Recent Themes in Case-Based Reasoning and Knowledge Discovery

Isabelle Bichindaritz
Computer Science Department
State University of New York
Owego, New York, USA

Cindy Marling
School of Electrical Engineering
and Computer Science
Ohio University
Athens, Ohio, USA

Stefania Montani
DISIT
Computer Science Institute
Università del Piemonte Orientale
Alessandria, Italy

Abstract

Case-based reasoning (CBR) systems have tight connections with machine learning and knowledge discovery and often incorporate diverse knowledge discovery functionalities and algorithms. This article presents themes identified in work presented at recent workshops on synergies between CBR and knowledge discovery. Among the main themes appear Big Data, with cases involving signals, images, texts, and other complex types of data; similarity metric discovery, in the form of weight spaces, feature weights, and feature selection; adaptation knowledge; explainability and transparency; and user centeredness and interactivity. Researchers highlight the advantages of case-based reasoning in terms of its lazy learning, explainability, user centeredness, and interactivity when performing knowledge discovery, as well as how diverse knowledge discovery methods can improve CBR.

Introduction

Case-based reasoning (CBR) systems have tight connections with machine learning and knowledge discovery (KD) (Bichindaritz 2015). In 2014 and 2016, workshops on Synergies between Case-based Reasoning and Knowledge Discovery were held at the International Conference on Case-based Reasoning. This article summarizes the major themes illustrated by the papers presented at these workshops. As instance-based learners, CBR systems possess the advantage of being easy - aka *lazy* - learners, because they defer the decision of how to generalize beyond the training set until a target new case is encountered. This is in opposition to the way in which most other knowledge discovery systems operate by abstracting models from cases. CBR is also known for its knowledge containers - vocabulary, similarity measure, case base, and adaptation. The case base in and of itself is often a major focus of knowledge discovery in CBR, with its cases, structures, and organization. After a review of major knowledge discovery tasks and how they relate to CBR, recent themes are highlighted.

Background: Knowledge Discovery

Knowledge discovery may be defined as the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Hand, Mannila, and Smyth 2001). Broadly viewed, it encompasses the automated learning of new trends and associations from data as well as novel characterizations and explanations of data. Some functionalities are well defined and researched, including:

- Classification / Prediction. Classification is a supervised knowledge discovery method applied to datasets containing instances which have been labeled with the categories, or classes, to which they belong. Examples of classifiers include neural networks, support vector machines (SVMs), naive Bayes, and decision trees. Classifiers may be used for categorization (e.g., diagnosis) or for prediction (e.g., prognosis). Datasets containing instances labeled by numeric values for a variable of interest, rather than by a category, may be used for prediction.
- Association Mining. Association mining identifies frequent itemsets in a dataset, from which it derives rules associating the items, as in market basket analysis. It is an unsupervised method. The best known algorithm in this category is the Apriori algorithm.
- Clustering. Clustering finds groups of similar objects in a dataset, which are also dissimilar from the objects in other groups. In addition to similarity-based methods like k-means, some clustering methods use density-based algorithms or hierarchical algorithms.

These core functionalities can be combined, applied to different types of data, including multimedia data, and augmented by other functionalities, e.g., feature selection.

The relationship between CBR and knowledge discovery is bidirectional. On the one hand, CBR systems may be thought of as knowledge discovery systems because they can perform classification or prediction tasks (Bichindaritz 2015), which is a consequence of their being instance-based learners (IBL). The classification or prediction achieved in CBR gives the case base a competency beyond that provided by the data alone. An important distinction is that CBR systems start their reasoning from comprehensive knowledge units (i.e., cases), while most knowledge discovery systems

start from raw data. This is why case mining, which consists of mining raw data to build cases, is a knowledge discovery task often used in CBR. We have already noted that CBR systems are instance-based learners; Mitchell views the complex, symbolic representations of instances (cases) as distinguishing CBR from other IBL approaches (Mitchell 1997). Furthermore, CBR systems follow a Retrieve, Reuse, Revise, Retain reasoning cycle (Aamodt and Plaza 1994). In the Retain step, new cases learned by the system are added to the case base for future use. This type of machine learning may be viewed as incremental learning or online learning.

