Study of Data Mining Techniques

Swapnali Bhuite

Aishwarya Suryawanshi

Dept. of computer application, University of Mumbai, Mumbai – 400068, India

Abstract:

Data mining is one of the fastest growing areas of computer science, using multiple mining algorithms to discover hidden knowledge and patterns. Data mining is the process of finding useful patterns in large amounts of data. Various areas of DM include stock analysis, economics, image processing, data mining in education, and weather forecasting. Data mining in education is one of the well-known areas of application for discovering using interesting and hidden knowledge from academic data. Various data mining-technologies. Predictive and descriptive mining techniques apply to extracted academic knowledge that supports educational organizations.

Keyword:

Data mining tasks and methods, Knowledge discovery database data mining, techniques Data mining tools, Areas where data mining has good / bad effects.

Introduction

Overview:

Today, the main purpose of academic institutions is to provide students with specialized knowledge and improve their academic practice. This can be done by observing hidden knowledge from academic activities that directly affect student performance in the learning environment. The observed information can be used to make better decisions, enhance learning activities, and better understand student behavior. Data mining is a way to investigate the facts. Discover knowledge using a variety of supervised and unsupervised techniques, explore hidden knowledge and patterns, and discover unexpected knowledge and the relationships between them. There are various data mining techniques such as classification. Used by researchers to examine sequential pattern mining, graph mining, association rule mining, and data. Educational data mining is becoming a new field for researchers. This is a subset of data mining and is an evolving area where learning institutions and data mining techniques can be used to make meaningful predictions as described by S. Ventura and C. Remeo.

Main areas of EDM:

The most relevant area of EDM is learning analytics, also known as learning analytics. This can be described as analysing, collecting, reporting, and measuring datasets to determine the learning environment. According to Baker and Yasef (2017), there are four main goals of educational data mining in education.

1. Examining the learning performance of educators is the first goal of EDM that can be achieved through student modelling.

2. Improving the domain model is EDM's second goal, and various EDM techniques can be used to predict new knowledge in the area of interest.

3. Given the impact of academic support, this goal can be achieved through a learning management system.

4. EDM's ultimate goal is to improve the knowledge of responsible educators, learners and scholars, which can be achieved through the construction and integration of student models.

To achieve the above goals, researchers used data mining techniques that were basically split into two models. In other words, there are two methods, a prediction method and a description method. Predictive approaches use current training sets to predict future outcomes, and these predictions are probability-based. These do not provide accurate results, they only show what might happen, for example, a company's sales or sales report, as explained in Data Analysis (2019).

Problem Statement:

Recently, academic institutions collect and organize vast amounts of data, such as student attendance records and exam results records. Student demographic information, academic information, student registration information, etc. Extracting such information yields meaningful knowledge that helps academic leaders make better decisions. Educational data is constantly evolving and needs to be transformed into useful information, requiring advanced mining techniques. There are various methods available on the market that can be applied to the classification of educational datasets. This study runs five classification algorithms on educational datasets to assess whether more accurate results can be obtained.

Motivation:

We live in a society where technology is an important tool for public development. Recent developments have brought about many changes in various areas of society. Advances in machine learning, artificial intelligence, and educational data mining have led researchers to develop appropriate mock-up’s and approaches to predict student performance and discover properties that improve performance.

Aim and Objectives:

Educational data in academic databases has grown rapidly in recent years, and dealing with this raw data is a daunting task and needs to be transformed into meaningful knowledge. Most researchers work in academia and do a lot of research in education. The purpose of this study is to provide a large and comprehensive study of data mining in education and to compare different mining algorithms. The main objectives of this study are:

- Identify the techniques used in connection with data mining in education.
- Comparison of EDM technology and its strengths and weaknesses.
- Identify the tools used for EDM.
- Identify the main application areas of EDM.
- Identify the current challenges facing your research.
- Compare classification algorithms.

Methodology:

The purpose of EDM is to transform the raw data that comes from educational databases into valid and meaning full knowledge as describe by L.A.

- A comprehensive study of EDM
- Applying data mining techniques on educational data set.

This kind of literature review where we identify educational data mining phases, techniques , tools ,main modules of EDM , application areas and major challenges , whereas the second part of research apply data mining technique classification on the data set and compare the evaluated results to show that which particular techniques provides more accuracy.

