Available online at www.sciencedirect.com

ScienceDirect

Physics Procedia 62 (2015) 18 – 22

Physics Procedia

3rd International Conference Frontiers in Diagnostic Technologies, ICFDT3 2013

From Patterns in Data to Knowledge Discovery: What Data Mining Can Do

Francesco Gullo*

Yahoo Labs
Barcelona, Spain
gullo@yahoo-inc.com

Abstract

Data mining is defined as the computational process of analyzing large amounts of data in order to extract patterns and useful information. In the last few decades, data mining has been widely recognized as a powerful yet versatile data-analysis tool in a variety of fields: information technology in primis, but also clinical medicine, sociology, physics.

In this technical note we provide a high-level overview of the most prominent tasks and methods that form the basis of data mining. The note also focuses on some of the most recent yet promising interdisciplinary aspects of data mining.

Peer-review under responsibility of the ENEA Fusion Technical Unit

Keywords: data mining, knowledge discovery, graph mining

1. Knowledge Discovery in Databases and Data Mining

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying novel, valid, potentially useful, and ultimately understandable patterns in data Fayyad et al. (1996a). The term "pattern" refers to a subset of the data expressed in some language or a model exploited for representing such a subset. KDD aims at discovering patterns that *(i)* do not result in straightforwardly computing predefined quantities (i.e., non-trivial), *(ii)* can apply to new data with some degree of certainty (i.e., valid), *(iii)* have been unknown so far (i.e., novel), *(iv)* provide some benefit to the user or to further tasks (i.e., potentially useful), and *(v)* lead to useful insights, immediately or after some post-processing (i.e., understandable).

The KDD process is an iterative and interactive sequence of the following main steps (Figure 1):

- *selection*, whose main goal is to create a target data set from the original data, i.e., selecting a subset of variables or data samples, on which discovery has to be performed;
- *preprocessing*, which aims to "clean" data by performing various operations, such as noise modeling and removal, defining proper strategies for handling missing data fields, accounting for time-sequence information;

* Corresponding author. Tel.: +34-93-183-8891 ; fax: +34-93-183-8901.
E-mail address: gullo@yahoo-inc.com

Peer-review under responsibility of the ENEA Fusion Technical Unit
doi:10.1016/j.phpro.2015.02.005

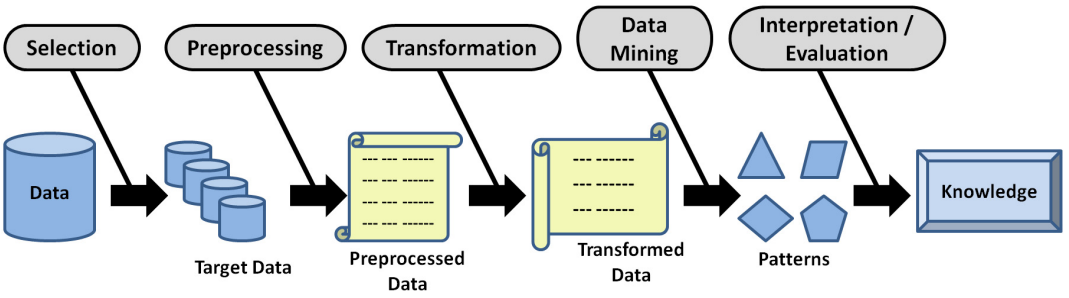

Fig. 1. The Knowledge Discovery in Databases (KDD) process

- *transformation*, which is in charge of reducing and projecting the data, in order to derive a representation suitable for the specific task to be performed; it is typically accomplished by involving transformation techniques or methods that are able to find invariant representations of the data;
- *data mining*, which deals with extracting interesting patterns by choosing *(i)* a specific data-mining *method* or *task* (e.g, summarization, classification, clustering, regression, and so on), *(ii)* proper *algorithm(s)* for performing the task at hand, and *(iii)* an appropriate representation of the output results;
- *interpretation/evaluation*, which is exploited by the user to interpret and extract knowledge from the mined patterns, by visualizing the patterns; this interpretation is typically carried out by visualizing the patterns, the models, or the data given such models and, in case, iteratively looking back at the previous steps of the process.

