unexpected) patterns are interesting because they contradict a person’s existing knowledge or expectations and may suggest an aspect of the data that needs further study. According to a survey by Kontonasios, et al. [12], most studies that measure the unexpectedness of patterns focus on association rules. Moreover, to the best of our knowledge, none of the studies use interestingness measures to find unexpected patterns in decision trees.

The research aims of this study include the following:

- We define our interestingness measures based on the D3M concept to detect unexpected patterns that contrast with domain knowledge.
II. THE CONCEPT OF DOMAIN DRIVEN DATAMINING

Discovering interesting patterns in data is an important objective of data mining [13]. By applying interestingness measures, experts can find interesting patterns [14]. Most researchers divide interestingness measures into objective and subjective measures [11-13, 15-20]. The objective interestingness measures depend only on raw data; they are data-driven and domain-independent [17, 18]. Most objective interestingness measures are based on theories from statistics, probability, and information theory [21]. Subjective interestingness measures should consider both the data and users. To define subjective measures, we have acquired users’ insight on data and their background knowledge. Consequently, the subjective measures are user-driven and domain-dependent [17, 21].

Based on domain-driven data mining, Cao and Zhang [22] claimed that patterns extracted from a database must simultaneously satisfy technical and business interestingness. In other words, the patterns have to satisfy (1):

$$\forall x \in I, \exists P : x_{tech}(P) \land x_{biz}(P) \rightarrow x_{act}(P). \quad (1)$$

In (1), $I$ represents a set of items; $x$ is an item-set in a database $DB$ that consists of a set of transactions. A pattern $P$ (composed of item-set $x$) is an interesting pattern discovered in $DB$ through a modeling method. $Tech()$ is the technical interestingness measure that implies how interesting the pattern is from a technical viewpoint. It often utilizes specific technical metrics for data mining. $Biz()$ is the business interestingness measure, which indicates how interesting the pattern is from the users’ point of view. It is determined by the domain-oriented criteria accepted by domain users. $Act()$ represents the actionability of a pattern. Thus, if a pattern $P()$ satisfies both $tech()$ and $biz()$, it is both interesting and actionable.

To be integrated with the subjective and objective concepts described above, interesting domain knowledge should satisfy $tech_{obj}()$ (the technical objective interestingness measures), $tech_{sub}()$ (the technical subjective interestingness measures), $biz_{obj}()$ (the business objective interestingness measures), and $biz_{sub}()$ (the business subjective interestingness measures) [23]. Therefore, the output knowledge should satisfy (2):

$$\forall x \in I, \exists P : x_{tech_{obj}}(P) \land x_{tech_{sub}}(P) \land x_{biz_{obj}}(P) \land x_{biz_{sub}}(P) \rightarrow x_{act}(P). \quad (2)$$

III. THE UNEXPECTED KNOWLEDGE DISCOVERY MODEL

From a subjective point of view, an interesting pattern is either unexpected or actionable. Therefore, as described by Silberschatz and Tuzhilin [11], both unexpectedness and actionability are important for subjective interestingness measures. Unexpectedness is formalized with respect to background knowledge which either is explicitly defined by a user or represents common sense domain knowledge [12]. Unexpected patterns are interesting because they cannot be identified by previous knowledge and may suggest a particular status of the data that needs further study [21]. Furthermore, the patterns that contradict prior knowledge can be used to build theories about the domain [13].

To detect whether the extracted patterns are both unexpected and actionable, we propose an unexpected knowledge discovery model based on interestingness measures, as shown in Fig. 1. During data preprocessing, domain experts input and select variables for decision tree induction. A decision tree uses a tree branch structure to produce easily understandable classification rules. In current practice, the decision tree can be considered to be a fairly mature technique. The most famous decision tree algorithms are ID3 [24], C4.5 [25], CART [26], and CHAID [27]. Since the ID3 algorithm was only designed to handle categorical variables, continuous variables must be divided into discrete categorical values before the decision tree construction process. During the construction processes of the CART, C4.5, and CHAID models, continuous variables are automatically divided into discrete categorical values.

