ISSN: 2582-8118 ***Volume 3, Issue 6; Nov 2023***

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYSIS IN SCIENCE AND ENGINEERING

Web: https://www.iarj.in/index.php/ijrase/index

Interaction and Scalability in Data Mining

Dr. Sumangala Patil

Professor, Computer Science & Engineering Department,
Faculty of Engineering & Technology (Co-education),
Sharanabasva University, Kalaburagi, Karnataka..

ABSTRACT

We are able to gather, store, and share vast amounts of data on a massive scale because to technological advancements in hardware and software. Data mining is the process of automatically identifying patterns in these massive amounts of data and extracting hidden knowledge. Research, marketing, financial analytics, and other application fields are just a few of the domains in which data mining technology finds usage outside of business intelligence. Finding designs in large record sets using computer is a process known as records mining. Extraction of information from a record collection and its subsequent transformation into a comprehensible framework for additional usage is the overall objective of the records mining technique. In actuality, information exploration is the analytical step of the KDD, or "understanding invention in data banks" technique. However, there are several computational obstacles in terms of processing time, memory, bandwidth, and power consumption when attempting to extract knowledge in the form of patterns from large data volumes. These difficulties have prompted the creation of distributed and parallel data analysis techniques as well as the use of cloud and grid computing. The purpose of knowledge discovery/data mining (KDD) and human-computer interaction (HCI) is to support human intelligence with machine intelligence. We shall talk about scalability and interaction in data mining in this paper.

KEYWORDS:

Interaction, Scalability, Data Mining, Cloud Computing, Human-Computer, Interaction, Knowledge Discovery/Data Mining, Vertical Scalability, Horizontal Scalability.

1. Introduction:

Data Mining:

Data mining is the process of removing information from large data sets in order to find patterns, trends, and pertinent data that would enable the company to make data-driven decisions.

Data mining is also known as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging in the area of computer science. [1]

In the context of data mining, scalability pertains to an algorithm's capacity to process vast volumes of data in an efficient and productive manner. This implies that the algorithm should be able to handle the input quickly without compromising the accuracy of the output. To put it another way, a scalable data mining algorithm should be capable of processing ever-larger data sets without requiring a major increase in processing power.

Scalabilities come in a variety of forms that are significant when it comes to data mining.

Vertical Scalability:

The ability of a system or algorithm to manage an increase in workload by adding additional computational resources, such as faster processors or more memory, is referred to as vertical scalability, also known as scale-up. In contrast, horizontal scalability entails expanding a distributed computing system's machine count to accommodate a rise in workload.

Vertical scalability is a useful technique for enhancing an algorithm's or system's performance, especially for applications whose computational resources are constrained. A system's ability to process more data and carry out more complicated calculations can be increased by adding additional resources, which can enhance the speed and precision of the output.

Horizontal Scalability:

The ability of a system or algorithm to manage an increase in workload by adding more machines to a distributed computing system is referred to as horizontal scalability, also known as scale-out. Vertical scalability, on the other hand, refers to expanding a single machine's computational capacity by adding additional RAM or faster processors.

An algorithm's or system's performance can be enhanced via horizontal scalability, especially for applications that demand a lot of processing power. The system's speed and accuracy can be increased by adding more machines so that the workload is split over more machines.

In reality, data mining is the process of uncovering previously hidden relationships between data objects through automated data evaluation techniques. Information stored in an information storehouse is often studied as part of information mining. Regression, differentiation, and focus are actually three of the main techniques for data exploration. Expertise Breakthrough in Databases (KDD), another name for records exploration, is the process of nontrivially extracting significant information from database data that is implied, previously unknown, and probably useful. Although knowledge invention and data exploration are frequently seen as synonymous terms in data banks (or KDD), data exploration is actually a component of the know-how discovery process. [2]

Mining Methodology and User Interaction Issues:

User Interface:

The information that is discovered by data mining tools is only valuable if the customer finds it interesting or more importantly, reasonable.