The other side of the relationship between CBR and knowledge discovery is that CBR systems make efficient use of knowledge discovery techniques for descriptive modeling. In descriptive modeling, as opposed to predictive modeling, the goal is to characterize, explain, and/or better understand the underlying nature of observed data or cases. Among the main techniques encountered are cluster analysis, rule induction, hierarchical cluster analysis, and decision tree induction.

It should be noted that, while *knowledge discovery, data mining* and *machine learning* are closely related, there is no clear consensus on the distinctions between them or on when each particular term should be used. In an early seminal paper on knowledge discovery in databases (Fayyad, Piatetsky-Shapiro, and Smyth 1996), it was suggested that knowledge discovery referred to the overall process of discovering useful knowledge from data, while data mining was just one step of the overarching process, in which pattern extraction algorithms were applied. These pattern extraction algorithms, for classification, regression and clustering, were then borrowed from machine learning or from other disciplines partaking in data mining such as pattern recognition, statistics, or mathematics. However, the frontier between machine learning, statistics and pattern recognition has receded over the years. In the current vernacular, the terms knowledge discovery, data mining, and machine learning may be used almost interchangeably. While our 2014 workshop was entitled "Workshop on Synergies Between CBR and Data Mining," we switched to the term *knowledge discovery* for our 2016 workshop, as being more inclusive. We take the broad perspective that, whenever a system learns knowledge that is new to it, through the analysis of raw data or cases, for external application or its own internal use, it is engaged in knowledge discovery.

The primary motivations for performing knowledge discovery during CBR are to:

- Increase efficiency, mainly of the Retrieve step, but also of the Reuse, Revise, and Retain steps
- Increase robustness, tolerance to noise
- Increase reasoning accuracy and effectiveness
- Decrease storage needs
- Follow a cognitive model
- Add functionality
- Perform metareasoning, such as knowledge discovery to learn new adaptation rules

We also foresee synergies with Big Data for processing large datasets in distributed memory that can make efficient use of knowledge discovery while processing on a larger scale.

Recent Themes

Table 1 shows projects exemplifying recent themes in synergies between CBR and knowledge discovery. Identified themes include:

- Big Data. Most systems apply CBR not only to large volumes of data but also to data presenting high levels of complexity, challenge, and opportunity. Dileep and Chakraborti propose a multi-step similarity assessment functioning at different levels of granularity for textual CBR (Dileep and Chakraborti 2014). Hromic and Hayes apply a signal processing method based on attack-decay-sustain-release (ADSR) envelopes, which are commonly learned in acoustics signal modeling, to Twitter hashtags in order to construct Twitter datasets suitable for event detection research, particularly research involving CBR (Hromic and Hayes 2014). Zhang, Zhang, and Leake introduce sieve streaming, a recent method for massive data summarization, and adopt it for continuous incremental case base maintenance of cases from a case stream, without access to the full case base (Zhang, Zhang, and Leake 2016). Barua et al. combine independent component analysis (ICA), hierarchical clustering, and CBR to detect ocular artifacts in electroencephalogram (EEG) signals. In empirical experiments, ICA was used in case formulation, hierarchical clustering was used to build the initial case library, and CBR was used to classify signals as EEG or ocular artifact. An overall classification accuracy of 95% was achieved (Barua et al. 2014). Olsson et al. combine signal processing, statistics, machine learning and CBR for diagnosing problems in heavy duty machines. The first three approaches are used for continuous monitoring of machines to detect anomalies, while CBR is used offline, to classify the problems identified and to retrieve cases for manual decision making (Olsson et al. 2014).
- Similarity Metrics. Effective feature selection and weight learning are essential to similarity assessment. Sekar and Chakraborti present a conversational recommender system in which product cases are indexed by their region in preference weight space in order to predict the prospective buyers of any particular product (Sekar and Chakraborti 2016). Guo, Jerbi, and O'Mahony showcase a job recommender system personalized to users based on a combination of machine learning approaches for feature selection and feature weighting. Jobs in this system are represented in a structured representation with well-defined features, amenable to varied similarity assessment functions using weighted aggregates of features (Guo, Jerbi, and O'Mahony 2014).
- Adaptation Knowledge. Improvements in adaptation can have a large impact, especially on classification and prediction performance. Tomasic and Funk explore synergies between CBR and regression analysis in the domain of quality control for automotive assembly. Regression models are used to improve CBR performance by:

Table 1: Projects Illustrating Recent Themes in Case-Based Reasoning and Knowledge Discovery