The CART model constructs binary trees: each internal node has exactly two outgoing edges [26]. A post-pruning process will sequentially collapse nodes that result in the smallest change in purity. Therefore, it can maintain the significant difference between two branches of a node. All of the process can be automatically completed by mining tools. Thus, we can interpret the corresponding rules directly from the tree. When dealing with continuous variables, a multi-branch decision tree might divide continuous variables into several ranges, which could make the decision tree difficult to interpret by domain experts (doctors). In this situation, the CART model is useful for producing binary cutoff points for continuous variables. Therefore, during the data decentralization stage, the CART algorithm is applied, and the cutoff points of the continuous variables will be generated during the decision tree construction, which can help to divide patients into different conditions.

After the decision tree is generated, we search potentially unexpected nodes in the decision tree. The corresponding rules of these nodes will be extracted and examined using interestingness measures. We define interestingness measures based on the D*M concept and use them to determine whether the node has domain interestingness. We are looking for potentially unexpected patterns that pass the interestingness examination. If the potentially unexpected interesting patterns satisfy the technical objectives and subjective interestingness measures
but not the business objective interestingness measure, the closed-loop and iterative refinement processes should be further commenced. We would feed the initial results back to the data decentralization stage, readjust the input parameters, and conduct mining again to obtain in-depth patterns. This process is repeated until no more unexpected patterns are found.

\[ \text{RecoveryRate(Treatment)} = \frac{\text{Treatment,success}}{\text{Treatment,success} + \text{Treatment,failure}} \]  \hspace{1cm} (3)

\section{B. Unexpected Nodes and Rule Interestingness Examination}

For each node \( i \) in the decision tree, the interestingness measures are formulated as follows:

\[ IM_{\text{tech.obj}}(i): \text{RecoveryRate}(\neg T_i) > \text{RecoveryRate}(T_i). \hspace{1cm} (4) \]

In (4), when \( \text{RecoveryRate}(T_i) > \text{RecoveryRate}(\neg T_i) \), it signifies that treatment \( T \) has a better curative effect than treatment \( \neg T \) in node \( i \); this is consistent with prior knowledge. In contrast, when \( \text{RecoveryRate}(T_i) < \text{RecoveryRate}(\neg T_i) \), treatment \( \neg T \) has a better curative effect than treatment \( T \) in node \( i \); this is in contrast to the prior knowledge. Therefore, this node may be a potentially unexpected node, it may need further investigation.

\[ IM_{\text{tech.sub}}(i): \left| \frac{T_i,\text{success}}{T_i,\text{success}} + \frac{T_i,\text{failure}}{\neg T_i,\text{failure}} \right| > 1 \text{ and } \left| \frac{\neg T_i,\text{success}}{\neg T_i,\text{success}} + \frac{T_i,\text{failure}}{\neg T_i,\text{failure}} \right| > 1. \hspace{1cm} (5) \]

In (5), we recognize that when there is only one \( T (\neg T) \) treatment in the node of the decision tree, this node cannot be compared; moreover, when there is merely one sample of \( T \) or \( \neg T \) in the tree node, the recovery rate for \( T \) or \( \neg T \) will either be 0% or 100%, due to relying on the outcome of a single sample; thus, it is “over fit.” As a result, such exceptional conditions should not be included in our discussion. When a node satisfies \( IM_{\text{tech.obj}} \) and \( IM_{\text{tech.sub}} \), then this node is a useful potentially unexpected node.

\[ IM_{\text{hi.obj}}(i): \left\{ \text{subnode}_i = \emptyset \text{ or } \neg \text{subnode}_{i,\text{left}} \rightarrow \neg \text{X} \text{ and } \text{subnode}_{i,\text{right}} \rightarrow \neg \text{X} \right\} \hspace{1cm} (6) \]

and \( \neg X',\text{success} \geq X',\text{success} \) and \( p < 0.05 \).

