The Knowledge Engineering Review, Vol. 26:1, 25–29. © Cambridge University Press, 2011
doi:10.1017/S0269888910000378

Data mining: past, present and future

FRANS COENEN

Department of Computer Science, The University of Liverpool, Liverpool L693BX, UK;
e-mail: coenen@liverpool.ac.uk

Abstract

Data mining has become a well-established discipline within the domain of artificial intelligence (AI) and knowledge engineering (KE). It has its roots in machine learning and statistics, but encompasses other areas of computer science. It has received much interest over the last decade as advances in computer hardware have provided the processing power to enable large-scale data mining to be conducted. Unlike other innovations in AI and KE, data mining can be argued to be an application rather then a technology and thus can be expected to remain topical for the foreseeable future. This paper presents a brief review of the history of data mining, up to the present day, and some insights into future directions.

1 Introduction

Data mining has become an established discipline within the scope of computer science. The origins of data mining can be traced back to the late 80s when the term began to be used, at least within the research community. In the early days, there was little agreement on what the term data mining encompassed, and it can be argued that in some sense this is still the case. Broadly, data mining can be defined as a set of mechanisms and techniques, realized in software, to extract hidden information from data. The word *hidden* in this definition is important; SQL style querying, however sophisticated, is not data mining. In addition, the term *information* should be interpreted in its widest sense. By the early 1990s, data mining was commonly recognized as a sub-process within a larger process called knowledge discovery in databases or KDD (although in the modern context of data mining Knowledge Discovery in Data would be more apt, as we are no longer preoccupied solely by databases). The most commonly used definition of KDD is that attributed to Fayyad *et al.*: 'The non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data' (Fayyad *et al.*, 1996). As such data mining should be viewed as the sub-process, within the overall KDD process, concerned with the discovery of 'hidden information'. Other sub-processes that form part of the KDD process are data preparation (warehousing, data cleaning, pre-processing, etc.) and the analysis/visualization of results. For may practical purposes, KDD and data mining are seen as synonymous, but technically one is a sub-process of the other.

The data that data mining techniques were originally directed at was tabular data and, given the processing power available at the time, computational efficiency (and particular the number databases accesses) was of significant concern. As the amount of processing power generally available increased, processing time (although still an issue) became less of a concern and was replaced with a desire for accuracy and a desire to mine ever larger data collections. Today, in the context of tabular data, we have a well-established range of data mining techniques available. It is well within the capabilities of many commercial enterprises and researchers to mine tabular data, using software such as SPSS clementine or Weka, on standard desktop machines. However, the amount of electronic data collected by all kinds of institutions and commercial enterprises, year on year, continues to grow and thus there is still a need for effective mechanisms to mine ever larger data sets. A second current focus of the data mining community is the application of data mining to non-standard data sets (i.e. non-tabular data sets). Examples include: image sets, document collections, video, multimedia data of all kinds and graph and network data.

The popularity of data mining increased significantly in the 1990s, notably with the establishment of a number of dedicated conferences; the ACM SIGKDD annual conference in 1995, and the European PKDD and the Pacific/Asia PAKDD conferences in 1997 (The IEEE ICDM conference was not introduced till 2001 as was the first SIAM conference on data mining). This increase in popularity can be attributed to advances in technology; the computer processing power and data storage capabilities available meant that the processing of large volumes of data using desk top machines was a realistic possibility. It became common place for commercial enterprises to maintain data in computer readable form, in most cases this was primarily to support commercial activities, the idea that this data could be mined often came second. The 1990s also saw the introduction of customer loyalty cards (particularly with respect to large super market chains) that allowed enterprises to record customer purchases, the resulting data could then be mined to identify customer purchasing patterns. The popularity of data mining has continued to grow over the last decade with a particular current emphasis on mining non-standard data (i.e. non-tabular data).

2 Data mining mechanism and techniques

The mechanisms and techniques within the remit of data mining can be described as an amalgamation of approaches to machine learning and statistics; from this perspective, data mining can be said to have ‘grown’ out of the disciplines of machine learning and statistics. Indeed the data mining community is dominated by a mix of computer scientists and statisticians. The European conference on machine learning and the European conference on principles and practice of knowledge discovery in databases (PKDD) came together on 2001 and have stayed together ever since. There is, however, a distinction between data mining and machine learning. Data mining is focused on data (in all its formats) and as such can be viewed as an application domain; while machine learning, at least in its traditional form, is focussed on mechanisms whereby computers can learn (e.g. one focus of early work on machine learning was computer programmes that could learn to play chess). Machine learning can thus be viewed as a technology, whereas data mining, and by extension KDD, as an application.

Traditionally, data mining techniques can very broadly be categorized as being directed as either: (i) pattern extraction/identification, (ii) data clustering or (iii) classification/categorization. Each is briefly considered in more detail in the following subsections. Within the current data mining literature, we can also find reference to many other techniques that have been adopted from fields such as statistics and mathematics, for example, linear regression and principal component analysis.