Diverse forms of knowledge extraction from databases: This problem is in charge of solving the challenges of covering a large amount of data to satisfy the needs of the client or customer. For a user, covering a wide range of knowledge discovery jobs becomes challenging due to varied information or methods.Knowledge mining that is interactive at several levels of abstraction: Interactive mining is essential because it allows the user to narrow down the search for patterns by supplying and fine-tuning data mining requests in response to the returned results.Including background knowledge: The primary function of background knowledge is to carry on the exploration process and point out any patterns or trends that were seen along the way.

Ad hoc data mining and data mining query languages: Ad hoc data mining tasks must be described using data mining query languages, which must be coupled with data warehouse query languages to provide users with access to them. Results of data mining are to be presented and visualized. In this case, the patterns or trends that are found must be represented visually and in high-level languages. The data mining system's user interface module facilitates communication between users and the system. Using a user interface enables the following features:

- Enter a data mining query task to communicate with the system.
- supplying details to aid in narrowing the search's focus.
- mining according to the findings of the intermediate data mining.
- Examine the schemas and data structures in databases and data warehouses.
- Analyze the patterns that were mined.
- See the patterns in various formats. [3]

Performance Challenges in Data Mining:

The creation and enhancement of business intelligence and data mining applications is highly profitable due to the need to extract valuable information in the form of patterns from massive amounts of data. Credit card transactions are just one way that computer systems record our lives. Loyalty reward programs keep track of our shopping preferences. CCTV cameras, GPS units in smartphones, and navigation apps track our movements and locations. The internet also records information about us through programs like Facebook, Twitter, Email, and blogs. According to the authors' estimates, the size of our digital universe will increase 44 times by 2020 compared to 2009. The scalability of data mining techniques to these massive data volumes is a common challenge in the field of data mining, as advancements in storage technology have made it possible to store all these data volumes at a relatively low cost. The issue of data mining techniques' scalability affects several scientific fields as well.

Approaches to Scaling up Data Stream Mining in Resource Constrained Environments:

We live in a time when tiny sensors and portable electronics can do tasks that not so long ago required highly capable systems. Such tiny devices can generate and/or receive data streams that are used for a variety of crucial purposes in fields like national security, astrophysics, and stock market analysis, to mention a few.

Two crucial realities necessitate that data streams be processed locally on-board small devices with little computing capability. In this chapter, we'll refer to these gadgets as resource-constrained environments. Experimental evidence has demonstrated that local data processing is a more energy-efficient option than transmitting data streams to a cloud or other high-power computational service.

Complex operations like data mining can be completed in environments with limited resources thanks to their current computational capabilities. There has been a constant need to do more complicated computing jobs, even with the computational capabilities of machines with limited resources continuing to increase.

As a result, we face what are known as relative resource limits, which occur when hardware technology advancements are insufficient to meet the demands of the needs of present applications. The big data phenomenon is actually a result of the integration of the application needs with the growth of massive streaming data volumes. [4]

Review of Literature:

Data interaction research aims to address the following issues: resolving the "information silo" issue in many enterprise application systems; facilitating the interoperability of diverse data resources among application systems; and offering global data views, global data authority views, and ideal data exchange services for the enterprise's application systems. Jensen Z (2021) [5]

We examine efforts to scale up data mining techniques in this research. We incorporate concepts that have been applied to many data mining applications, like clustering, classification, and feature and instance selection. Rather than particular strategies for a given algorithm, we are interested in approaches that may be used broadly to a variety of data mining algorithms.

We will focus particularly on evolutionary methods among the various data mining techniques because they have proven to be highly effective in numerous data mining tasks, but they are also highly susceptible to scaling up issues. There is a significant bottleneck in the evolutionary approach's fitness function evaluation. (Pal, 2006) [6] Scalability is by its very nature practical.

While theoretical studies help us understand scaling difficulties better, many key factors remain simply practical, such as delay for storage access and memory thrashing when huge amounts of memory are used.