Citation	Themes	CBR Phase	Domain
(Adedoyin et al. 2016)	Big Data Explanation Application	Retrieval	Fraud Detection
(Barua et al. 2014)	Signal Processing Application	Case Acquisition Retrieval	Classifying Ocular Artifacts in EEGs
(Canensi et al. 2014) (Canensi et al. 2016)	Big Data Explanation Interactivity	Retrieval	Medical Processes
(Dileep and Chakraborti 2014)	Big Data Knowledge Rich CBR	Retrieval Similarity	Textual CBR
(Eyorokon et al. 2016)	Explanation Interactivity	Retrieval	Conversational CBR Dialogue
(Guo, Jerbi, and O'Mahony 2014)	Big Data Application	Retrieval Similarity Weight Learning	Job Recommendation
(Hromic and Hayes 2014)	Big Data Signal Processing Application	Case Mining	Twitter Datasets for Event Detection
(Olsson et al. 2014)	Signal Processing Application	Retrieval Similarity	Fault Diagnosis in Heavy Duty Machines
(Sekar and Chakraborti 2016)	Big Data Interactivity Application	Retrieval	Product Recommendation
(Tomasic and Funk 2014)	Regression Analysis Application	Reuse Revise	Quality Control in Manufacturing
(Zhang, Zhang, and Leake 2016)	Big Data	Retain	Streaming Data

verifying that CBR-recommended adjustments will lead to improved manufacturing outcomes; and adapting retrieved adjustments to ensure that only relevant variables are modified (Tomasic and Funk 2014).

- Explainability and Transparency. Canensi et al. mine processes from traces of executed actions in a medical domain while keeping references to the traces through a novel algorithm. The algorithm shows very high precision, i.e., it never includes incorrect paths. Moreover, since the traces from which a model branch was mined are explicitly referenced, the model can be used as an indexing structure, for flexible and efficient trace retrieval (Canensi et al. 2014; 2016). Adedoyin et al. note CBR's advantage in explainability and transparency, when compared to logistic regression and neural networks, for the task of identifying fraudulent patterns among financial transactions (Adedoyin et al. 2016).
- User Centeredness and Interactivity. Eyorokon et al. present a system that supports knowledge discovery using a conversational CBR process. Investigation involves the creation of a knowledge goal trajectory, where an initial knowledge goal is refined, focused, or changed, through a variety of phases in the investigative sequence, in a creative manner in reaction to retrieved information (Eyorokon et al. 2016).

It is worth noting that many papers address more than one theme, as illustrated in Table 1.

Conclusion

In this paper, we have reported on the work presented at two recent workshops on Synergies between Case-based Reasoning and Knowledge Discovery, which were held at the International Conference on Case Based Reasoning. The analysis we have carried out has allowed us to identify different roles that knowledge discovery can play to support CBR. As can be expected, knowledge discovery techniques have been adopted for case mining. Moreover, they have been frequently integrated with the Retrieve step, improving similar-

ity calculation, weight learning and indexing. Interestingly enough, examples also exist of the adoption of these techniques in all other steps of the CBR cycle, i.e., Reuse, Revise and Retain.

Moreover, we were able to discover a set of common trends. We found a significant number of applications in Big Data domains (or, at least, in data intensive domains, such as business process management, e-commerce and signal processing). The adoption of knowledge discovery techniques can quite naturally tackle some of the issues presented by this kind of data, and thus facilitate and enhance the implementation of CBR systems in such contexts. We observed an increasing attention to transparency and context-awareness. Knowledge discovery techniques can be used to preserve contextual information in cases, thus facilitating the generation of explanations and justifications of system output. Finally, we noted an increasing interest in the implementation of interactive tools. Here, the integration of CBR and knowledge discovery techniques provides an initial solution to the user, who is then allowed to ask for progressive refinements, being directly involved in the reasoning process.

In summary, knowledge discovery techniques can be deployed in multiple ways, in a wide variety of domains, helping CBR researchers, system implementers and users to manage complex issues. Indeed, as exemplified by the works we have examined, CBR can significantly benefit from integration with knowledge discovery, gaining in reasoning accuracy, effectiveness and transparency. In the future, we plan to extend our analysis, examining how, conversely, the CBR methodology can improve and enhance the performance of systems based on knowledge discovery techniques. This will conclude our investigation of the multiple synergies that can be achieved between these two fields of artificial intelligence research.