Comprehensive Study Of Educational Data Mining +

Applying DM techniques for Predicting Students Academic Performance =

A Brief Analysis on EDM

Tasks and Methods of data mining:

According to R. Sathya and A. Abraham (2013) , data mining tasks are divided into two different categories that are :

- Supervised /Directed data mining
- Un Supervise /Un Directed data mining

Supervised data mining approach characterizes instance according to the present facts. Targeted data is nominated and directs the program to create the model that will present the instance according to the nominated data as mentioned by C. Donalek.

The objective of unsupervised / un-directed data mining techniques is to make associations among distributed facts available in the training set. Here operator may request the program to the detect important connection that may occur between instances as mentioned by C. Donalek.

Data mining methods could be categorized into two major models or we can say into two main methods as mentioned by Nikita and Vishal.

- Predictive mining Method
- Descriptive mining Method

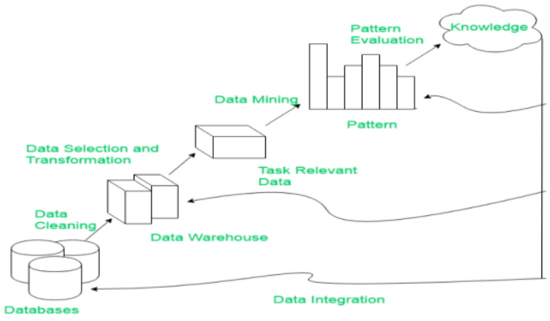
Predictive data mining techniques can be defined as what will happens in the future by analysing the historical data , it is a process of predicting the future of an organization example predicting reports for a company as described in the data analytics tutorial. The process involved here is predicting techniques and statistics. It allows a proactive approach.

Descriptive data mining techniques can be defined as what happened in the past , it is procedure of discovering valid information by analysing large databases and this technique is only limited to the historical data. The proceed involved here is data mining and data aggregation and it allows reactive approach as described in the data analytics tutorial.

Knowledge Discovery Database:

The term KDD stands for Knowledge Discovery in Databases. It refers to the broad procedure of discovering knowledge in data and emphasizes the high-level applications of specific Data Mining techniques. It is a field of interest to researchers in various fields, including artificial intelligence, machine learning, pattern recognition, databases, statistics, knowledge acquisition for expert systems, and data visualization.

The main objective of the KDD process is to extract information from data in the context of large databases. It does this by using Data Mining algorithms to identify what is deemed knowledge.



. 1. Data Cleaning: Data cleansing is described as elimination of noisy and inappropriate information from collection.

• Cleaning in case of Missing values.

• Cleaning noisy information, wherein noise is a random or variance error.

• Cleaning with Data discrepancy detection and Data transformation tools.

2. Data Integration: Data integration is described as heterogeneous information from a couple of reasserts blended in a not unusual place source (Data Warehouse).

• Data integration the use of Data Migration tools. • Data integration the use of Data Synchronization tools.

• Data integration the use of ETL (Extract-Load-Transformation) system.

3. Data Selection: Data choice is described because the system wherein information applicable to the evaluation is determined and retrieved from the information collection.

• Data selection using neural network.

• Data selection using Decision Trees.

• Data selection using Clustering, Regression, etc.

4. Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

Data Transformation is a two-step process:

• Data Mapping: Assigning elements from source base to destination to capture transformations.

• Code generation: Creation of the actual transformation program.

Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.

• Transforms task relevant data into patterns.

• Decides purpose of model using classification or characterization.

5. Pattern Evaluation: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.

• Find interestingness score of each pattern.

• Uses summarization and Visualization to make data understandable by user.

6. Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

• Generate reports.

• Generate tables.

• Generate discriminant rules, classification rules, characterization rules, etc.

Data Mining Techniques:

Data mining techniques are mainly categorized into major categories:

1) Association
2) Classification
3) Prediction
4) Clustering
5) Regression
6) Artificial Neural network (ANN) Classifier Method
7) Outlier Detection

. **1. Association:**

Association evaluation is the locating of affiliation regulations displaying attribute-fee situations that arise regularly collectively in a given set of records. Association evaluation is extensively used for a marketplace basket or transaction records evaluation.

