Abstract— Knowledge discovery is an emerging field where we combine techniques from algorithmics, artificial intelligence, mathematics and statistics to deal with the theoretical and practical issues of extracting knowledge, i.e., new concepts or concept relationships hidden in volumes of raw data. Knowledge discovery offers the capacity to automate complex search and data analysis tasks.

Data mining is the main phase in the knowledge discovery process. It consists in extracting nuggets of knowledge, i.e., pertinent patterns, pattern correlations, estimation or rules, hidden in bodies of data. The extracted nuggets of knowledge will be used in the verification of hypothesis or the prediction and explanation of knowledge.

In this paper, we present a new survey on Knowledge Discovery and Data Mining (KDD).

Index Terms— Data Mining, Knowledge Discovery.

I. INTRODUCTION

Knowledge Discovery and Data Mining (KDD) is the process of non-trivial discovering, from data, of implicit informations, previously unknown and potentially interesting [39]. By non-trivial is meant that the identification was done thanks to a research or an induction.

In many fields, such as Health, Geology, Marketing, Finance and Molecular Biology, the size and the complexity of data is growing exponentially [12,35,39,49,60,68,93,105,111]. Indeed, concerning the biological data, for example, the Human Genome Project is providing the sequence of the three billion DNA bases that constitute the human genome. Consequently, we are also provided with the sequences of about 100 000 proteins. Therefore, we are entering the post-genomic era: after having focused so much efforts on the accumulation of data, we have now to focus as much effort, and even more, on the analysis of these data. This will enable us to learn more about gene expression, protein inter-actions and other biological mechanisms. Analyzing this huge volume of data is a challenging task because, not only, of its complexity and its multiple numerous correlated factors, but also, because of the continuous evolution of our understanding of the biological mechanisms. Classical approaches of biological data analysis are no longer efficient and produce only a very limited amount of information, compared to the numerous and complex biological mechanisms under study. Actually, these approaches use only a very limited number of parameters, to represent the so-many correlated factors involved in the biological mechanisms. From here comes the necessity to use computer tools and develop new in silico high performance approaches to support us in the analysis of biological data and, hence, to help us in our understanding of the correlations that exist between, on one hand, structures and functional patterns of biological sequences, i.e. DNA, RNA and proteins, and, on the other hand, genetic and biochemical mechanisms. KDD are a response to these new trends [12,29,59,102,103,107].

KDD is a process made up of three main phases:
- Phase of data preprocessing,
- Phase of data processing, also called data mining,
- And phase of data postprocessing [12, 35, 39, 49, 60, 68, 93,105,111].

In this paper, we present a new survey on KDD. In the second section, we discuss the phase of data preprocessing. In the third one, we discuss the phase of data mining. In the fourth one, we discuss the phase of data postprocessing. Finally, in the last one, we present a conclusion.

II. DATA PREPROCESSING

Data preprocessing [30,34,191] consists in:
- Understanding the domain of application and the acquired knowledge, and identifying the goal of the KDD process from the user’s point of view.
- Creating a target data set, i.e., selecting a sample of data on which the KDD must be applied.
- Cleaning the data, e.g., removing the errors, eliminating the redundant data and/or completing the missing data.
Indeed, data may contain errors coming from the experiments and/or the measurements done [9,47].

- Coding the data, i.e., extracting attributes and selecting from them those which are pertinent, in order to code the data according to the goal to reach [72,112].
- Choosing the data mining task to achieve, e.g., prediction [107,109], clustering [4,74,104,106,108], estimation [86,94] and classification [65,100,101].
- Choosing the data mining algorithm to apply, according to the task to achieve.

III. DATA MINING

Data mining is the application of specific algorithms for the extraction of nuggets of knowledge, i.e., pertinent patterns, pattern correlations, estimation or rules, from data [35,125,37]. According to Fayyad et al. [35], data mining tasks are:
- Detection of the changes and deviations: this task consists in locating the changes and deviations occurred on the starting data.
- Modeling of the dependences: this task consists in modeling the dependences existing between attributes.
- Abbreviation: this task consists in defining a shortened description of a subset of data.
- Regression: this task consists in examining the dependence between random variables, called dependent variables, and other random ones, called independent variables [94,54]. This dependence is represented by an equation called regression equation [29,38]. Among the techniques allowing to make regression, we quote Bayesian techniques [14,15,17,82,83].
- Clustering: this task, also called unsupervised classification, consists in forming classes starting from a set of data. The classes are formed so that two elements of the same class resemble each other much more than two elements of two different classes. We distinguish two main types of clustering [4,8,13,16,24,57]:
  - Hierarchical clustering: By making hierarchical clustering, we build a tree of classes called dendrogram. Each node in this tree represents a class. The children of a node represent the subclasses of the class represented by the node father. Among the techniques allowing to make hierarchical clustering, we quote the technique based on link metrics [25,26,87,98], the one of hierarchical classes of arbitrary forms [41,45,46] and the one of divisive binary partitioning [107,99,100].
  - Unhierarchical clustering: By making unhierarchical clustering, we build classes that are not subclasses of each other. Among the techniques allowing to make unhierarchical clustering, we quote the technique of k-means [50,51], the one of k-medoids [32,62,113,115], the one of probabilistic clustering [114,116,117,118] and density-based one [119,120,121,122].
- Classification: this task, also called supervised classification, consists in assigning a new element to one of the predefined classes.

The main techniques used to carry out the tasks presented above are:

---

**Fig. 2 Case-based reasoning cycle**

- Case-Based Reasoning: it is a matter of a reasoning which is made starting from old experiments in order to solve a new problem. We operate as follows [123,124,125]:
  - During the first step, we propose a solution.
  - During the second step, either we adapt this solution to the needs for the new problem or we justify it.
  - During the third step, we criticize the solution obtained then we evaluate it. If this solution is satisfactory then we save it, if not we return to the second step, in order to adapt it.