It is extremely challenging to include problems with file systems, virtual memory managers, and job schedulers into a formal theory from a statistical perspective. Naturally, this does not imply that a certain theory is unimportant; rather, it indicates that, frequently, the programmer who is scaling up a data mining technique is just as crucial as the theory.

Provost and Kolluri (1999) distinguished three primary approaches for scaling up learning techniques in the setting of inductive algorithms: (i) creating quick algorithms, (ii) splitting the data, and (iii) applying a relational representation.

Since these three methods are independent of one another, they can be incorporated into any data mining technique. We suggest a novel taxonomy that shares certain similarities with Provost and Kolluri's. This work does not discuss the strategy based on the relational representation of the information because it is not applicable to most data mining techniques in general. [7]

Objectives:

- Data mining converts information into knowledge.
- Data gathering and preparation.
- sApplying data mining algorithms, and evaluating results.
- As data sets grow in size, variety, and complexity, expert systems for data mining need to be scalable, meaning they can handle increasing amounts of data and tasks without compromising performance or quality.

Research Methodology: This study's overall design was exploratory. The research paper is an endeavor that is founded on secondary data that was obtained from reliable online resources, newspapers, textbooks, journals, and publications. The research design of the study is mostly descriptive in nature.

Result and Discussion:

Scalability:

Elasticity in a system is characterized by its scalability. It's not limited to this significance, even though we frequently use it to describe a system's capacity for growth. As necessary, we can scale up, down, and out. How much network traffic your website, online service, or application receives will determine how successful it is? Oftentimes, especially in the beginning, people underestimate the amount of traffic that their system would receive. A server crash and/or a reduction in the quality of your service could come from this.

Horizontal Scaling:

Adding more nodes or machines to your infrastructure in order to meet growing demands is known as horizontal scaling (aka scaling out). Adding a server could be the answer if the program you are hosting on a server has outgrown its capabilities or capacity to handle traffic.

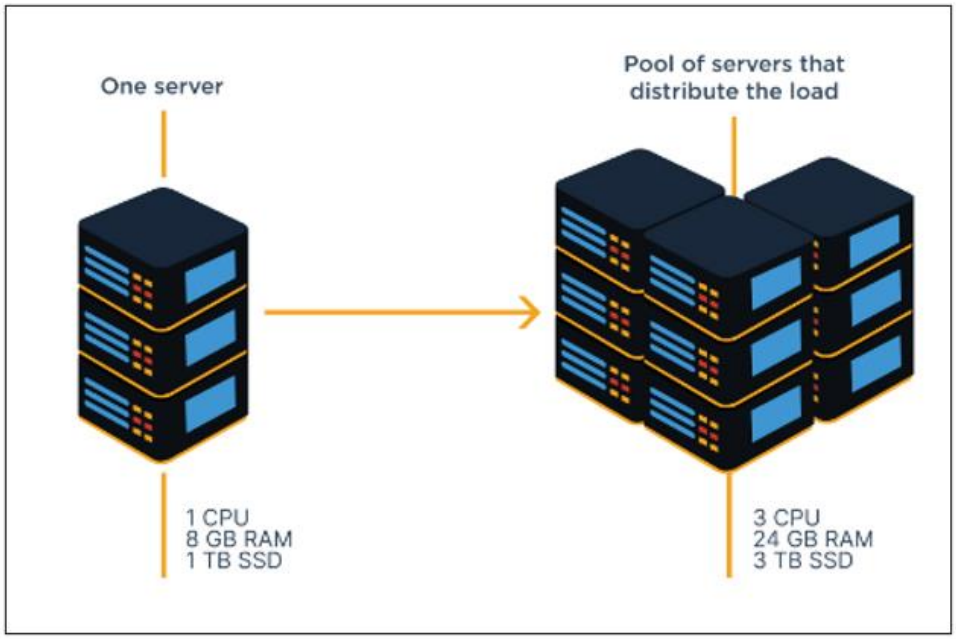

Figure 1: How Horizontal Scaling Works

Vertical Scaling:

Vertical scaling (aka scaling up), is the process of enhancing a system's capacity to meet demand. What distinguishes this from scaling horizontally.Vertical scaling describes increasing the power of your current machines, and horizontal scaling refers to adding more nodes. For instance, vertical scaling would include updating the CPUs if your server needed additional processing capacity. Moreover, you can scale memory, storage, or network speed vertically.

