Some Properties of Most Relevant Explanation

Changhe Yuan

Uncertainty Reasoning Laboratory
Department of Computer Science and Engineering
Mississippi State University, Mississippi State, MS 39762, USA
cyuan@cse.msstate.edu

Abstract. This paper provides a study of the theoretical properties of Most Relevant Explanation (MRE) [12]. The study shows that MRE defines an implicit soft relevance measure that enables automatic pruning of less relevant or irrelevant variables when generating explanations. The measure also allows MRE to capture the intuitive phenomenon of explaining away encoded in Bayesian networks. Furthermore, we show that the solution space of MRE has a special lattice structure which yields interesting dominance relations among the candidate solutions.

1 Introduction

Bayesian networks have become the basis for many diagnostic expert systems. However, many of these systems focus on disambiguating single-fault diagnostic hypotheses because it is hard to generate “just right” multiple-fault hypotheses that contain only the most relevant faults. Maximum a Posteriori (MAP) assignment and Most Probable Explanation (MPE) are two explanation methods for Bayesian networks that find a complete assignment to a set of target variables as the best explanation for given evidence and can be applied to generate multiple-fault hypotheses. A priori, the set of target variables is often large and may be in tens or hundreds for a real-world diagnostic system. Given that so many variables are involved, even the best solution by MAP or MPE may have an extremely low probability, say in the order of $10^{-6}$. It is hard to make any decision based on such hypotheses.

Recently, Yuan and Lu [12] propose an approach called Most Relevant Explanation (MRE) to generate explanations containing only the most relevant target variables for given evidence in Bayesian networks. Its main idea is to traverse a trans-dimensional space containing all the partial instantiations of the target variables and find one instantiation that maximizes a relevance measure called generalized Bayes factor [4]. The approach was shown in [12] to be able to find concise and intuitive explanations. This paper provides a study of the theoretical properties of MRE. The study shows that MRE defines an implicit soft relevance measure that enables automatic pruning of less relevant or irrelevant variables when generating explanations. The measure also allows MRE to capture the intuitive phenomenon of explaining away encoded in Bayesian networks. Furthermore, we show that the solution space of MRE has a special lattice structure which yields interesting dominance relations among the candidate solutions.
2 Related Work

Figure 1 shows a probabilistic circuit and its corresponding diagnostic Bayesian network [10, 12]. Suppose we observe that current flows through the circuit, which means that nodes Input and Total Output in the Bayesian network are both in the state “current”. The task is to diagnose the system and find the best fault hypotheses. Based on our knowledge of the domain, we know there are three scenarios that most likely lead to the observation: (1) A is defective; (2) B and C are defective; and (3) B and D are defective.

There are many existing approaches for generating multivariate explanations in Bayesian networks. However, they often generate explanations are either too complex (overspecified) or too simple (underspecified). They fail to find explanations that contain the most relevant target variables.

Many existing methods make simplifying assumptions for computational convenience and assume that the target variables are mutually exclusive and collectively exhaustive, and there is conditional independence of evidence given any hypothesis [6, 8]. Therefore, we only need to consider the singleton explanations. However, such singleton explanations may be underspecified and are unable to fully explain given evidence. For the running example, the posterior probabilities of A, B, C, and D failing independently are 0.391, 0.649, 0.446, and 0.301 respectively. Therefore, (¬B) is the best singleton explanation (We use a variable and its negation to stand for its “ok” and “defective” states respectively). However, B alone does not fully explain the evidence. C or D has to be involved.

For a domain in which target variables are interdependent, multivariate explanations are often more natural for explaining given observations. However, existing methods often produce hypotheses that are overspecified. MAP finds a configuration of a set of target variables with the maximum joint posterior probability given partial or complete evidence on the other variables. For the running example, if we set A, B, C and D as the target variables, MAP will
find \((A \land \neg B \land \neg C \land D)\) as the best explanation. However, given that \(B\) and \(C\) are faulty, \(A\) and \(D\) are somewhat redundant for explaining the observation. MPE finds explanations with even more variables. Several other approaches make use of the conditional independence relations encoded in Bayesian networks to identify the best multivariate explanations [11]. They will find the same explanation as MAP because all the target variables are dependent on the evidence. Yet several other methods measure the quality of explanation candidates using likelihood of the evidence [1]. Unfortunately they will overfit and choose \((\neg A \land \neg B \land \neg C \land \neg D)\) as the best explanation, because the likelihood of the evidence given that all the target variables fail is almost 1.0.