Case-based reasoning is used to make classification [126,127], to make prediction [128] and to make clustering [129].

**Fig. 3 Example of a decision tree.**

- Links analysis: it is a matter of an analysis which consists in connecting entities (customers, companies ...) with each other by links. With each link a weight, which quantifies the force of this connection, is assigned [130]. Links analysis is used to make prediction and to make classification [130,131,132,133,134].
- Decision trees: a decision tree is a tree such that each internal node represents a test on the value of one or more attributes. According to the result of this test, we pass to one of the children of this node. In order to process an element, we start from the root of the tree. We stop when a leave is reached, i.e., a node representing no test. We say then that a decision is taken. Hence the name of this tree [43]. Decision trees are used to make classification [10,135,95,136,76,42] and to make regression [10].
- **Association rules**: an association rule is a rule of the form « if \( C \) then \( P \) », where \( C \) is the condition of the rule and \( P \) is the prediction. The condition \( C \) is a conjunction of terms and the prediction \( P \) is a simple term. A term is of the form (attribute operator value) \([59,66]\). Association rules are used to make classification \([71,73,99,137]\) and to make clustering \([138]\).

**Example**: In the decision tree of Fig. 3, a patient having a temperature strictly lower than 37.5 and having an irritated throat, for example, will be classified as sick by this tree. Thus, the translation of this decision tree in association rules is as follows:

\[
\begin{align*}
\text{if (temperature < 37.5) and (throat = irritated)} & \quad \text{then patient = sick} \\
\text{if (temperature < 37.5) and (throat ≠ irritated)} & \quad \text{then patient = in good health} \\
\text{if (temperature ≥ 37.5) then patient = sick}
\end{align*}
\]

- **Genetic algorithms**: A genetic algorithm is an algorithm whose principle of functioning is as follows: We start with an initial population of potential solutions, arbitrarily selected. We evaluate their respective performances. On the basis of these performances, we create a new population of potential solutions by using simple evolutionary operators, i.e., selection, crossing-over and mutation. We reiterate this process until we obtain a satisfactory solution \([40,44,55,59]\). Genetic algorithms are used to make classification \([3,64,88]\) and to make clustering \([18, 31,33,58]\).

- **Artificial neural networks**: an Artificial Neural network (ANN) is a system made up of several simple calculating units functioning in parallel, whose function is determined by the structure of the network, the solidity of the connections and the operation carried out by the elements or nodes. An ANN resembles the human brain on two aspects \([52]\):

(a) Knowledge is acquired by the ANN through a learning process.

(b) Connections between the neurons, known under the name of synaptic weights, are used to store knowledge.

ANN are used to make classification \([5,27,41,56,96]\), to make clustering \([41,53]\), to make prediction \([111]\) and to make regression \([56,75]\).

- **Support Vector Machines**: a Support Vector Machine (SVM) is a discrimination technique. It consists in separating two, or more, sets of points by a hyperplane. It is based on the use of functions, known as core (kernel), which allow an optimal separation of the points of the plane in various categories. Unlike the learning approaches, which are based on the minimization of the empirical risk, this technique is based on the principle of minimization of the structured risk. SVM are used to make classification \([19,22,23]\) and to make regression \([67,89,97]\).
During the first step, we compute the distance between a sample \( y \) and each training sample \( x \), by using an Euclidean distance.

During the second step, we identify the \( K \) samples which are the closest to \( y \) and choose the majority class among these \( K \) samples. KNN algorithms are used to make classification [20,48].

IV. DATA POSTPROCESSING

Data postprocessing of the nuggets of knowledge extracted during the data mining phase consists in [11]:

(a) Filtering the nuggets of knowledge in order to keep only those which are pertinent compared to the problem in question [69,70,95].

(b) Interpreting and explaining the nuggets of knowledge in order to make them understandable, if they have to be used by a human, or to express them in a suitable formalism, if they have to be used by a machine. In the case where the nuggets of knowledge have to be used by a human, techniques of visualization are used [2,6,21,36,63,77,78,90].

(c) Evaluating the nuggets of knowledge. This can be made, for example, by calculating the classification error rate or by evaluating the association rules via numerical indicators, e.g. measurements of interest [1].

(d) Integrating the nuggets of knowledge. This consists in combining these nuggets in order to obtain knowledge, i.e. new concepts or concept relationships hidden in volumes of raw data [11,95].

V. CONCLUSION

In this paper, we presented a new survey on Knowledge Discovery and Data Mining (KDD). We presented the different phases of KDD: data pre-processing, data processing, also called data mining, and data post-processing. The most important phase among these three phases is data mining. We provided the different tasks of data mining and the main techniques used to carry out these tasks. The techniques presented in this paper are quite general. There are other data mining techniques specialized for particular types of data and domain.

REFERENCES


[93] P. N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson Addison Wesley (Publish.) : (May, 2005).


[113] R. Ng, J. Han, Efficient and Effective Clustering Methods for Spatial Databases, in Proc. VLDB’94, (Santiago, Chile) : (1994), p144-155.


[137] X. Yin, J. Han, CPAR: Classification based on Predictive Association Rules, SDM’03 : (2003).


Fauzi Mhamdi received the Licence’s Degree in Computer Science in 1999 from the Faculty of Sciences of Tunis, Tunisia. Then, he received a Master’s Degree in Computer Science from the National School of Computer Science, Tunis, Tunisia. Now, he is preparing a PhD Degree in Computer Science at the Faculty of Sciences of Tunis. His main research area is Knowledge Discovery in Databases and its application in Bioinformatics.