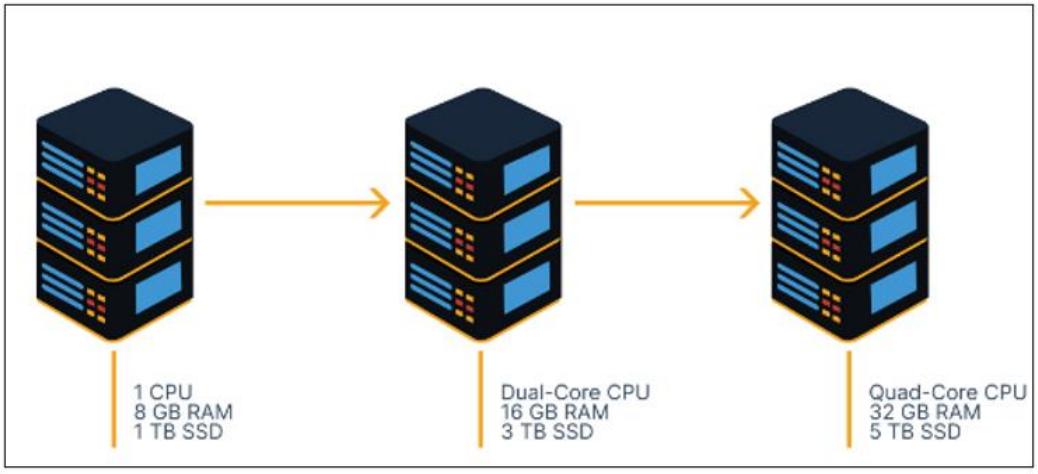

Figure 2: How Vertical Scaling Works

Table 1: Horizontal Vs. Vertical Scaling: [8]

	Horizontal scaling	Vertical scaling
Description	Increase or decrease the number of nodes in a cluster or system to handle an increase or decrease in workload	Increase or decrease the power of a system to handle increased or reduced workload
Example	Add or reduce the number of virtual machines (VM) in a cluster of VMs	Add or reduce the CPU or memory capacity of the existing VM
Execution	Scale in/out	Scale up/down
Workload distribution	Workload is distributed across multiple nodes. Parts of the workload reside on these different nodes	A single node handles the entire workload.
Concurrency	Distributes multiple jobs across multiple machines over the network, at a go. This reduces the workload on each machine	Relies on multi-threading on the existing machine to handle multiple requests at the same time
Required architecture	Distributed	Any
Implementation	Takes more time, expertise, and effort	Takes less time, expertise, and effort
Complexity and maintenance	Higher	Lower
Configuration	This requires modifying a sequential piece of logic in order to run workloads concurrently on multiple machines	No need to change the logic. The same code can run on a higher-spec device
Downtime	No	Yes
Load balancing	Necessary to actively distribute workload across the multiple nodes	Not required in the single node
Failure resilience	Low because other machines in the cluster offer backup	High since it's a single source of failure
Costs	High costs initially; optimal over time	Low-cost initially; less cost-effective over time

Purposes of Data Mining: Finding patterns and trends, forecasting future results, and enhancing decision-making are just a few of the vital functions that data mining performs. By maximizing marketing tactics, comprehending consumer behavior, and spotting irregularities or fraud, it gives organizations a competitive advantage. Data mining helps in disease prediction and therapy optimization in the healthcare industry. In scientific research, it plays a crucial role in deriving significant insights from intricate datasets. By converting unprocessed data into usable knowledge, data mining essentially spurs creativity and helps people make wise decisions in a variety of fields. [9]