Data mining represents the "core" step of the KDD process, so much so that the "data mining" and "KDD" terms are often treated as synonyms Han and Kamber (2001). Several definitions of what data mining is have been used, e.g., "*automated yet non-trivial extraction of implicit, previously unknown, and potentially useful information from data*", "*automated exploration and analysis of large quantities of data in order to discover meaningful patterns*", "*computational process of automatically extracting useful knowledge from large amounts of data*". All definitions are all roughly equivalent to each other. They all agree on the main aspects of data mining, which are: *(i)* huge quantity of data that *(ii)* should be analyzed so as to *(iii)* extract what is called "knowledge", or "useful information", or "patterns", i.e., *(iv)* something that can be processed and profitably exploited by human beings.

The importance of data mining nowadays is mainly motivated by the lots of data that is collected and stored by a variety of today's prominent applications. This data includes Web data, e-commerce data, purchases, bank transactions, and so on. Also, the number of applications dealing with data that needs to be processed at enormous speeds (GB/seconds or even more) is rapidly increasing; examples include remote sensors on satellite, telescopes scanning the skies, microarray generating gene-expression data, scientific simulations. Due to the peculiarity of the underlying data, it is apparent that data analysis in such challenging contexts cannot be performed with traditional data-analysis techniques, either manual or automated. Data mining aims at filling this gap, with its intrinsic interdisciplinary nature that poses it at the intersection of a number of more classical fields, such as artificial intelligence, statistics, database systems, machine learning.

2. Data-Mining Tasks

Data mining comprises a number of tasks that can be used, even in combination, based on the requirements of the specific application context. Data-mining tasks are usually classified into *predictive* and *descriptive* Fayyad et al. (1996b). Predictive tasks refer to building a model useful for predicting future behavior or values for certain features. Among others, these include *classification* and *prediction*, i.e., deriving some models (or functions) that describe data classes or concepts by a set of data objects whose class label is known (i.e., the *training set*), so as to predict the class of objects whose class label is unknown; *deviation detection*, i.e., dealing with *deviations* in data, which are defined as differences between measured values and corresponding references such as previous values or normative values; *evolution analysis*, i.e., detecting and describing regular patterns in data whose behavior changes over time. On the

other hand, in a descriptive data-mining task, the built-in model aims at describing the data in an understandable, effective, and efficient form. Relevant examples of descriptive tasks are *data characterization*, whose main goal is to summarize the general characteristics or features of a target class of data; *data discrimination*, i.e., a comparison of the general features of a target class of data objects with the general features of objects from a set of contrasting classes; *association-rule discovery*, i.e., discovering rules that show attribute-value conditions occurring frequently together in a given set of data; and *clustering*, which aims at forming high-cohesive and well-separated groups of objects from the input set of data objects.

2.1. Classification

The classification task takes as input a collection of records, called *training set*, where each record is composed of a set of *attributes*, and one of the attributes denotes the *class* of the record. The goal is to find a *model* for the class attribute as a function of the values of the other attributes. The model is then used to predict the class attribute of previously unobserved records.

As an example, consider a collection of records describing the position held by the academic staff of some university. Assume that each record has the following attributes: *(i) name* of the professor, *(ii) position* (i.e., assistant professor, associate professor, or full professor), *(iii)* number of *years* she has been affiliated to such a university, and *(iv)* the class attribute, that is a boolean attribute that indicates whether the professor holds a *tenured* position or not. Assume also that the input collection contains the following four records: ⟨Mike, Assistant Prof, 3, no⟩, ⟨Mary, Assistant Prof, 7, yes⟩, ⟨Bill, Full Prof, 2, yes⟩, ⟨Anne, Associate Prof, 7, yes⟩. Based on this input, a classification algorithm would likely find a model expressed by the following (set of) rule(s): "`IF` *position*=Full Prof `OR` *years* > 3 `THEN` *tenured*=yes". Thus, given a new record ⟨Barbara, Full Prof, 4, ?⟩, the model would predict the missing class value as yes.