In (6), \( \text{subnode}_i \) is the node that immediately follows unexpected node \( i \). \( \text{subnode}_{i,\text{left}} \) and \( \text{subnode}_{i,\text{right}} \) represent the immediate left and right subnodes, respectively, of unexpected node \( i \); \( p_i \) represents the \( p \)-value of node \( i \). In \( \text{subnode}_{i,\text{left}} \) and \( \text{subnode}_{i,\text{right}} \), when the recovery rate from treatment \( \neg T \) is higher than that of treatment \( T \), it is denoted as \( \text{subnode}_{i,\text{left}} \rightarrow \neg T \) or \( \text{subnode}_{i,\text{right}} \rightarrow \neg T \). When the immediate subnodes of the potentially unexpected node are both unexpected, \( \neg T \) is the general case of these nodes. Thus, this potentially unexpected node can also be a terminal node. When the total amount of patients with treatment \( T \) is more than that of treatment \( \neg T \), a low total number of patients with treatment \( \neg T \) can distort \( \text{RecoveryRate}(\neg T_i) \). In other words, even if the recovery rate of treatment \( \neg T \) is better than that of treatment \( T \), the

\[ \text{RecoveryRate}(\neg T_i) > \text{RecoveryRate}(T_i). \]

\[ \text{RecoveryRate}(T_i) < \text{RecoveryRate}(\neg T_i). \]

\[ IM_{\text{tech.obj}}(i): \text{RecoveryRate}(\neg T_i) > \text{RecoveryRate}(T_i). \hspace{1cm} (4) \]

\[ IM_{\text{tech.sub}}(i): \left| \frac{T_i,\text{success}}{T_i,\text{success}} + \frac{T_i,\text{failure}}{\neg T_i,\text{failure}} \right| > 1 \text{ and } \left| \frac{\neg T_i,\text{success}}{\neg T_i,\text{success}} + \frac{T_i,\text{failure}}{\neg T_i,\text{failure}} \right| > 1. \hspace{1cm} (5) \]

\[ IM_{\text{hi.obj}}(i): \left\{ \text{subnode}_i = \emptyset \text{ or } \neg \text{subnode}_{i,\text{left}} \rightarrow \neg \text{X} \text{ and } \text{subnode}_{i,\text{right}} \rightarrow \neg \text{X} \right\} \hspace{1cm} (6) \]

and \( \neg X',\text{success} \geq X',\text{success} \) and \( p < 0.05 \).

\[ \text{RecoveryRate}(T_i) < \text{RecoveryRate}(\neg T_i). \]

\[ \text{RecoveryRate}(T_i) > \text{RecoveryRate}(\neg T_i). \]
patterns of the unexpected node may not be meaningful. Therefore, one still has to consider whether the number of recovering patients who received treatment \( \neg T \) is also more than that of \( T \). Our aim is to find patterns that contrast with domain experts’ knowledge: \( \neg T \). Therefore, the threshold in each node is that the sample number of \( \neg T \) should greater than \( T \). Also, in medical research, we often use \( t \)-tests to understand whether there is a significant difference among different groups. When the \( p_i < 0.05 \), it implies that a remarkable variance exists among groups. Therefore, we use a \( t \)-test to confirm whether there is significant difference between the recovery rate of treatment \( T \) and treatment \( \neg T \) at each node.

\[
IM_{biz\_sub}(i): \left[ X_i, success ] + |X_i, failure ] + |X_i, success ] + |X_i, failure ] \right] > \text{expertDefine.} \\
\tag{7}
\]

In (7), expertDefine represents the threshold value determined by domain experts. The business subjective interestingness, \( IM_{biz\_sub}(i) \), serves as the minimum threshold value in each node. According to the prevalence of a disease and the sample size, domain experts set a minimum amount of patients for each node so that they can be meaningful. In other words, the amount of patients for an unexpected node must satisfy the threshold value, so it can pass \( IM_{biz\_sub} \). In this stage of our study, we do not discuss the threshold of \( IM_{biz\_sub} \). Therefore, our algorithm does not include \( IM_{biz\_sub} \).

