Luminis Applied Science and Engineering
ISSN: 3030-1831
Vol 2, No 3
2025

Mathematical and Statistical Methods in the Application Fields of Data Mining Technology

Mesume Seyidova
[1] Seyidova, M. Nakhchivan State University, Azerbaijan. Email: mensumeseyidova@ndu.edu.az. ORCID: https://orcid.org/0009-0008-7616-7268
https://doi.org/10.69760/lumin.2025003006

Abstract; In today's digital age, enormous volumes of data are generated every moment. Data mining leverages mathematical and statistical methods—alongside machine learning, database management, and data visualization—to extract valuable knowledge from these large data sets. Its techniques are applied successfully in diverse areas, including business, government, healthcare, science, and sports, with use cases such as database marketing, fraud detection, retail analytics, credit scoring, astronomy, and molecular biology. Mathematics provides the backbone of these processes through statistics, optimization, linear algebra, probability theory, and other fields. Together, mathematical and statistical methods enable efficient preprocessing, accurate modeling, reliable forecasting, and optimization. Their integration makes large-scale data easier to process and analyze, ultimately supporting informed decision-making across multiple domains.

***Keywords**: data mining; statistics; linear algebra; optimization; machine learning*

Application Areas of Data Mining

Data mining is applied across a wide range of industries. Its methods have achieved significant success in **healthcare, finance, retail, telecommunications, and marketing**. In fact, data mining technology can be implemented in all areas of human activity where retrospective data (researchers' reliance on past records or previously collected information) exists (Савченко & Бежитский, 2015). The main fields include:

- **Business problem-solving:** banking, finance, insurance, manufacturing, telecommunications, e-commerce, marketing, stock markets, and more.
- **Public sector:** detecting tax evasion, developing counter-terrorism tools, and other state-level applications.
- **Scientific research:** medicine, biology, molecular genetics and genetic engineering, bioinformatics, astronomy, and other sciences.
- **Web applications:** information-retrieval systems, web analytics, and traffic counters.

Through mathematics, data mining enables the analysis of information, the construction of models, and the discovery of hidden patterns. There is a close relationship between mathematics and data mining, as the field frequently combines **statistics, geometry, algebra, optimization, and probability theory**. Several core examples illustrate how mathematics is applied in data mining (Гаврилова & Хорошевский, 2001).

The Role of Mathematics in Data Mining

Mathematics is a widely used tool in **data mining**, applied across a variety of tasks ranging from data analysis to forecasting and optimization. Mathematical models and methods help data mining achieve more efficient and accurate outcomes. Their application facilitates the processing and analysis of large-scale data and provides essential information for making sound decisions.

In many cases, specialists need to understand the basics of mathematics to apply methods correctly. For other purposes, ready-made APIs or algorithms may be sufficient, but deeper insights require mathematical knowledge.

The **first stage of data mining** is **data preparation (preprocessing)**. At this stage, mathematical and statistical techniques are used to detect inexact, missing, or erroneous data. For example, faulty data can be identified and corrected through statistical or algebraic methods. Different mathematical approaches are applied to transform raw inputs into a usable format. In fact, data analysis and statistical modeling form the very foundation of data science, covering a wide spectrum of mathematical techniques and methods.

Core Methods Used in Data Mining

1. Mathematical methods:

- Linear algebra (vectors, matrices)
- Optimization methods
- Graph theory
- Discrete mathematics

2. Statistical methods:

- Regression analysis (linear and nonlinear)
- Correlation and covariance
- Hypothesis testing
- Probability distributions

Main functions of data mining include:

- **Classification:** assigning objects to defined categories.
- **Clustering:** grouping similar objects together.
- **Association rules:** uncovering hidden relationships among data.
- **Prediction/forecasting:** estimating future outcomes.
- **Anomaly detection:** identifying unusual or rare events.

Key Mathematical Foundations

Probability theory and statistics form the core of data analysis. Probability theory is essential for reasoning under uncertainty, such as in Bayes' theorem, which provides a framework for updating probabilities as new evidence becomes available.

Mathematical statistics covers the systematization, processing, and interpretation of data, and provides both scientific and practical conclusions. Here, "statistical data" refers to a comprehensive set of indicators describing an object. The foundation of mathematical statistics rests upon probability theory.

Meanwhile, statistics as a discipline is concerned with the collection, organization, analysis, interpretation, and presentation of data. This dual role—probability as the theoretical backbone and statistics as the applied practice—makes them indispensable for effective data mining.

