Adaptation Knowledge Discovery from a Case Base

Mathieu d'Aquin, Fadi Badra, Sandrine Lafrogne, Jean Lieber, Amedeo Napoli, Laszlo Szathmary

Adaptation Knowledge Discovery from a Case Base

M. d'Aquin[1] and **F. Badra**[1] and **S. Lafrogne**[1] and **J. Lieber**[1] and **A. Napoli**[1] and **L. Szathmary**[1]

Abstract. In case-based reasoning, the adaptation step depends in general on domain-dependent knowledge, which motivates studies on adaptation knowledge acquisition (AKA). CABAMAKA is an AKA system based on principles of knowledge discovery from databases. This system explores the variations within the case base to elicit adaptation knowledge. It has been successfully tested in an application of case-based decision support to breast cancer treatment.

1 INTRODUCTION

Case-based reasoning (CBR [4]) aims at solving a target problem thanks to a case base. A case represents a previously solved problem. A CBR system selects a case from the case base and then adapts the associated solution, requiring domain-dependent knowledge for adaptation. The goal of adaptation knowledge acquisition (AKA) is to extract this knowledge. The system CABAMAKA applies principles of knowledge discovery from databases (KDD) to AKA. The originality of CABAMAKA lies essentially in the approach to AKA that uses a powerful learning technique that is guided by a domain expert, according to the spirit of KDD. This paper proposes an original and working approach to AKA, based on KDD techniques.

CBR and adaptation. A case in a given CBR application is usually represented by a pair $(\mathtt{pb}, \mathtt{Sol}(\mathtt{pb}))$ where pb represents a problem statement and $\mathtt{Sol}(\mathtt{pb})$, a solution of pb. CBR relies on the *source cases* $(\mathtt{srce}, \mathtt{Sol}(\mathtt{srce}))$ that constitute the *case base* CB. In a particular CBR session, the problem to be solved is called *target problem*, denoted by tgt. A case-based inference associates to tgt a solution $\mathtt{Sol}(\mathtt{tgt})$, with respect to the case base CB and to additional knowledge bases, in particular $\mathcal{O}$, the *domain ontology* that usually introduces the concepts and terms used to represent the cases.

A classical decomposition of CBR consists in the steps of retrieval and adaptation. *Retrieval* selects $(\mathtt{srce}, \mathtt{Sol}(\mathtt{srce})) \in \mathtt{CB}$ such that srce is judged to be similar to tgt. The goal of adaptation is to solve tgt by modifying $\mathtt{Sol}(\mathtt{srce})$ accordingly.

The work presented hereafter is based on the following model of adaptation, similar to *transformational analogy* [1]:

① $(\mathtt{srce}, \mathtt{tgt}) \mapsto \Delta\mathtt{pb}$, where $\Delta\mathtt{pb}$ encodes the similarities and dissimilarities of the problems srce and tgt.
② $(\Delta\mathtt{pb}, \mathtt{AK}) \mapsto \Delta\mathtt{sol}$, where AK is the adaptation knowledge and where $\Delta\mathtt{sol}$ encodes the similarities and dissimilarities of $\mathtt{Sol}(\mathtt{srce})$ and the forthcoming $\mathtt{Sol}(\mathtt{tgt})$.
③ $(\mathtt{Sol}(\mathtt{srce}), \Delta\mathtt{sol}) \mapsto \mathtt{Sol}(\mathtt{tgt})$, $\mathtt{Sol}(\mathtt{srce})$ is modified into $\mathtt{Sol}(\mathtt{tgt})$ according to $\Delta\mathtt{sol}$.

Adaptation is generally supposed to be domain-dependent in the sense that it relies on domain-specific adaptation knowledge. Therefore, this knowledge has to be acquired. This is the purpose of *adaptation knowledge acquisition* (AKA).

