Prediction and Explanation by Combined Model-Based and Case-Based Reasoning

Hoda Nikpour

Nowegian University of Science and Technology, Department of Computer and Information Science
hodan@idi.ntnu.no
https://www.ntnu.edu/idi

1 Introduction

Case-based reasoning is suitable for capturing and reusing human experiences for complex problem solving, and has earlier shown its success also in the oil drilling domain[1]. However, a pure CBR system suffers from the inability to justify a solution - beyond referring to the best matching case or cases. Further, CBR represents in itself a knowledge-lean method for case retrieval. A model of general domain knowledge would enable cases to be matched based on semantic rather than purely syntactic criteria. Hence, a general domain model combined with CBR will enable the system to generate targeted explanations for the user as well as for its internal reasoning process. Earlier work in our group have addressed this problem by combining CBR with a semantic network of multi relational domain knowledge [2], which implementation is called TrollCreek. A problem with that method was the lack of a formal basis for the semantic network that was used, which made the inference processes within the network difficult to develop and less powerful than wanted. The need for a more formal treatment of uncertainty leads to some initial investigations into how a Bayesian Network (BN) model could be incorporated [3, 4].
Bayesian Network has shown its feasibility to build probabilistic models without introducing unrealistic assumptions of independencies [4]. The probability distribution provided by BN enables the conditioning over any of the variables and supports any direction of reasoning [5]. Also, the Bayesian Networks framework includes an inference engine, which, given some evidence, is capable of updating its beliefs [6]. Moreover, the nature of Bayesian Networks allows for some explanations to be given regarding the reasoning process [4]. All these make BNs a proper candidate for my PhD work.

2 Related Work

The literature study done so far addresses the two main aspects of this project, namely the combination of CBR and a multi-relational domain knowledge model, and the combination of CBR with a BN.
The TrollCreek system is an implementation based on the Creek architecture for knowledge-intensive case-based problem solving and learning, targeted at

In Proceedings of the ICCBR 2016 Workshops. Atlanta, Georgia, United States of America

addressing problems in open and weak-theory domains [2]. In TrollCreek, case-based reasoning is supported by a model-based reasoning component that utilizes general domain knowledge. The model of general knowledge constitutes a combined frame system and semantic network, where each node and each link in the network is explicitly defined in its own frame object. Each node in the network corresponds to a concept in the knowledge model, and each link corresponds to a relation between concepts. A frame represents a node in the network, i.e. a concept in the knowledge model. Each concept is defined by its relations to other concepts, represented by the list of slots in the concept's frame definition. A case is also viewed as a concept (a situation-specific concept), and hence it is a node in the network, linked into the rest of the network by its case features. Fig. 1 illustrates the three main types of knowledge in TrollCreek, a top-level ontology of generic, domain-independent concepts, the general domain knowledge, and the set of cases. The case retrieval process in TrollCreek is a two-step

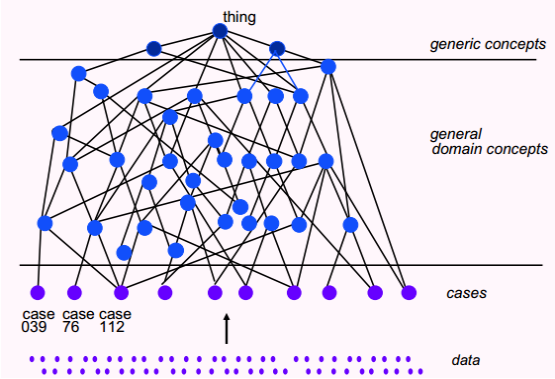

Fig. 1. The three-level Creek knowledge structure. Uses general domain knowledge as a knowledge model.

process, in line with the two-step MAC-FAC model [7], in which the first step is a computationally cheap, syntactic matching process, and the second step is a knowledge-based inference process that attempts to create correspondences between structured representations in the semantic network.
Additionally, a study over a number of literatures which are investigated the combination of CBR and BN and their applications has been done.

3 Research Plan

While the existing TrollCreek ontology may be useful for human interpretation, its current inference methods are too weak to address the needs of automated data analysis and decision support targeted by my Ph.D. A new approach is called for, which defines three central objectives of this project:

– The design of a domain model representation, with suitable inference methods, which captures the essential content of the existing ontology.