Classification is a long-standing area of research where a plethora of different approaches and algorithms have been defined, including k Nearest Neighbors (KNN), decision trees, Support Vector Machines (SVM), neural networks, Gradient Boosted Decision Trees (GDBT) Kotsiantis (2007).

2.2. Clustering

Given a set of data objects, clustering aims at identifying a finite set of groups of objects, i.e., *clusters*, so that the objects within the same cluster are "similar" to each other, whereas the objects belonging to different clusters are "dissimilar". The degrees of (dis)similarity among data objects are computed and evaluated according to a proximity measure that can be either specified by the user or inherently incorporated in the specific clustering algorithm. In a clustering task there is no prior knowledge of the class labels associated to the objects to be grouped; for this reason, clustering is often also referred to as *unsupervised classification*, to emphasize the difference from the (supervised) classification task, in which the class labels of the objects in the training set are known.

A clustering of the input set of objects is thus built in such a way that cluster cohesiveness and separation, measured in terms of the underlying proximity measure, are maximized. More precisely, clustering methods typically define a specific objective function to be optimized, in order to formally define clusters that are compact and well-separated from each other. Since these formulations usually lead to computational problems too hard to be optimally solved for large-scale inputs (the so-called *NP-hard* problems), any specific clustering method should define the corresponding approximation/heuristic algorithm(s) to find good approximations of the optimal solution.

The literature abounds with different clustering approaches and algorithms, which differ to each other for the optimization criterion, the resolution strategy, and the computation of the distance between the input objects Aggarwal and Reddy (2014). These algorithms can be classified according to a lot of different taxonomies, which however usually all agree on the top-level division in two main categories, i.e., *partitional* (or *partitioning*) and *hierarchical*. Broadly, partitional-clustering algorithms compute a single partition of the input dataset. A considerable number of partitional algorithms exploits the *relocation* scheme, i.e., the objects are iteratively re-assigned to the clusters, until a stop criterion has met. Such a scheme is at the basis, e.g., of the popular *K-Means* algorithm MacQueen (1967). Rather than a single partition of the input dataset, hierarchical-clustering approaches output instead a *hierarchy* of clustering solutions that are organized into a tree-like structure known as *dendrogram* Aggarwal and Reddy (2014).

2.3. Association-rule discovery

Given a set of records (i.e., *transactions*), each of which containing a number of items from a given collection, the goal of association-rule discovery is to produce dependency rules that can predict occurrence of an item based on occurrences of other items.

As an example, think about an electronic-device shop where, for marketing reasons, one is interested in understanding the best way to expose items to customers in order to increase purchases. In this case, one can analyze the past purchasing history in order to discover association rules like $\{camera, tripod\} \rightarrow \{SD\ memory\}$, which informally states that, when customers buy a camera and a tripod, it is very likely that they buy an SD memory as well. Such a rule can be used in several ways. For instance, cameras and tripods can be used to boost the sales of SD memories by, e.g., storing the cameras and tripods close to SD memories or putting cameras in bundle promotion with tripods.

A preliminary step commonly required by association-rule-discovery algorithms corresponds to another classical data-mining task know as *frequent pattern mining* Han et al. (2007), whose main goal is to find subsets of items that co-occur frequently in a set of transactions. For instance, the above example association rule $\{camera, tripod\} \rightarrow \{SD\ memory\}$ would derive from the preliminary discovery that cameras, tripods and SD memories frequently appear together in a purchasing data log.