A related work in AKA. The idea of the research presented in [3] is to exploit the variations between source cases to learn adaptation rules. These rules compute variations on solutions from variations on problems. More precisely, ordered pairs $(\mathtt{srce\text{-}case}_1, \mathtt{srce\text{-}case}_2)$ of similar source cases are formed. Then, for each of these pairs, the variations between the problems $\mathtt{srce}_1$ and $\mathtt{srce}_2$ and the solutions $\mathtt{Sol}(\mathtt{srce}_1)$ and $\mathtt{Sol}(\mathtt{srce}_2)$ are represented (Δpb and Δsol). Finally, the adaptation rules are learned, using as training set the set of the input-output pairs $(\Delta\mathtt{pb}, \Delta\mathtt{sol})$. The experiments have shown that the CBR system using the adaptation knowledge acquired from the automatic system of AKA shows a better performance compared to the CBR system working without adaptation. This research has strongly influenced our work that is globally based on similar ideas.

2 CABAMAKA

Principles. CABAMAKA deals with case base mining for AKA. Although the main ideas underlying CABAMAKA are shared with those presented in [3], the followings are original ones. The adaptation knowledge that is mined has to be validated by experts and has to be associated with explanations that make it understandable by the user. In this way, CABAMAKA may be considered as a semi-automated (or interactive) learning system. Another difference with [3] lies in the volume of the cases that are examined: given a case base CB where $|\mathtt{CB}| = n$, the CABAMAKA system takes into account every ordered pair $(\mathtt{srce\text{-}case}_1, \mathtt{srce\text{-}case}_2)$ with $\mathtt{srce\text{-}case}_1 \neq \mathtt{srce\text{-}case}_2$ (whereas in [3], only the pairs of *similar* source cases are considered, according to a fixed criterion). Thus, the CABAMAKA system has to cope with $n(n-1)$ pairs, a rather large number of elements, since in our application $n \simeq 750$. ($n(n-1) \simeq 5 \cdot 10^5$). This is why efficient techniques of knowledge discovery from databases (KDD [2]) have been chosen for this system.

Principles of KDD. The goal of KDD is to discover knowledge from databases, with the supervision of an analyst (expert of the domain). A KDD session usually relies on three main steps: data preparation, data-mining and interpretation.

Data preparation is based on formatting and filtering operations. The formatting operations transform the data into a form allowing the application of the chosen data-mining operations. The filtering operations are used for removing noisy data and for focusing the data-mining operation on special subsets of objects and/or attributes.

Data-mining methods are applied to extract pieces of information from the data. These pieces of information have some regular

[1] LORIA (UMR 7503 CNRS–INPL–INRIA-Nancy 2–UHP), BP 239, 54506 Vandœuvre-lès-Nancy, FRANCE

properties allowing their extraction. For example, CHARM [5] is a data-mining algorithm that performs efficiently the extraction of *frequent closed itemsets* (*FCIs*). CHARM inputs a database in the form of a set of transactions, each *transaction* T being a set of boolean properties or *items*. An *itemset* I is a set of items. The support of I, $\texttt{support}(I)$, is the proportion of transactions T of the database possessing I ($I \subseteq T$). I is frequent, with respect to a threshold $\sigma \in [0;1]$, whenever $\texttt{support}(I) \geq \sigma$. I is closed if it has no proper superset J ($I \subsetneq J$) with the same support.

Interpretation aims at interpretating the output of data-mining i.e. the FCIs in the present case, with the help of an analyst. In this way, the interpretation step produces new knowledge units (e.g. rules).