2.1 Pattern extraction

Throughout its history, data mining has had a substantial focus on finding patterns in data. These patterns can take many forms, we have already mentioned customer purchasing patterns; alternative patterns may be trends in temporal or longitudinal data, frequently occurring subgraphs in graph data and so on. A patten is any frequently occurring combination of entities, events, objects, etc. The exemplar pattern mining technique is association rule mining (ARM) as first proposed by Agrawal *et al.* in the context of super market basket analysis (Agrawal *et al.*, 1993). The aim here was to identify frequent occurring patterns in the data and then, from these patterns, extract association rules (ARs). An AR is a probabilistic rule that states that if some set of data attributes occur together then some other (disjoint) set of attributes is also likely to occur. The fundamental challenge of ARM is that given a data set with N attributes (field-value pairs), there are 2^N-1 candidate patterns. ARM has attracted much attention from the data mining community over the years. Many extensions have been proposed such as weighted and utility ARM, spatio-temporal ARM, incremental ARM, fuzzy ARM, etc. Frequent pattern mining remains a common area of investigation within the domain of data mining. Resent work on frequent pattern mining has been directed at recommender systems (people who bought *x* also bought *y*). The most popular current frequent pattern mining algorithm is arguably frequent pattern growth (Han *et al.*, 2000).

2.2 Clustering

Clustering is concerned with the grouping of data into categories. This is particularly desirable in the context of customer data where it is useful to group similar customers together for the purpose of (say) targeted advertising. For many concerns clustering is an exploratory activity. Typically, we wish to cluster data into either a specified number of clusters, as in the case of the well-known K-means algorithm (MacQueen, 1967); or according to some proximity threshold, as in the case of the well-established KNN algorithm (Hastie and Tibshirani, 1996). An alternative approach is to adopt some form of hierarchical clustering where the data are iteratively partitioned to form a set of clusters. The most frequently sighted hierarchical clustering algorithm is arguably BIRCH (Zhang *et al.*, 1996). The 'goodness' of a cluster configuration is usually measured in terms of intra-cluster cohesion and inter-cluster separation. The issues with established clustering algorithms, such as K-means and KNN, are that the generated clusters are represented as hyperspheres when this may not be the ideal shape. Further issues are: the frequently encountered high dimensionality of the input data and the treatment of noise (*outliers*) and categorical data. Clustering is a well-established data mining (and before that machine learning) technique. Interestingly, there is no 'best' clustering algorithm applicable to all data; instead, for reasons that are not entirely clear, some algorithms work better on some data sets than others.

2.3 Classification

Classification is concerned with the construction of 'classifiers' that can be applied to 'unseen' data so as to categorize that data into groups (*classes*). As such classification has parallels with clustering. The distinction, however, is that classification requires pre-labelled training data from which the classifiers can be built. As such classification is sometimes referred to as *supervised learning* while clustering is considered to represent *unsupervised learning*. The desired classifiers, can take many forms: decision trees, support vector machines (SVMs) as first proposed by Vapnik (1995), rules, etc. Decision trees are the simplest. The most influential decision tree generation algorithm with respect to data mining is Quinlan's C4.5 algorithm (Quinlan, 1993). The advantage of rule-based classifiers is that they offer a ready explanation to end users. In the context of rule-based classifiers, classification rules can be considered to be a special form of AR and as such ARM techniques (see above) can be used to generate such rules. The most frequently referenced classification ARM algorithm is arguably the CBA algorithm (Liu *et al.*, 1998). Other notable classification techniques include regression, for example, the CART algorithm (Breiman *et al.*, 1984) and Naive Bayes (Hand and Yu, 2001). Classifiers can be either (i) binary classifiers (select between two alternatives), (ii) multi-class classifiers (select between more than two alternatives); or (iii) multi-labelled (assign unseen data to one or more classes). Binary classifiers are the simplest to generate. The quality of a generated classifier is usually measured in terms of accuracy, sensitivity and specificity. To an extent similarities can be drawn between classification and case-based reasoning, both operate using previous cases or knowledge. Classification continuous to receive attention from the data mining community. One extension is the concept of ordinal classifiers where the possible classes are ordered in some way. There is also significant interest in dynamic classification, for example, classifying video sequences.

3 Applications

From the foregoing, the original focus of data mining was tabular data; an extremely effective set of techniques has been established directed at the mining of tabular data, however data miners wish to mine everything! This section briefly reviews some current applications of the technology beyond simple tabular mining. There are, of course, many more.

3.1 Text mining

A Natural next step from traditional tabular data mining was text mining. A typical application is to build classifiers to categorize or cluster large document collections (news articles are a popular example, another is web pages). Another application is opinion or questionnaire ming where the objective is to obtain useful informations, that is, 'opinions', from the free text element of questionnaire style data. A further application is text summarization, an application that starts to 'blur' into the domain of information retrieval. In the context of text classification, SVMs operate well (but offer no explanation of resulting classifications). Generally speaking, the issue with text mining is how best to represent textual data so as to allow the application of data mining techniques. The most common representation is the bag-of-words representation where documents are represented in terms of a collection of keywords. The question then is what keywords to include? These can be defined by experts, or extracted using other data mining techniques or natural language processing (NLP) techniques. An alternative to the bag-of-words representation is the bag-of-phrases representation. However, in both cases, the ordering of words/phrases is lost. Alternative techniques attempt to maintain this knowledge, however this entails a significant increase in computational complexity. Text mining, in all its forms, continues to be a popular data mining activity.