3. Graph Mining

Even being general enough to handle any type of data, data mining can however be "customized" to deal with a specific typology of data and be focused on specific data peculiarities. For instance, when data mining meets the Web, we talk about *web mining* Liu (2006). Similar examples include *XML mining* Tagarelli (2011) and *uncertain data mining* Aggarwal (2009).

Specifically, *graph mining* Aggarwal and Wang (2010) is the set of data-mining methods and algorithms used to deal with *graphs*. In its basic form, a graph is a pair $G = (V, E)$, where V is a set of vertices and $E \subseteq V \times V$ is a set of edges expressing relationships between vertices. This basic model can be enhanced in several ways: the edges can have an orientation (i.e., they can be *directed*), and vertices and/or edges can be coupled with additional information such as weights, labels, timestamps, probability of existence, feature vectors, and so on. Graphs provide a general framework for modeling real-world data. They are routinely used to represent data in a wide variety of contexts, such as computational biology (e.g., protein-interaction networks), chemical data analysis (e.g., chemical compounds), communication networking (e.g., device networks, road networks), social network analysis, Web link analysis, and so on.

Due to this large availability of graph data, graph mining has become a prominent data mining sub-field whose appeal is continuously increasing. Prominent graph-mining tasks include graph clustering, graph search, dense-subgraph extraction, graph classification, graph pattern mining, graph matching, graph querying, influence maximization Aggarwal and Wang (2010). Particularly, graph clustering is the problem of partitioning a given graph into a set of subgraphs so as to optimize some criterion that takes into account intra-cluster density and/or cluster separation. The notions of density usually adopted include modularity, average degree, ratio cut, normalized cut, conductance, and many more Leskovec et al. (2010). Graph clustering finds application into a plethora of scenarios, such as community detection in social networks, identification of high-cohesive structures in biological networks, packet delivery in communication networks, detecting highly-correlated stocks. Graph search instead deals with the following problem: given a set of graphs $\mathcal{G} = \{G_1, \ldots, G_n\}$ and a query graph Q, find all graphs in $\mathcal{G}$ that are supergraphs of Q. Graph search is commonly required in several contexts too, e.g., chemical compound search, context-based image retrieval, and 3D protein structure search.

4. Conclusions

Data mining is a powerful tool that has been used for decades for advanced analysis of large quantities of data. It is defined as the automated yet non-trivial extraction of implicit, previously unknown, and potentially useful information from data.

This technical note provided a broad overview of the main data-mining principles and its interdisciplinary aspects. We also focused on an emerging data-mining subfield, that is graph mining, which deals with the problem of mining data represented as graphs, i.e., as sets of interconnected objects.

References

Aggarwal, C.C., 2009. Managing and Mining Uncertain Data. Springer.

Aggarwal, C.C., Reddy, C.K., 2014. Data Clustering: Algorithms and Applications. CRC Press.

Aggarwal, C.C., Wang, H., 2010. Managing and Mining Graph Data. Springer.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., 1996a. Knowledge Discovery and Data Mining: Towards a Unifying Framework, in: Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), pp. 82–88.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., 1996b. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

Han, J., Cheng, H., Xin, D., Yan, X., 2007. Frequent Pattern Mining: Current Status and Future Directions. Data Mining and Knowledge Discovery (DAMI) 15, 55–86.

Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques. Academic Press.

Kotsiantis, S.B., 2007. Supervised Machine Learning: A Review of Classification Techniques, in: Proc. Conf. Emerging Artificial Intelligence Applications in Computer Engineering, pp. 3–24.

Leskovec, J., Lang, K.J., Mahoney, M., 2010. Empirical Comparison of Algorithms for Network Community Detection, in: Proc. Int. Conf. on World Wide Web (WWW), pp. 631–640.

Liu, B., 2006. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer.

MacQueen, J.B., 1967. Some Methods for Classification and Analysis of MultiVariate Observations, in: Proc. Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297.

Tagarelli, A., 2011. XML Data Mining: Models, Methods, and Applications. IGI Global.