A Framework for Integrating Data Mining and Case-Based Reasoning

Torgeir Dingsøyr*

January 15, 1998

Data Mining and Case-Based Reasoning are methods used for *decision support*; to organize and process information to make it available for improving the quality of decisions. It is likely that integration of the two methods will lead to a better usage of information. Here, we give a quick introduction to the methods, briefly describe existing integrated methods, and outline a framework for different integration approaches.

Data Mining (DM) has become a popular method for extracting information from large databases, in the form of patterns. The patterns can be *informative* or *predictive*, and some types are classification, regression, clustering, dependency modeling and change and deviation analysis. The whole process of discovering knowledge in databases is referred to as *Knowledge Discovery in Databases* [6].

Case-Based Reasoning (CBR) is a method for solving problems by comparing a problem situation – a case – to previously experienced ones. The aim is to store information about earlier situations, and when new ones arrive, *retrieve* the situation that is most similar, and *reuse* it – or *revise* it to match the new problem if the most similar problem does not match sufficiently. If the new case provides new insight it should be *retained*. An introduction to CBR is given in [1].

At present, there are some integrated systems under development, like Case-Method, developed by the NEC corporation for handling corporate memory [7], and a system for forecasting of epidemics, developed at the University of Rostock [5]. Most of the research have been to integrate Bayesian networks with CBR. Microsoft has developed two prototype systems with codename "Aladdin" to diagnose problems in customer support [4]. A system named INBANCA (Integrating Bayes networks with CBR for Planning) for planning in a simulated soccer environment has been outlined in [2], but it has been discontinued due to lack of funding. At the University of Salford, a system using Bayesian networks for indexing cases have been developed [8], and at the University of Helsinki,

*Dept. of Computer and Information Science, Norwegian University of Science and Technology, 7034 Trondheim, Norway. E–mail: dingsoyr@idi.ntnu.no. Telephone: +47 91 87 19 62

CBR and Bayesian networks are used for classification [9]. Tools for combining CBR and decisions trees were developed in the Esprit project INRECA [3].

Integration of DM and CBR can be done with either of the methods as the master and the other as the slave, depending on which method that uses information provided by the other. First we outline methods for integration with CBR as the master, then with DM as the master. We assume that there exists a casebase (database of cases) and in some scenarios also an external database from which further information can be mined.

With CBR as the master, we can: *Find features for a case* (from casebase) – Classify the cases in the casebase for use. This might speed up the computation of similarity metrics in the *retrieve* step of the CBR cycle, if the case(s) that will be retrieved are known to be in the same class as the new case.

Find features for a case (from a database) – A database can be searched to supplement the information given in a case. For instance, Gibbs sampling can be used to fill inn missing features in case-data.

Find domain knowledge (from a database or a casebase) – Domain knowledge might be mined from the data in the form of functions, rules or causal graphs which can later be used by the Case-Based Reasoner when identifying features and explaining away unimportant features in the *retrieve* step, adapting in the *reuse* step, or explaining cases in the *retain* step.

Construct "artificial cases" – we can imagine that it is possible to construct cases from a database, that is not present in a casebase. This would require a deep integration where the DM algorithm would search for patterns in the form of cases, which could be evaluated by a novelty function which gives high values to cases not present in the casebase.

With DM as the master: *DM searches is the case* – DM is only one part of the KDD process which can involve accessing several files, cleaning data and interpreting results. The DM search may be time-consuming. The information about the search results and the whole knowledge discovery process might be stored in a case so that extra time will not be spent on mining the same information more than once.

CBR provides info – CBR can be used in providing some background knowledge about features in a database, for instance, the weight of features for a classifier can be learned from the CBR tool. In a Bayesian network, the structure of the network might be set up by the CBR tool (model construction), using its "expert knowledge" and the parameters learned using DM algorithms. CBR can also be used to provide utility, validity and novelty functions for the DM algorithm from the domain that the CBR tool is working in (model evaluation).

We have now given the status of research on integration of DM and CBR, outlined a framework for different methods of integration, and stated areas for further research. In the future, we aim to implement a prototype system which combines CBR and domain knowledge mined from a database as a Bayesian network for doing causal inference, and that also clusters the cases to compute similarity metrics fast.

References

[1] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

[2] David W. Aha and Li Wu Chang. Cooperative bayesian and case-based reasoning for solving multiagent planning tasks. Technical report, Navy Center for Applied Research in AI, Naval Research Laboratory, Washington, DC, USA, 1996.

[3] Klaus-Dieter Althof, Eric Auriol, Ralph Traphöner, and Stefan Wess. Inreca – a seamlessly integrated system based on inductive inference and case-based reasoning. In Agnar Aamodt and Manuela Veloso, editors, *Case-Based Reasoning Research and Development, ICCBR-95*, pages 371–380, 1995.

[4] John S. Breese and David Heckerman. Decision-theoretic case-based reasoning. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 56–63, 1995.

[5] M. Bull, G. Kundt, and L. Gierl. Discovering of health risks and case-based forecasting of epidemics in a health surveillance system. In Jan Komorowski and Jan Zytkow, editors, *Principles of Data Mining and Knowledge Discovery. Proceedings*, pages 68–77, 1997.

[6] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. *Advances in Knowledge Discovery and Data Mining*, chapter From Data Mining to Knowledge Discovery: An overview. AAAI/MIT Press, 1996.

[7] Hiroaki Kitano, Hideo Shimazu, and Akihiro Shibata. Case-method: A methodology for building large-scale case-based systems. In *Proceedings of the AAAI*, pages 303–308, 1993.

[8] Andres F. Rodriguez, Sunil Vadera, and L. Enrique Sucar. A probabilistic model for case-based reasoning. In I Smith and B Faltings, editors, *Case-Based Reasoning Research and Development, ICCBR-97. Proceedings*, pages 623–632, 1997.

[9] Henry Tirri, Perti Kontkanen, and Petri Myllymäksi. A bayesian framework for case-based reasoning. In I Smith and B Faltings, editors, *Advances in Case-Based Reasoning, EWCBR-96*, pages 413–427, 1996.