There have been efforts trying to generate more appropriate explanations. Henrion and Druzdzel [7] assume that a system has a set of pre-defined scenarios as potential explanations and find the scenario with the highest posterior probability. Flores et al. propose to grow an explanation tree incrementally by branching the most informative variable at each step while maintaining the probability of each explanation above certain threshold [5]. Nielsen et al. [9] use a different measure called causal information flow to grow the explanation trees. Because the explanations in the trees have to branch on the same variable(s), they may still contain redundant variables. Finding more concise hypotheses also have been studied in model-based diagnosis [2]. The approach focus on truth-based systems and is not easily generalized to deal with Bayesian networks.

3 Most Relevant Explanation

Yuan and Lu recently propose a method called Most Relevant Explanation to automatically identify the most relevant faults for given evidence in Bayesian networks [12]. First, explanation in Bayesian networks is formally defined as follows.

**Definition 1.** Given a set of target variables \(X\) in a Bayesian network and evidence \(e\) on the remaining variables, an explanation \(x_{1:k}\) for the evidence is a partial instantiation of the target variables, i.e., \(X_{1:k} \subseteq X\) and \(X_{1:k} \neq \emptyset\).

Most Relevant Explanation (MRE) is then defined as follows.

**Definition 2.** Let \(X\) be a set of target variables, and \(e\) be the evidence on the remaining variables in a Bayesian network. Most Relevant Explanation is the problem of finding an explanation \(x_{1:k}\) that has the maximum Generalized Bayes Factor score \(GBF(x_{1:k}; e)\), i.e.,

\[
MRE(X, e) \equiv \arg \max_{x_{1:k}} GBF(x_{1:k}; e),
\]

where \(GBF\) is defined as

\[
GBF(x_{1:k}; e) \equiv \frac{P(e|x_{1:k})}{P(e|x_{1:k})},
\]
Potentially, MRE can use any measure that provides a common ground for comparing the partial instantiations of the target variables. GBF is chosen because it is shown to provide a plausible measure for representing the degree of evidential support in recent studies on Bayesian confirmation theory [3, 4].

MRE was shown to be able to generate precise and concise explanations for the running example [12]. The best explanation according to MRE is:

\[ GBF(\neg B, \neg C; e) = 42.62. \] (3)

For simplicity we often omit \( e \) and write \( GBF(\neg B, \neg C) \). \((\neg B, \neg C)\) is a better explanation than both \((\neg A)\) (39.44) and \((\neg B, \neg D)\) (35.88), because its prior and posterior probabilities are both relatively high. Therefore, MRE seems able to automatically identify the most relevant target variables and states as its explanations for given evidence.

4 A Theoretical Study

4.1 Theoretical properties of MRE

We now discuss several theoretical properties of MRE. Since MRE relies heavily on GBF in generating its explanations, it is not surprising that these properties are mostly originated from the GBF measure. The proofs of these properties are omitted due to space limit.

The most important property is that MRE is able to weigh the relative importance of multiple variables and only include the most relevant variables in explaining the given evidence. The degree of relevance is evaluated using a measure called conditional Bayes factor (CBF) encoded in the GBF measure and defined as follows.

**Definition 3.** Conditional Bayes factor for hypothesis \( y_{1:m} \) given evidence \( e \) conditional on \( x_{1:k} \) is defined as

\[ CBF(y_{1:m}; e | x_{1:k}) \equiv \frac{P(e|y_{1:m}, x_{1:k})}{P(e|\bar{y}_{1:m}, x_{1:k})}. \] (4)

We also define belief update ratio \( r(x_{1:k}; e) \) of \( x_{1:k_1} \) given \( e \) as

\[ r(x_{1:k}; e) \equiv \frac{P(x_{1:k}|e)}{P(x_{1:k})}. \] (5)

Then, we have the following theorem.