3.2 Image Mining

There are many large collections of digital images that have been generated with respect to many applications. As in the case of text mining, image mining is concerned with the representation of images (both 2D and 3D) so that mining techniques may be applied. For this purpose, images can be represented in many different ways, popular techniques include the generation of histograms or trees/graphs (one per image). Alternatively, we can attempt to represent images in terms of sets of objects identified using segmentation and registration techniques. Image segmentation techniques have limited success, depending on the nature of the images, and are the subject of continuing research within the image analysis community. Image analysis remains a challenging research topic (we are still unable to get a machine to distinguish between a cat and a dog with any degree reliability). In certain fields, such as medical image mining, where the problem can be scoped in a specific way, image mining has had some successes. Examples include the classification of retina image data and magnetic resonance imaging scan data to identify disorders. Another popular area of application is satellite image mining. Current research in image mining continues to be focused on how best to represent images so that data mining techniques can be applied. In this respect, it is worth observing that for the application of data mining techniques, we do not need to have a representation that is interpretable by humans, as long as the data mining works (e.g. we do not necessarily need precise segmentation techniques).

3.3 Graph mining

Graph (and tree) mining is essentially an extension of frequent pattern mining (see above), what we are interested in is frequently occurring sub-graphs. Graph mining practitioners argue that everything can be represented as a graph. Indeed it is straight forward to see how entities such as documents, emails and images can be represented in this form. A common application area is chemical compound analysis. At a high level, we can identify two forms of the problem: (i) frequent sub-graphs that occur across a collection of graphs and (ii) frequent sub-graphs that occur in one very large graph. We can also distinguish between graph mining and tree mining; tree mining is more tractable as advantage can be taken of the inherent features of a tree (no cycles, etc.). Graph (and tree) mining require some canonical form with which to represent the graphs;

much early work was focussed on this. The main current issues with graph mining are candidate subgraph generation and sub-graph isomorphism testing. The most influential frequent sub-graph mining algorithm is arguably gSpan (Yan and Han, 2002). A popular extension of graph mining is social network mining. The motivation here is the popularity of social networking sites such as Facebook, and the consequent desire to identify groupings (communities) within these networks. However, there are many other forms of social networks, such as transport and co-authoring (bibliographic) networks, to which social network mining techniques can be applied.

4 Conclusions

Data mining has come to prominence over the last two decades as a discipline in its own right which offers benefits with respect to many domains, both commercial and academic. Broadly, data mining can be viewed as a application domain, as opposed to a technology. The increasing ability of institutions to collect electronic data, facilitated by advanced computer processing, means that the desire to 'mine' data is likely to expend. The data mining community has a well-established set of techniques available, which we are seeking to apply to an ever greater variety of data. Generally speaking, the actual data mining processes, in many cases, are readily available. Current issues are more concerned with the processing of data so that data mining techniques can be applied, and the post-processing (e.g. visualization, explanation generation, etc.) of the end result. Thus, although we are very good at the actual data mining, the 'end-to-end' process of data mining still requires significant research input. Another driver for research in data mining is the ever increasing size of the data we wish to work with. We are therefore also interested in techniques to mine ever larger data sets (and an ever greater variety of data).

References

Agrawal, R., Imielinski, T. & Swami, A. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'93)*, ACM Press, 207–216.

Breiman, L., Friedman, Y., Olshen, R. & Stone, C. 1984. *Classification and Regression Trees*. Wadsworth.

Fayyad, U., Piatetsky-Shapiro, H. & Smyth, P. 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* **39**(11), 27–34.

Han, J., Pei, J. & Yin, Y. 2000. Mining frequent patterns without candidate generation. In *Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD '00)*, ACM Press, 1–12.

Hand, D. J. & Yu, K. 2001. Idiot's Bayes: not so stupid after all? *International Statistical Review* **69**, 385–398.

Hastie, T. & Tibshirani, R. 1996. Discriminant adaptive nearest neighbor classification. *IEEE Transaction on Pattern Analysis and Machibe Intelligence* **18**(6), 607–616.

Liu, B., Hsu, W. & Ma, Y. M. 1998. Integrating classification and association rule mining. In *Proceedings of the Knowledge Discovery and Data Mining-98*, ACM Press, 80–86.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium Mathematical Statistics and Probability*. University of California Press, Berkeley, CA, USA, 281–297.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag.

Yan, X. & Han, J. 2002. gSpan: graph-based substructure pattern mining. In *Proceedings of the IEEE International Conference on Data Mining (ICDM '02)*, IEEE, 721–724.

Zhang, T., Ramakrishnan, R. & Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, ACM Press, 103–114.