Core Mathematical and Statistical Methods in Data Mining

Descriptive Statistics

Descriptive statistics include measures such as the mean, median, mode, standard deviation, variance, and range. These tools are used to summarize and understand data at a glance, providing a snapshot of key trends and variability.

Inferential Statistics

Inferential statistics involve drawing conclusions about a population from a sample. This is typically based on random sampling techniques. Common methods include hypothesis testing, chi-square tests, t-tests, and analysis of variance (ANOVA). These techniques allow researchers to generalize findings from limited data to larger populations with a defined level of confidence.

Linear Algebra

Linear algebra is the branch of mathematics dealing with vector spaces and linear mappings between them. It is a fundamental area widely applied in data analysis. Linear algebra focuses on matrices and vectors, solving linear equations, and studying linear functions. In simple terms, it helps to understand multidimensional geometric concepts and to perform calculations on them. Concepts from linear algebra are used in regression analysis, principal component analysis (PCA) for dimensionality reduction, and as the computational backbone of many machine learning (ML) models (Kharkovyna, 2019).

Calculus

Calculus, the study of continuous change, is applied in optimization tasks within data analysis and especially in machine learning. For instance, gradient descent is an optimization algorithm that iteratively moves in the direction of the steepest decrease in order to minimize a function. This relies on derivatives to determine the slope and direction of change, thereby helping to locate minima or maxima of loss functions.

Optimization Methods

Optimization techniques aim to find the best (or optimal) solution when constraints exist. Optimization problems arise naturally in data analysis and ML. For example, training a model often involves optimizing a loss function to improve predictive performance.

Graph Theory

Graph theory, a branch of discrete mathematics, studies the properties of graphs (nodes and edges). It is particularly useful in analyzing social networks, ranking web pages, and understanding relationships within large datasets. In data mining, graph-based methods help with network analysis and certain clustering algorithms.

Differential Equations

Differential equations describe the relationship between a function and its derivatives. Because such relationships are common across disciplines, differential equations play a central role in engineering, physics, economics, and biology. In data mining, they are used in **time-series analysis**, where data are modeled as functions of time. They also underpin processes such as **training neural networks**, where iterative updates can be represented as differential systems.

Matrix Operations and Decomposition

Matrix operations (multiplication, transposition, inversion) and decomposition techniques (e.g., Singular Value Decomposition, SVD) are fundamental in many data-analysis contexts. Machine learning projects often deal with complex objects such as audio, video, and images, where classifiers rely on linear algebraic techniques to extract features and reduce errors. Matrix decomposition serves as the computational engine behind the efficient processing of large, high-dimensional data sets. These methods provide the principles and tools necessary for converting raw data into useful, actionable knowledge.

Databases and Machine Learning in Data Mining

Data mining relies on two fundamental components: databases and machine learning (Brownlee, 2019a). The database component provides methods for storing, managing, and retrieving data, while the machine-learning component supplies methods for analyzing and interpreting that data.

It is important to note that data mining by itself does not "learn" independently. It follows predefined rules and algorithms to solve specific problems. In contrast, machine-learning algorithms can adapt, change their rules based on circumstances, and discover new solutions in flexible ways.

Types of Machine Learning

In the context of data mining, three main types of machine learning are commonly applied (Brownlee, 2019a):

- **Supervised learning.** In this approach, the algorithm is trained on input–output pairs that have been pre-labeled by humans. The system is provided with examples of the desired outcome, enabling it to map inputs to outputs and generalize to new data (Brownlee, 2019b).
- **Unsupervised learning.** Here, the input data are unlabeled. The algorithm searches for commonalities, patterns, and features within the data without predefined categories. Because unlabeled data are far more abundant than labeled data, this method is especially valuable (Mishra, 2017).
- **Reinforcement learning.** In this setting, the system learns by interacting with its environment, receiving rewards or penalties based on its actions, and gradually improving its performance.

Linear Algebra in Machine Learning

Linear algebra is one of the core mathematical tools in machine learning. Nearly all ML models are built on operations involving vectors and matrices. These concepts form the computational language of algorithms, making it possible to represent, manipulate, and optimize large and complex datasets.

The next section explores practical applications of linear algebra in ML, including data representation, linear models, neural networks, dimensionality reduction, similarity measures, and optimization.

Linear Algebra Applications in Machine Learning

1. **Data Representation**

   Observations (*samples*) and their properties (*features*) are often represented as **vectors** or **matrices**.
   Example: Exam scores of three students across three subjects:

$$x = \begin{bmatrix} 85 & 90 & 78 \\ 70 & 65 & 80 \\ 95 & 88 & 92 \end{bmatrix}$$

Here, each **row** corresponds to a sample (a student), and each **column** corresponds to a feature (a subject).