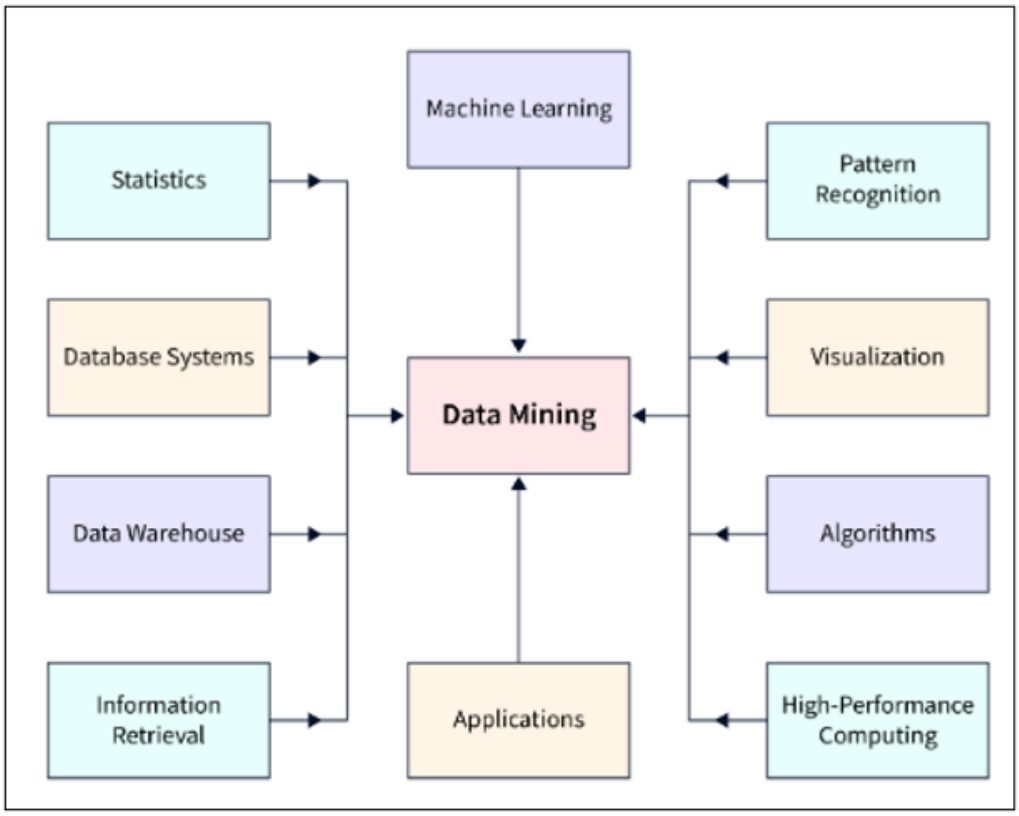

Figure 3: Purposes of Data Mining

The process of creating models or finding new patterns from a given dataset is collectively referred to as data mining. The KDD enterprise involves a number of phases, including data selection, data preparation and cleaning, data transformation and reduction, data-mining task and algorithm selection, post-processing, and knowledge interpretation. This KDD method is frequently quite interactive and iterative. A few of the KDD stages are shown in Figure 4. Prediction and description are the two main objectives of data mining. In prediction, our goal is to create a model that, using known values for some attributes in the database, will forecast future or unknown values of relevant attributes. In KDD applications, human-readable data description is just as crucial, if not more so, than prediction. Data mining can be divided into two primary categories. Verification-driven data mining involves the user formulating a hypothesis, which the system then attempts to verify.

Statistical analysis, multidimensional analysis, and query and reporting are examples of frequent verification-driven procedures. The focus of this essay is on discovery-driven mining, which automatically extracts fresh information.