IV. MINING UNEXPECTED PATTERNS FOR ENDOMETRIOSIS

Endometriosis is one of the most common problems in gynecology [28]. It is defined as the presence of endometrial-like tissue outside the uterus [29]; it causes pain (e.g., pelvic pain, dysmenorrhea, and dyspareunia) and infertility, although 20–25% of patients are asymptomatic [30]. The associated symptoms of endometriosis can impact a woman’s quality of life in many ways. Endometriosis should be considered to be a chronic disease characterized by high recurrence rates [30]. In addition, women who have their ovaries removed still risk further recurrence of endometriosis. The treatment for endometriosis should be chosen by each individual patient, depending on symptoms, age, and fertility [30]. In other words, treatment depends on how severe the patients’ symptoms are and whether the patient wants to get pregnant. Transvaginal ultrasound-guided aspiration with ethanol sclerotherapy is an alternative treatment that can minimize surgical risks and effectively decrease cyst size and the related symptoms of cyst compression [31, 32]. Ideally, if treatment time is sufficient, this procedure may result in total regression of the inflammatory cyst [2].

A. Materials and Settings

In this study, a retrospective review of 188 aspirations was collected from 2001 through 2010 at Taipei Chang Gung Memorial Hospital. These endometriosis patients with postsurgical recurrence of pelvic cysts received transvaginal ultrasound aspiration with 95% ethanol sclerotherapy at the outpatient gynecological department. The Institutional Review Board (IRB) of Chang Gung Medical Foundation has approved this study. To preserve patient confidentiality, direct patient identifiers were not collected. Therefore, each registry contains the record of therapy and follow-up episodes rather than the record of the patient. Our patients received 7 to 10 minutes ethanol instillation or total retention. Post-operative hormonal therapy was administered for 3-6 months in response to the patient’s will to extend the treatment effect, or based on the decision of the physician. Further repeat surgical interventions with either repeat sclerotherapy or abdominal operations were recorded. The immediate preoperative data of repeat operation patients were characterized as the end point data of the patient. They were followed with vaginal ultrasounds, CA-125 determinations, and pain score records every 3 to 6 months for at least one year. Twelve-month recovery was defined as: (a) any pregnancy achieved, (b) no repeat surgery, and (c) no cysts, if any, with a diameter of 3.0 cm or larger.

In this paper, domain experts recognize that post-operative hormonal therapy can extend the treatment effect. Therefore, the recovery rate of a treatment with post-operative hormonal therapy (\( T \)) should higher than that of a treatment with no post-operative hormonal therapy (\( \neg T \)). Accordingly, the patients were categorized into four groups as the target categories of the decision tree:

1) \( T, success \): Successful recovery with post-operative hormonal therapy.
2) \( T, failure \): Recovery failure with post-operative hormonal therapy.
3) \( \neg T, success \): Successful recovery with no post-operative hormonal therapy.
4) \( \neg T, failure \): Recovery failure with no post-operative hormonal therapy.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA-125</td>
<td>Preoperative CA-125 level</td>
</tr>
<tr>
<td>Uterus length</td>
<td>Length of preoperative uterus, mm</td>
</tr>
<tr>
<td>Uterus volume</td>
<td>Volume of preoperative uterus, mm³</td>
</tr>
<tr>
<td>Cyst Size</td>
<td>Total size of preoperative cysts, cm</td>
</tr>
<tr>
<td>Cyst Number</td>
<td>Total amount of preoperative cysts</td>
</tr>
<tr>
<td>Cyst Content</td>
<td>Type 1 / Type 0</td>
</tr>
<tr>
<td>Patient group (Target)</td>
<td>1: successful recovery with post-operative hormonal therapy 2: recovery failure with post-operative hormonal therapy 3: successful recovery with no post-operative hormonal therapy 4: recovery failure with no post-operative hormonal therapy.</td>
</tr>
</tbody>
</table>
Using preliminary statistics for the 12-month follow-up period, it was found that the recovery rate of Group $\neg T (N = 112)$ and Group $T (N = 76)$ were 48.68% and 49.11% ($p > 0.05$), respectively, which shows that the regimens of each group has no significant effect on curative outcomes. To further identify the factors that influence the curative outcomes, we adopted a decision tree to generate the cutoff point of numeric data for grouping. The original records contained more than 80 preoperative and postoperative examination follow-up fields; however, most were null values. Therefore, via discussion with domain experts, the variables that physicians consider influential on curative effect were selected to build the decision tree (for more detail, please refer to TABLE I).