  - In this level of study, Bayesian Network has been utilized as a knowledge model with a strong inference capability. But in this way, the knowledge that are added to the system by the multi relational knowledge models, will be lost. Therefore, we are looking for a way to incorporate CBR with multi relational knowledge and causal knowledge models.
- Integration of a case representation and CBR method, that is able to utilize the general domain ontology in its case modeling and reuse process.
- Adaptation of machine learning methods that builds abstract process signatures from data with the help of the ontology and the case based reasoning.

The existing ontology and CBR methods have been developed within a development environment that is now obsolete. An open source environment or a combination of environments will be studied, assessed and adapted to our needs. Candidates are MyCBR, Colibri Studio, and Protege.
We will use oil well drilling problems as our application domain and rely on the field expert evaluation as our evaluation method.

4 State of the project

In the line of my PhD plan as a start point of combining CBR and BN research, the prediction of root causes of failures and the generation of explanations given the observed symptoms or errors are studied.
The main structure of a Bayesian network has been designed in order to express the elements' relations and calculate the updated beliefs based on the prior probabilities assigned by a field expert. The domain concepts are presented by nodes and their causal relations are shown by arrows. A parent causes a child, and each node represents the current belief of the network given its parents. Our approach in the first place, views the BN as a different type of, and a replacement for the knowledge model in TrollCreek (BN-Creek). Then, integrates TrollCreek case retrieval results with the BN-Creek results to get benefit from the other type of relations that been considered in TrollCreek. Fig. 2 depicts the graphical structure for the proposed approach. The filled and not filled circles are indicators of Bayesian network nodes and cases, respectively. TrollCreek and the present approach are extracting the cases from the raw data, but the main difference between them is their knowledge models. TrollCreek uses a multi relational semantic based knowledge model while the new approach uses a probabilistic causal model as its knowledge model.
The field expert's knowledge has been exploited to create the aforementioned BN. The causal relations between the oil well drilling process' concepts were identified by the expert and were used as the prior probabilities of the BN.
The main task in this study is to answer the query of: "What is the whole probability distribution over variable X given evidence e?". In other words, the most plausible causes of the failure under study, given some observations, is desired. In our approach, the mentioned query will be answered in the following three steps.
Step one: This step utilizes BN to calculate a temporal probability distribution of the new case.

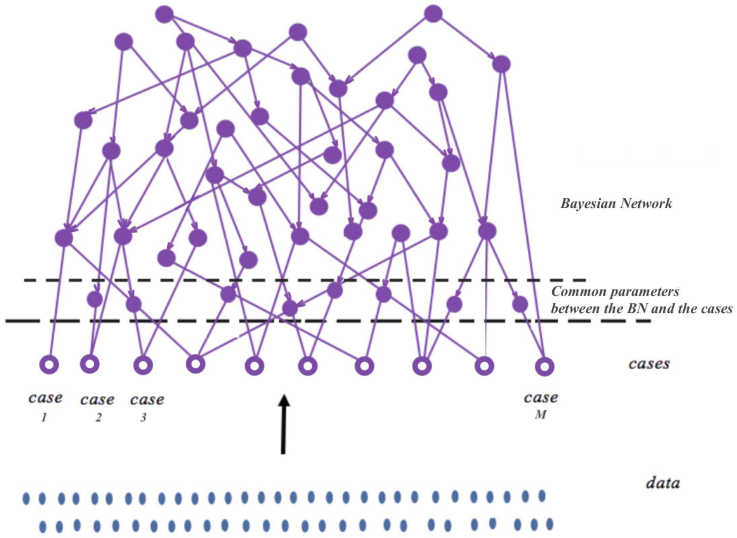

Fig. 2. The BNCreek knowledge structure

By creating a new case, a copy of the domain's prior Bayesian network is assigned to it. The inference process is started, by applying each of the observed concepts as evidence in the network. Then the network's beliefs will be updated given those evidence and the result will be shown as the posterior distribution (PD) of the network, corresponding to that specific case.