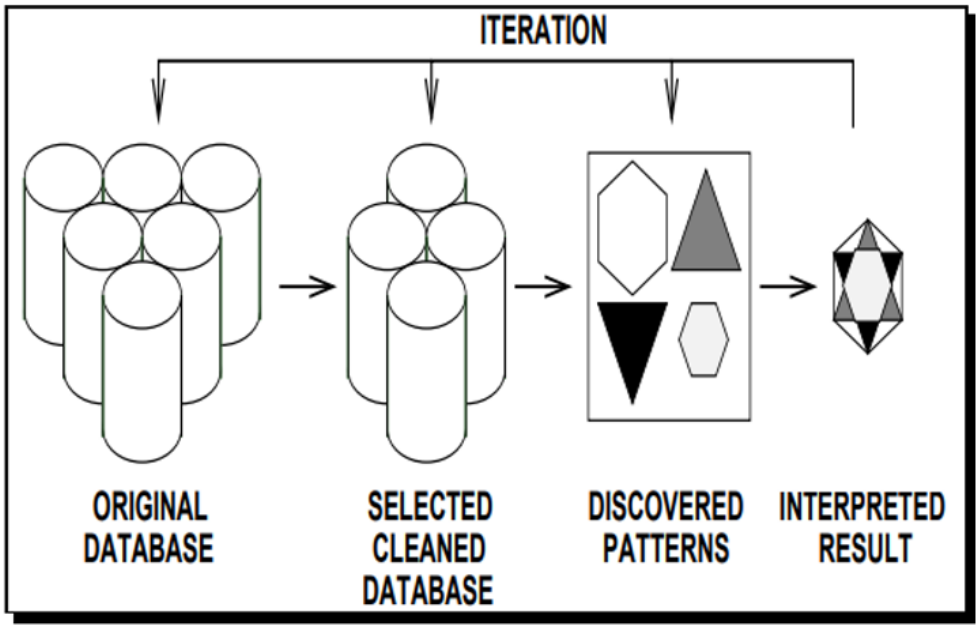

Figure 4: Data Mining Process

Applications such as customer segmentation, store layout, catalog design, telecom alarm diagnosis, etc., could be used. [10]

Data Mining Issues:

Let's explore three key data mining issues, as mentioned below -

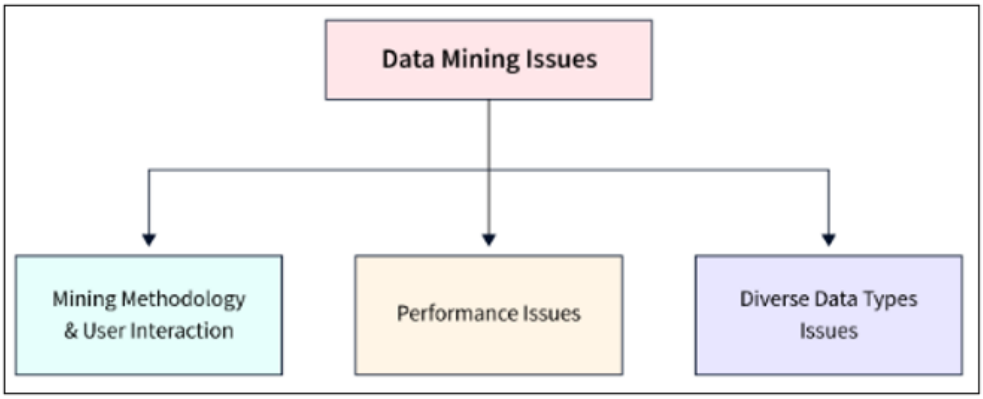

Figure 5: Data Mining Issues

Mining Methodology Issues:

Problems with the selection and use of mining algorithms and approaches are included under methodology-related data mining difficulties. Choosing the best approach for a given dataset and problem might be difficult. Achieving a balance between model complexity and accuracy is particularly important since problems such as overfitting, bias, and the requirement for interpretability frequently arise.

Performance Issues:

Large dataset handling, efficiency, and scalability are the main concerns in performance-related data mining. The exponential growth of data quantities necessitates the development of algorithms and infrastructure that can handle and analyze data quickly. Performance bottlenecks might make it difficult to use data mining techniques in situations that are actual.

Diverse Data Types Issue:

The various kinds of data The challenges associated with managing diverse data sources are brought to light by data mining problems. Integrating data from many formats, including text, photos, and structured databases, is a common task for data miners.