B. Results

![Fig. 2. Resulting decision tree](image)

Fig. 2 is the resulting tree that we utilized for the original setting of the CART algorithm, we find that node 3 is potentially unexpected. The result of the interestingness examination of this node is shown in TABLE II. As shown in TABLE II, in the right subtree of the root, all the other nodes depict treatment $T$ as having a better curative effect than treatment $\neg T$; this is consistent with prior knowledge. Therefore, the right subtree of the root does not need to proceed with in-depth pattern mining. On the other hand, in the left subtree of the root, node 3 satisfies $IM_{tech\_obj}$ and $IM_{tech\_sub}$ but not $IM_{biz\_obj}$. Therefore, the in-depth pattern mining strategy is used to further analyze the data. According to the closed-loop method, we utilize the feedback results to adjust the parameters.

Our experts consider that different type of cysts content may result in different curative effect. Thus, we reselect data by the condition that the cyst content = 1 to grow a new tree. The resultant tree is shown in Fig. 3. We find that node 3 and node 5 are potentially unexpected. The results of the interestingness examination of these nodes are shown in TABLE III.

As shown in TABLE III, the corresponding rules of node 3 and 5 in Fig. 3 pass the interestingness examinations of $IM_{tech\_obj}$, $IM_{tech\_sub}$, and $IM_{biz\_obj}$. Therefore, the corresponding rules of node 3 and 5 are unexpected and interesting patterns we are looking for. The experts agreed that these patterns are interesting and unexpected. However, they tend to be reserved. They expect more samples to support the threshold of business subjective interestingness. Therefore, further research needs to be conducted to define the appropriate threshold.

### Table II: Interestingness Check of Potentially Unexpected Rule for Fig. 2

<table>
<thead>
<tr>
<th>Node</th>
<th>Rule</th>
<th>$IM_{tech_obj}$</th>
<th>$IM_{tech_sub}$</th>
<th>$IM_{biz_obj}$</th>
<th>$IM_{biz_sub}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Retention = 7-10 minutes and CA-125 &lt;= 52.75 cm and Uterus volume &lt;= 67.319 cm$^3$ $\neg T$</td>
<td>Pass</td>
<td>Pass</td>
<td>Pass</td>
<td>Pass</td>
</tr>
</tbody>
</table>

### Table III: Interestingness Check of Fig. 3

<table>
<thead>
<tr>
<th>Node</th>
<th>Rule</th>
<th>$IM_{tech_obj}$</th>
<th>$IM_{tech_sub}$</th>
<th>$IM_{biz_obj}$</th>
<th>$IM_{biz_sub}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Retention = 7-10 minutes and Cyst Content = 1 and Cyst Size &lt;= 5.05 cm and Uterus length &lt;= 5.44 cm $\neg T$</td>
<td>Pass</td>
<td>Pass</td>
<td>Pass</td>
<td>Pass</td>
</tr>
<tr>
<td>5</td>
<td>Retention = 7-10 minutes and Cyst Content = 1 and &gt; 5.05 cm and CA-125 &lt;= 50.06 $\neg T$</td>
<td>Pass</td>
<td>Pass</td>
<td>Pass</td>
<td>Pass</td>
</tr>
</tbody>
</table>

Fig. 3. Resultant decision tree with alcohol retention 7-10 minutes and content type = 1.

108
V. CONCLUSIONS

Since traditional decision trees compare the differences between sibling nodes, they usually use binary targets to induce classification trees. In this paper, we use four target categories to induce classification trees. Using a pure decision tree without interestingness measures, users need to calculate each individual node and search the whole tree manually to find unexpected patterns. In this study, the interestingness measures are design for comparing treatment at individual nodes, and retrieving the unexpected patterns. Thus, we can conduct a comparison at each individual node automatically. Also, we use retrospective data on transvaginal ultrasound-guided aspirations to show that the proposed model can compare different treatments and retrieve an unexpected pattern by using a decision tree. Since clinical studies are conducted in complex environments, we believe that it is important to develop an interactive mining model that involves prior domain knowledge, constraints, and expert knowledge. For the future research, this method can be used in industrial processes or other fields.

REFERENCES