**Theorem 1.** Let \( x_{1:k} \) be an explanation with \( r(x_{1:k}; e) \geq 1.0 \), and \( y \) be a state of a variable \( Y \) whose conditional Bayes factor given \( x_{1:k} \) is less than or equal to the inverse of the belief update ratio of the alternative explanations \( x_{1:k_1} \), i.e.,

\[ CBF(y; e|x_{1:k}) \leq \frac{1}{r(x_{1:k}; e)}. \] (6)
the following holds

\[ GBF(x_1:k \cup \{y\}; e) \leq GBF(x_1:k; e). \quad (7) \]

Therefore, \( CBF(y, e|x_1:k) \) provides a soft measure on the relevance of a new variable state and can be used to decide whether or not to include it in an existing explanation. \( GBF \) also encodes a threshold, the inverse belief update ratio of alternative explanations \( x_1:k \) given \( e \), which provides a threshold on how important the remaining variables should be in order to be included in the current explanation.

Note that we focus on the explanations with belief update ratio greater than or equal to 1.0. We believe that an explanation whose probability decreases given the evidence is unlikely to be a good explanation for the evidence.

Theorem 1 has several intuitive corollaries. First, the following corollary shows that, for any explanation \( x_1:k \) with belief update ratio greater than or equal to 1.0, adding any independent variable to the explanation will decrease its explanatory score measured by \( GBF \) [12].

**Corollary 1.** Let \( x_1:k \) be an explanation with \( r(x_1:k; e) \geq 1.0 \), and \( y \) be a state of variable \( Y \) independent from variables in \( x_1:k \) and \( e \). Then

\[ GBF(x_1:k \cup \{y\}; e) \leq GBF(x_1:k; e). \quad (8) \]

Therefore, adding an irrelevant variable dilutes the explanatory power of an explanation. MRE is able to automatically prune such variables. This is clearly a desirable property.

Corollary 1 requires the additional variable \( Y \) to be independent from both \( X_1:k \) and \( E \). The assumption is rather strong. The following corollary relaxes it to be that \( Y \) is conditionally independent from \( E \) given \( X_1:k \) and shows the same result still holds.

**Corollary 2.** Let \( x_1:k \) be an explanation with \( r(x_1:k; e) \geq 1.0 \), and \( y \) be a state of a variable \( Y \) conditionally independent from variables in \( e \) given \( x_1:k \). Then

\[ GBF(x_1:k \cup \{y\}; e) \leq GBF(x_1:k; e). \quad (9) \]

Corollary 2 is a slightly more general result than corollary 1 and captures the intuition that conditionally independent variables add no additional information to an explanation in explaining given evidence, even though the variable may be marginally dependent on the evidence.

The above theoretical results can be partially verified using the running example. For example,

\[ GBF(\neg B, \neg C) > GBF(\neg B, \neg C, A) \& GBF(\neg B, \neg C, D) > GBF(\neg B, \neg C, A, D). \]

The results suggest that \( GBF \) has the intrinsic capability to penalize higher-dimensional explanations and prune less relevant variables.
4.2 Explaining away

One unique property of Bayesian networks is that they can model the so called explaining away phenomenon using the V structure, i.e., a single variable with two or more parents. This structure intuitively captures the situation where an effect has multiple causes. Observing the presence of the effect and one of the causes reduces the likelihood of the presence of the other causes. It is desirable to capture this phenomenon when generating explanations.

MRE seems able to capture the explaining away effect using CBF. Again for the running example, (¬B, ¬C) (42.62) and (¬A) (39.44) are both good explanations for the evidence by themselves. When given only e (the effect), $\text{CBF}(\neg A; e)$ is equal to $\text{GBF}(\neg A; e)$ and is rather high. However, $\text{CBF}(\neg A, e|\neg B, \neg C)$ is equal to 1.03, which means when (¬B, ¬C) (one of the causes) is also observed, the CBF of ¬A becomes rather low. Therefore, CBF measures how relevant a new variable is to an existing explanation. In a situation of explaining-away, if one of the causes is already present in the current explanation, other causes become much less relevant for explaining the evidence according to CBF. This clearly agrees with the theoretical finding in Theorem 1.