Step two: In this step, we utilize the CBR's capability of employing the past experiences, aimed to improve the BN's accuracy in suggesting the root causes. TrollCreek is used to retrieve the most similar cases to the new one. The best-matched case is considered and an impact factor is assigned to its recorded PD, based on its similarity degree.

TrollCreek uses MAC-FAC method [7] to retrieve the cases. As the MAC phase each of the findings from the testing case are compared to all the findings from the retrieved case aimed to find similar findings as many as possible. The Eq.1 illustrates the similarity assessment formula:

$$sim(C_{IN}, C_{RE}) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} sim(f_i, f_j) * relevancefactor_{f_j}}{\sum_{j=1}^{m} relevancefactor_{f_j}} \qquad (1)$$

In Eq.1 the C_{IN} stands for the under study case and and C_{RE} demonstrates the retrieved case. n and f_i, m and f_j are the number of findings and the finding's number in the C_{IN} and C_{RE}, respectively. The $sim(C_{IN}, C_{RE})$ is equal to 1 if $f_i = f_j$, otherwise it's value would be 0. The relevance factor is a number that combines the predictive strength and importance of a feature for a stored case and comes from the expert [2].

The FAC phase considers the paths in the semantic network that represent relation sequences between un-identical features. Based on a method for calculating

the closeness between two features at each end of such a sequence, the two features are given a local similarity score.

Step three: This step integrates the probability distributions from the first two steps and calculates the new case's finalized probability distribution. In other words, in this step we have added the CBR's capability in employing the past experiences, to improve the BN's accuracy in suggesting the root causes. The result of step three is the system's outcome.

Eq.2 integrates the effect of the pure BN and CBR from the first two steps and generates the finalized posterior distribution for the new case.

$$PP_{jf} = \frac{\sum_{i=1}^{k} PP_{ji} * \alpha_i}{\sum_{i=1}^{k} \alpha_i} \tag{2}$$

In Eq.2 the PP stands for the posterior probabilities which are the elements of the Posterior distribution. The $0 < \alpha < 1$ is the impact factor that is larger for the cases with higher similarity. The 'k' is the number of PDs that are integrated together and would be higher than two in a situation that the expert wants to involve the effect of less matched cases. 'j' and 'i' a're the indicators of a specific PP in a PD and the PD's number, respectively. Consequently, the PP_{jf} stands for the finalized posterior probability of the PP number 'j'. The index 'f' stands for finalized PP.

After completion of the third step, the finalized updated network's beliefs (PD) are achieved. Using the final PD, the strengths of the potential root causes are listed and are given to the expert for assessment.

References

1. Chuah, Edward, Shyh-hao Kuo, Paul Hiew, William-Chandra Tjhi, Gary Lee, John Hammond, Marek T. Michalewicz, Terence Hung, and James C. Browne. "Diagnosing the root-causes of failures from cluster log files." InHigh Performance Computing (HiPC), 2010 International Conference on, pp. 1-10. IEEE, 2010.
2. Aamodt, Agnar. "Knowledge-intensive case-based reasoning in creek." In Advances in Case-Based Reasoning, pp. 1-15. Springer Berlin Heidelberg, 2004.
3. Okes, Duke. Root cause analysis: The core of problem solving and corrective action. ASQ Quality Press, 2009.
4. Kofod-Petersen, Anders, Helge Langseth, and Agnar Aamodt. "Explanations in Bayesian Networks using Provenance through Case-based Reasoning." InWorkshop Proceedings, p. 79. 2010.
5. Odd Erik Gundersen, Frode Srmo, Agnar Aamodt, Pl Skalle: A real-time decision support system for high cost oil-well drilling operations. AI Magazine, Volume 34, Number 1, Spring 2013. ISSN-0738-4602. pgs. 21-32.
6. Agnar Aamodt, Helge Langseth: Integrating Bayesian Networks into knowledge-intensive CBR. In Case-based reasoning integrations; Papers from theAAAI workshop. David Aha, Jody J. Daniels (eds.).Technical Report WS-98-15. AAAI Press, Menlo Park, 1998. ISBN 1-57735-068-5. pp1-6.
7. Forbus, K., Gentner, D. and Law, K.: MAC/FAC: A model of Similarity-based Retrieval. Cognitive Science, 19(2), April-June, 1995, pgs. 141-205.