2. Classification:

Classification is the processing of locating a fixed of fashions that describe and distinguish records lessons or concepts, for the cause of being capable of use the version to are expecting the elegance of gadgets whose elegance label is unknown. Data Mining has a one of a kind of classifier:

• Decision Tree
• SVM (Support Vector Machine)
• Generalized Linear Models
• Bayesian type
• K-NN Classifier
• Rule-Based Classification
• Frequent-Pattern Based Classification • Rough set theory
• Fuzzy Logic

3. Prediction:

Predictive analytics makes use of ancient records to be expecting destiny events. Typically, ancient records are used to construct a mathematical version that captures essential trends. That predictive version is then used on cutting-edge records to be expecting what is going to take place next, or to signify movements to take for surest outcomes.

4. Clustering:

Unlike type and prediction, which examine elegance-labelled records gadgets or attributes, clustering analyses records gadgets without consulting a diagnosed elegance label. In general, the elegance labels do now no longer exist with inside the education records surely due to the fact they're now no longer regarded to start with. Clustering may be used to generate those labels. The gadgets are clustered primarily based totally at the precept of maximizing the intra-elegance similarity and minimizing the interclass similarity

5. Regression:

Regression may be described as a statistical modelling approach wherein formerly received statistics is used to predicting a non-stop amount for brand new observations. This classifier is likewise called the Continuous Value Classifier. There are sorts of regression models: Linear regression and a couple of linear regression models.

6. Artificial Neural network (ANN) Classifier Method:

A synthetic neural network (ANN) additionally known as truly a “Neural Network” (NN), will be a system version supported with the aid of using organic neural networks.

7. Outlier Detection:

A database might also additionally include records items that don't follow the overall behaviour or version of the records. These records items are Outliers. The research of OUTLIER records is referred to as OUTLIER MINING. An outlier can be detected the usage of statistical tests.

Tools of Data Mining:

1. **WEKA** is one of the maximum used ML devices written in java programming language brought in New Zealand through Waikato University. Weka is a set of system gaining knowledge of algorithms for information mining tasks. The algorithms can both be carried out without delay to a dataset or referred to as out of your personal Java code. Weka incorporates gear for information pre-processing, classification, regression, clustering, affiliation rules, and visualization.
2. **Rapid Miner:** Rapid Miner is one of the first-rate predictive evaluation machine evolved through the organization with the equal call because the Rapid Miner. It is written in JAVA programming language. It affords an included surroundings for deep gaining knowledge of, textual content mining, system gaining knowledge of & predictive evaluation.
3. **Orange:** Orange is a super software program suite for system gaining knowledge of & information mining. It first-rate aids the information visualization and is part primarily based totally software program. It has been written in Python computing language. As its miles component-primarily based totally software program, the additives of orange are referred to as ‘widgets’. These widgets variety from information visualization & pre-processing to an assessment of algorithms and predictive modelling.

Oracle Data Mining:

A issue of Oracle Advance Analytics, Oracle information mining software program presents great information mining algorithms for information classification, prediction, regression and specialised analytics that allows analysts to examine insights, make higher predictions, goal great customers, become aware of cross-promoting opportunities & locate fraud.

AREAS WHERE DATA MINING HAS GOOD/BAD EFFECTS:

Areas, where data mining has good effects, are:

- Future Predicting trend
- Decision Making
- Fraud Detection
- Improve Organization Revenue
- Weather Forecasting
- Stock Trade Analysis

Areas, where data mining has bad effects, are:

- Organization security/use privacy
- Required high implementation cost
- Possibility of misuse of information
- Possibility of misuse data

Conclusion:

Data mining has significance concerning locating the styles, forecasting, discovery of understanding etc., in extraordinary commercial enterprise domains. Data mining strategies and algorithms including classification, clustering etc., facilitates in locating the styles to determine upon the destiny traits in agencies to grow. Data mining has huge utility area nearly in each enterprise in which the facts is generated that's why facts mining is taken into consideration one of the maximum crucial frontiers in database and statistics structures and one of the maximum promising interdisciplinary trends in Information Technology.

Reference:

- C. Zuo, "Defence of computer network viruses based on data mining technology," International Journal on Network Security,vol. 20, no. 4, pp. 805–810, 2018.View at: Publisher Site | Google Scholar

- Data Mining Concepts and Techniques, published by Morgan Kauffman.