4.3 Dominance relations

MRE has a solution space with an interesting lattice structure similar to the graph in Figure 2 for three binary target variables. The graph contains all the partial instantiations of the target variables. Two explanations are linked together if they only have a local difference, meaning they either have the same set of variables with one variable in different states, or one explanation has one fewer variable than the other explanation with all the other variables being in the same states.

We now define several concepts to capture the dominance relations among these potential solutions implied by Figure 2. The first concept is strong dominance.

**Definition 4.** An explanation $x_{1:k}$ strongly dominates another explanation $y_{1:m}$ if and only if $x_{1:k} \subseteq y_{1:m}$ and $\text{GBF}(x_{1:k}) \geq \text{GBF}(y_{1:m})$.

If $x_{1:k}$ strongly dominates $y_{1:m}$, $x_{1:k}$ is clearly a better explanation than $y_{1:m}$, because it not only has a no-worse explanatory score but also is more concise.
We only need to consider \( x_{1:k} \) in generating multiple top MRE explanations. The second concept is *weak dominance*.

**Definition 5.** An explanation \( x_{1:k} \) weakly dominates another explanation \( y_{1:m} \) if and only if \( x_{1:k} \supset y_{1:m} \) and \( GBF(x_{1:k}) > GBF(y_{1:m}) \).

In this case, \( x_{1:k} \) has a strictly larger \( GBF \) score than \( y_{1:m} \), but the latter is more concise. It is possible that we can include them both and let the decision makers to decide whether they prefer higher score or conciseness. However, we believe that we only need to include \( x_{1:k} \), because its higher \( GBF \) score indicates that the extra variable states are relevant to explain given evidence and should be included in the explanation.

Based on the two kinds of dominance relations, we define the concept of *minimal*.

**Definition 6.** An explanation is minimal if it is neither strongly nor weakly dominated by any other explanation.

In case we want to find multiple top explanations, we only need to consider the minimal explanations, because they are the most representative ones. Let us look at the running example again to illustrate the idea. The explanations in Table 1 have the highest \( GBF \) scores.

<table>
<thead>
<tr>
<th>Explanation</th>
<th>GBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \neg B, \neg C )</td>
<td>42.62</td>
</tr>
<tr>
<td>( A, \neg B, \neg C )</td>
<td>42.15</td>
</tr>
<tr>
<td>( \neg B, \neg C, D )</td>
<td>39.93</td>
</tr>
<tr>
<td>( \neg B, \neg D )</td>
<td>35.88</td>
</tr>
</tbody>
</table>

*Table 1.* The top solutions ranked by \( GBF \). The solutions in boldface are the top minimal solutions.

If we select the top three explanations solely based on \( GBF \), we will obtain \( (\neg B, \neg C), (A, \neg B, \neg C), \) and \( (\neg B, \neg C, D) \), which are rather similar to each other. However, since \( (A, \neg B, \neg C), (\neg B, \neg C, D) \), and \( (A, \neg B, \neg C, D) \) are strongly dominated by \( (\neg B, \neg C) \), we should only consider \( (\neg B, \neg C) \) out of these four explanations. Similarly, \( (\neg A, B) \) and \( (\neg A, C) \) are strongly dominated by \( (\neg A) \). These dominated explanations should be excluded from top MRE solution set. In the end, we get the set of top explanations shown in boldface in Table 1, which are clearly more diverse and representative than the original set.

### 5 Concluding Remarks

In this paper, we studied the theoretical properties of Most Relevant Explanation (MRE) [12]. Our study shows that MRE defines an implicit soft relevance measure that enables automatic pruning of less relevant variables when generating explanations and, consequently, can automatically prune (conditionally)
independent variables. The relevance measure also seems able to capture the explaining away phenomenon often encoded in Bayesian networks. Furthermore, we identified two dominance relations among the potential explanations that are implied by the structure of MRE solution space. These relations allow a more representative set of top solutions to be found.

Acknowledgements This research was supported by the National Science Foundation grant IIS-0842480.

References