Formatting. The formatting step of CABAMAKA inputs the case base $\texttt{CB}$ and outputs a set of transactions obtained from the pairs $(\texttt{srce-case}_1, \texttt{srce-case}_2)$. It is composed of two substeps. During the first substep, each $\texttt{srce-case} = (\texttt{srce}, \texttt{Sol}(\texttt{srce})) \in \texttt{CB}$ is formatted in two sets of boolean properties: $\Phi(\texttt{srce})$ and $\Phi(\texttt{Sol}(\texttt{srce}))$. The computation of $\Phi(\texttt{srce})$ consists in translating $\texttt{srce}$ from the problem representation formalism to $2^{\mathcal{P}}$, $\mathcal{P}$ being a set of boolean properties. Possibly, some information may be lost during this translation, but this loss has to be minimized. Now, this translation formats an expression $\texttt{srce}$ expressed in the framework of the domain ontology $\mathcal{O}$ to an expression $\Phi(\texttt{srce})$ that will be manipulated as data, i.e. without the use of a reasoning process. Therefore, in order to minimize the translation loss, it is assumed that if $p \in \Phi(\texttt{srce})$ and p entails q (given $\mathcal{O}$) then $q \in \Phi(\texttt{srce})$. In other words, $\Phi(\texttt{srce})$ is assumed to be deductively closed given $\mathcal{O}$ in the set $\mathcal{P}$. The same assumption is made for $\Phi(\texttt{Sol}(\texttt{srce}))$. How this first substep of formatting is computed in practice depends heavily on the representation formalism of the cases.

The second substep of formatting produces a transaction $T = \Phi((\texttt{srce-case}_1, \texttt{srce-case}_2))$ for each ordered pair of distinct source cases, based on the sets of items $\Phi(\texttt{srce}_1)$, $\Phi(\texttt{srce}_2)$, $\Phi(\texttt{Sol}(\texttt{srce}_1))$ and $\Phi(\texttt{Sol}(\texttt{srce}_2))$. Following the model of adaptation presented in introduction (items ①, ② and ③), T has to encode the properties of $\Delta\texttt{pb}$ and $\Delta\texttt{sol}$. $\Delta\texttt{pb}$ encodes the similarities and dissimilarities of $\texttt{srce}_1$ and $\texttt{srce}_2$, i.e.:

- The properties common to $\texttt{srce}_1$ and $\texttt{srce}_2$ (marked by "$=$"),
- The properties of $\texttt{srce}_1$ that $\texttt{srce}_2$ does not share ("$-$") and
- The properties of $\texttt{srce}_2$ that $\texttt{srce}_1$ does not share ("$+$").

All these properties are related to problems and thus are marked by $\texttt{pb}$. $\Delta\texttt{sol}$ is computed in a similar way and $\Phi(T) = \Delta\texttt{pb} \cup \Delta\texttt{sol}$. For example,

$$\text{if} \quad \begin{cases} \Phi(\texttt{srce}_1) = \{a, b, c\} & \Phi(\texttt{Sol}(\texttt{srce}_1)) = \{A, B\} \\ \Phi(\texttt{srce}_2) = \{b, c, d\} & \Phi(\texttt{Sol}(\texttt{srce}_2)) = \{B, C\} \end{cases}$$
$$\text{then} \quad T = \left\{a^{-}_{\texttt{pb}}, b^{=}_{\texttt{pb}}, c^{=}_{\texttt{pb}}, d^{+}_{\texttt{pb}}, A^{-}_{\texttt{sol}}, B^{=}_{\texttt{sol}}, C^{+}_{\texttt{sol}}\right\} \quad (1)$$

Mining. The extraction of FCIs is computed thanks to CHARM (in fact, thanks to a tool based on a CHARM-like algorithm) from the set of transactions. A transaction $T = \Phi((\texttt{srce-case}_1, \texttt{srce-case}_2))$ encodes a specific adaptation $((\texttt{srce}_1, \texttt{Sol}(\texttt{srce}_1)), \texttt{srce}_2) \mapsto \texttt{Sol}(\texttt{srce}_2)$. An FCI extracted may be considered as a generalization of a set of transactions. For example, if $I_{ex} = \left\{a^{-}_{\texttt{pb}}, c^{=}_{\texttt{pb}}, d^{+}_{\texttt{pb}}, A^{-}_{\texttt{sol}}, B^{=}_{\texttt{sol}}, C^{+}_{\texttt{sol}}\right\}$ is an FCI, I_{ex} is a generalization of a subset of the transactions including the transaction T of equation (1): $I_{ex} \subseteq T$. The interpretation of this FCI as an adaptation rule is explained below.