Every form of data has different preprocessing, feature extraction, and modeling issues, and to effectively address these complexities, specialized methods and tools are needed.

Scalability - Making sure data mining techniques and infrastructure can handle large-scale datasets efficiently becomes a critical challenge as data volumes continue to rise. [11]

Eliciting user requirements for Interaction Mining:

This section aims to present a case for interaction mining by examining the shortcomings of existing business analytics methods, such as text mining and opinion mining. We examine the fundamental elements of consumer interactions in greater detail for that reason. We differentiate between three major categories of consumer interactions based on our observations:

1. Straight communication between the client and the business. This kind of communication is either requested by the business through surveys or feedback forms, or it is started by the client by calling the contact center, for example. With the exception of email exchanges, these conversations are usually synchronous, far more focused, and issue-oriented.
2. A customer and company interaction that is indirect. Usually, this is accomplished by broadcasting corporate communications in public or by using user-generated material in public spaces. This is frequently an asynchronous exchange with a goal beyond resolving a specific problem. These kinds of contacts might be started by the business or by the client. Usually, social media, blogs, forums, and company websites are the channels employed.

3. Interaction between customers. These kinds of talks are recorded for public consumption in chat rooms, discussion boards, and other Web 2.0 collaborative platforms. The goal is to debate the goods and services that businesses offer by exchanging best practices and experiences. [12]

Conclusion:

Since it is still in its infancy, data mining is a relatively new technique. Even so, it's already being used often by a variety of businesses. Hospitals, banks, insurance companies, and retail establishments are a few of these entities. A lot of these businesses also use other critical techniques like statistics, pattern recognition, and data mining. Finding patterns and relationships that would be challenging to find without data mining is possible. Nevertheless, data mining is still an active research field with open problems and fresh, creative ideas, even with the recent advances in scalability.

References:

1. Zliobaite, A. Bifet, Mohamed Gaber, B. Gabrys, J. Gama, L. Minku, and K. Musial. Next challenges for adaptive learning systems. SIGKDD Explorations Newsletter, 14(1), 2012.
2. Brodley, C., Smyth, P.: Applying Classification Algorithm in Practice. In: Proceedings of the workshop on Applying Machine Learning in Practice at the ICML 1995 (1995)
3. Cendrowska, J.: Prism: An algorithm for inducing modular rules. International Journal of Man-Machines Studies 27, 349–370 (1987)
4. Parthasarathy, S., Dwarkadas, S.: Shared State for Distributed interactive Data Mining Applications. International Journal on Distributed and Parallel Databases (2002).
5. Jensen Z., Kwon S., Schwalbe-Koda D., Gómez-Bombarelli R, Román-Leshkov Y. A. M. Discovering relationships between OSDAs and zeolites through data mining and generative neural networks. ACS Central Science. 2021;7(5):858–867. doi: 10.1021/acscentsci.1c00024.
6. Pal, S.K., Bandyopadhyay, S.: Evolutionary computation in bioinformatics: A review. IEEE Trans. Syst. Man Cybern. Part B Cybern. 36, 601–615 (2006)
7. Provost, F.J., Kolluri, V.: A survey of methods for scaling up inductive learning algorithms. Data Min. Knowl. Discov. 2, 131– 169 (1999).
8. Xin R., Zhang J., Shao Y. Complex network classification with convolutional neural network. Tsinghua Science and Technology. 2020;25(4):447–457. doi: 10.26599/tst.2019.9010055.
9. P. Atzeni, Conceptual modeling-er 2004: Er 2004: 23rd international conference on conceptual modeling, shanghai, China, November 8-12, 2004: Proceedings, Springer, 2004.
10. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
11. Blakemore, D. (2002). Meaning and relevance: The semantics and pragmatics of discourse markers. Cambridge University Press.
12. Kim J., Kim H.-J., Kim H. Fraud detection for job placement using hierarchical clusters-based deep neural networks. Applied Intelligence. 2019;49(8):2842–2861. Doi: 10.1007/s10489-019-01419-2.