2. **Linear Models**

   Linear regression can be expressed as:

$$y = X\omega + b$$

   where \(X\) is the data matrix, \(ω\) is the weight vector, \(b\) is the bias term, and \(y\) is the output. Predictions are calculated simply through the product of a matrix and a vector.

3. **Neural Networks**

   In each neural-network layer, the output is computed as:

$$z = Wx + b$$

   where \(W\) is the weight matrix, \(x\) the input vector, and \(b\) the bias. These are linear algebra operations, followed by a nonlinear activation function.

4. **Dimensionality Reduction**

   - **Principal Component Analysis (PCA):** identifies eigenvalues and eigenvectors to reduce dimensionality while retaining maximum variance.
   - **Singular Value Decomposition (SVD):** projects data onto its most significant directions.

5. **Similarity and Distances**

   Linear algebra enables computation of distances and similarities:

   - **Euclidean distance:**

$$d(A, B) = \sqrt{\sum (a_i - b_i)^2}$$

   - **Cosine similarity:**

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|}$$

These are widely applied in **natural language processing (NLP)** and **recommender systems**.

6. **Optimization (Gradient Descent)**

Training machine-learning models involves minimizing a loss function. Gradients, often represented as vectors or Jacobian matrices, are computed and used to iteratively update weights toward an optimal solution.

Conclusion

Linear algebra is the "language" of machine learning. Vectors, matrices, eigenvalues, orthogonality, and norms all stand at the foundation of ML. As a field, linear algebra—concerned with vectors, matrices, and linear transformations—is used not only in physics and engineering but also in computer science. Its applications in machine-learning algorithms include data analysis, prediction, classification, visualization, and regression (Mishra, 2017).

Thus, data mining combines mathematical methods and statistical analysis to evaluate and interpret information. These tools allow practitioners to understand customer behavior, make forecasts, discover relationships, and detect anomalies. From preprocessing to modeling and optimization, mathematics—particularly linear algebra—helps data mining achieve more accurate, efficient, and scalable results for decision-making.

References

Brownlee, J. (2019a, August 12). *Supervised and unsupervised machine learning algorithms*. Machine Learning Mastery. https://machinelearningmastery.com/linear-algebra-machine-learning-7-day-mini-course/

Brownlee, J. (2019b, December 5). *A tour of machine learning algorithms*. Academia.edu. https://www.academia.edu/44070982/IRJET_Application_of_Linear_Algebra_in_Machine_Learning

Hendrastuty, N. (2024). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa. *Jurnal Ilmiah Informatika dan Ilmu Komputer (JIMA-ILKOM)*, *3*(1), 46-56.

Javaid, H. A. (2024). Improving fraud detection and risk assessment in financial service using predictive analytics and data mining. *Integrated Journal of Science and Technology*, *1*(3).

Kharkovyna, O. (2019, October 12). *Mathematics for AI: Linear algebra and how to understand it better*.

Liu, H., Li, Y., Karsidag, M., Tu, T., & Wang, P. (2025). Technical and biological biases in bulk transcriptomic data mining for cancer research. *Journal of Cancer*, *16*(1), 34.

Mining, W. I. D. (2006). Introduction to data mining. *Mining Multimedia Databases, Mining Time Series and*.

Mishra, S. (2017). *Unsupervised learning and data clustering* [Master's thesis, Lund University Publications]. https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9066702&fileOId=9066703

Tsiu, S. V., Ngobeni, M., Mathabela, L., & Thango, B. (2025). Applications and competitive advantages of data mining and business intelligence in SMEs performance: A systematic review. *Businesses*, *5*(2), 22.

Zou, Z., Ohta, T., & Oki, S. (2024). ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. *Nucleic Acids Research*, *52*(W1), W45-W53.

Гаврилова, Т. А., & Хорошевский, В. Ф. (2001). *Какая-то интеллектуальная система* [An intelligent system]. Санкт-Петербург: Питер.

Савченко, Л. М., & Бежитский, С. С. (2015). Data mining и области его применения [Data mining and its application areas]. *Компьютерные и информационные науки*, 1. https://cyberleninka.ru/article/n/datamining-i-oblasti-ego-primeneniya

Received: 07.25.2025
Revised: 07.27.2025
Accepted: 08.10.2025
Published: 09.18.2025