Interpretation. The interpretation step is supervised by the analyst. The CABAMAKA system provides the analyst with the extracted FCIs and facilities for navigating among them. The analyst may select an FCI, say I, and interpret I as an adaptation rule. For example, the FCI I_{ex} may be interpreted in the following terms:

if a is a property of $\texttt{srce}$ but is not a property of $\texttt{tgt}$,
c is a property of both $\texttt{srce}$ and $\texttt{tgt}$,
d is not a property of $\texttt{srce}$ but is a property of $\texttt{tgt}$,
A and B are properties of $\texttt{Sol}(\texttt{srce})$ and
C is not a property of $\texttt{Sol}(\texttt{srce})$
then the properties of $\texttt{Sol}(\texttt{tgt})$ are
$\Phi(\texttt{Sol}(\texttt{tgt})) = (\Phi(\texttt{Sol}(\texttt{srce})) \setminus \{A\}) \cup \{C\}$.

This has to be translated as an adaptation rule r of the CBR system. Then the analyst corrects r and associates an explanation with it.

Implementation. The application domain of the CBR system we are developing is breast cancer treatment: in this application, a problem $\texttt{pb}$ describes a class of patients with a set of attributes and associated constraints (holding on the age of the patient, the size and the localization of the tumor, etc.). A solution $\texttt{Sol}(\texttt{pb})$ of $\texttt{pb}$ is a set of therapeutic decisions (in surgery, chemotherapy, etc.). The requested behavior of the CBR system is to provide a treatment and explanations on this treatment proposal. This is why the analyst is required to associate an explanation to a discovered adaptation rule.

The problems, solutions and the domain ontology of the application are represented in OWL DL (recommendation of the W3C).

3 CONCLUSION

The CABAMAKA system presented in this paper is inspired by the research presented in [3] and by the principles of KDD for the purpose of semi-automatic adaptation knowledge discovery. It has enabled to discover several useful adaptation rules for a medical CBR application. It has been designed to be reusable for other CBR applications: only a few modules of CABAMAKA are dependent on the formalism of the cases and of the domain ontology, and this formalism, OWL DL, is a well-known standard. One element of future work consists in searching for ways of simplifying the presentation of the numerous extracted FCIs to the analyst. This involves an organization of these FCIs for the purpose of navigation among them. Such an organization can be a hierarchy of FCIs according to their specificities or a clustering of the FCIs in themes.

REFERENCES

[1] J. G. Carbonell, 'Learning by analogy: Formulating and generalizing plans from past experience', in *Machine Learning, An Artificial Intelligence Approach*, ed., R. S. Michalski and J. G. Carbonell and T. M. Mitchell, chapter 5, 137–161, Morgan Kaufmann, Inc., (1983).

[2] M.H. Dunham, *Data Mining – Introductory and Advanced Topics*, Prentice Hall, Upper Saddle River, NJ, 2003.

[3] K. Hanney and M. T. Keane, 'Learning Adaptation Rules From a Case-Base', in *Advances in Case-Based Reasoning – Third European Workshop, EWCBR'96*, eds., I. Smith and B. Faltings, LNAI 1168, pp. 179–192. Springer Verlag, Berlin, (1996).

[4] C. K. Riesbeck and R. C. Schank, *Inside Case-Based Reasoning*, Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, 1989.

[5] M. J. Zaki and C.-J. Hsiao, 'CHARM: An Efficient Algorithm for Closed Itemset Mining', in *SIAM International Conference on Data Mining SDM'02*, pp. 33–43, (Apr 2002).