References

Aamodt, A., and Plaza, E. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications* 7(I):39–59.

Adedoyin, A.; Kapetanakis, S.; Petridis, M.; and Panaousis, E. 2016. Evaluating case-based reasoning knowledge discovery in fraud detection. In Coman, A., and Kapetanakis, S., eds., *Workshop Proceedings of the Twenty-Fourth International Conference on Case-Based Reasoning*, 182–191.

Barua, S.; Begum, S.; Ahmed, M. U.; and Funk, P. 2014. Classification of ocular artifacts in EEG signals using hierarchical clustering and case-based reasoning. In Leake, D. B., and Lieber, J., eds., *Workshop Proceedings of the Twenty-Second International Conference on Case-Based Reasoning*, 213–223.

Bichindaritz, I. 2015. Data mining methods for case-based reasoning in health sciences. In Kendall-Morwick, J., ed., *Workshop Proceedings of the Twenty-Third International Conference on Case-Based Reasoning*, 184–198.

Canensi, L.; Montani, S.; Leonardi, G.; and Terenziani, P. 2014. ChAPMaN: a context aware process miner. In Leake, D. B., and Lieber, J., eds., *Workshop Proceedings of the Twenty-Second International Conference on Case-Based Reasoning*, 202–212.

Canensi, L.; Leonardi, G.; Montani, S.; and Terenziani, P. 2016. A context-aware miner for medical processes. In Coman, A., and Kapetanakis, S., eds., *Workshop Proceedings of the Twenty-Fourth International Conference on Case-Based Reasoning*, 192–201.

Dileep, K. V. S., and Chakraborti, S. 2014. Intelligent integration of knowledge sources for TCBR. In Leake, D. B., and Lieber, J., eds., *Workshop Proceedings of the Twenty-Second International Conference on Case-Based Reasoning*, 224–234.

Eyorokon, V. B.; Bengfort, B.; Panjala, U. S.; and Cox, M. T. 2016. Goal trajectories for knowledge investigations. In Coman, A., and Kapetanakis, S., eds., *Workshop Proceedings of the Twenty-Fourth International Conference on Case-Based Reasoning*, 202–211.

Fayyad, U.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From data mining to knowledge discovery in databases. *AI Magazine* 17:37–54.

Guo, X.; Jerbi, H.; and O'Mahony, M. 2014. An analysis framework for content-based job recommendation. In Leake, D. B., and Lieber, J., eds., *Workshop Proceedings of the Twenty-Second International Conference on Case-Based Reasoning*, 235–244.

Hand, D. J.; Mannila, H.; and Smyth, P. 2001. *Principles of Data Mining*. Cambridge, Massachusetts: MIT Press.

Hromic, H., and Hayes, C. 2014. Constructing Twitter datasets using signals for event detection evaluation. In Leake, D. B., and Lieber, J., eds., *Workshop Proceedings of the Twenty-Second International Conference on Case-Based Reasoning*, 245–254.

Mitchell, T. 1997. *Machine Learning*. New York, New York: McGraw-Hill.

Olsson, T.; Källström, E.; Gillblad, D.; Funk, P.; Lindström, J.; Hakanssson, L.; Lundin, J.; Svensson, M.; and Larsson, J. 2014. Fault diagnosis of heavy duty machines: Automatic transmission clutches. In Leake, D. B., and Lieber, J., eds., *Workshop Proceedings of the Twenty-Second International Conference on Case-Based Reasoning*, 182–191.

Sekar, A., and Chakraborti, S. 2016. Learning a region of user's preference for product recommendation. In Coman, A., and Kapetanakis, S., eds., *Workshop Proceedings of the Twenty-Fourth International Conference on Case-Based Reasoning*, 212–221.

Tomasic, I., and Funk, P. 2014. Potential synergies between case-based reasoning and regression analysis in assembly processes. In Leake, D. B., and Lieber, J., eds., *Workshop Proceedings of the Twenty-Second International Conference on Case-Based Reasoning*, 192–201.

Zhang, Y.; Zhang, S.; and Leake, D. 2016. Case-base maintenance: A streaming approach. In Coman, A., and Kapetanakis, S., eds., *Workshop Proceedings of the Twenty-Fourth International Conference on Case-Based Reasoning*, 222–231.