Int. j. inf. tecnol. (December 2020) 12(4):1243–1257
https://doi.org/10.1007/s41870-020-00427-7

ORIGINAL RESEARCH

A comprehensive survey of data mining

Manoj Kumar Gupta[1] · Pravin Chandra[1]

Received: 29 June 2019 / Accepted: 20 January 2020 / Published online: 6 February 2020

Abstract Data mining plays an important role in various human activities because it extracts the unknown useful patterns (or knowledge). Due to its capabilities, data mining become an essential task in large number of application domains such as banking, retail, medical, insurance, bioinformatics, etc. To take a holistic view of the research trends in the area of data mining, a comprehensive survey is presented in this paper. This paper presents a systematic and comprehensive survey of various data mining tasks and techniques. Further, various real-life applications of data mining are presented in this paper. The challenges and issues in area of data mining research are also presented in this paper.

Keywords Data mining techniques · Data mining tasks · Data mining applications · Clustering · Classification · Survey

1 Introduction

Data mining, an essential and important step in knowledge discovery in databases, is used to discover useful unknown patterns from large repository of data [1–4]. Data mining consists of various functionalities, techniques and algorithms that are being used to discover and extract interesting patterns from the large repository of data [1, 2, 4]. Due to the importance in decision making, in the last two decades, data mining got a wide focus and has become an essential tool in performing variety of operations of the organizations [5].

> Data mining is a step in the knowledge discovery in databases process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data….. [1].

Han et al. [6] stated data mining as "data mining is a process of discovering or extracting interesting patterns, associations, changes, anomalies and significant structures from large amounts of data which is stored in multiple data sources such as file systems, databases, data warehouses or other information repositories."

Many techniques from other domains [6–8] such as statistics, database/data warehouse systems, machine learning, algorithms, pattern recognition, visualization, information retrieval, high-performance computing, etc. incorporated in data mining. First three techniques are the primary contributors of data mining [7].

2 Trends in data mining research

Through a survey of literature, it is identified that the data mining research can be broadly categorized into following types [9–12].

2.1 Data mining functions

Data mining functions or tasks can be used to specify the types of patterns or knowledge to be discovered during the data mining process. Some of the major data mining

✉ Manoj Kumar Gupta
manojkgupta5@gmail.com; manojgupta5@yahoo.com

[1] University School of Information, Communication and Technology, Guru Gobind Singh Indraprastha University, Sector-16C, Dwarka, Delhi 110078, India

functions are summarization, characterization and discrimination, association, clustering, classification, outlier analysis, regression and trend analysis, etc. [1, 2, 6, 13].

2.2 Data mining techniques

Data mining task(s) are performed based on number of data mining techniques or approaches. A wide range of techniques for data mining are investigated by the researchers so far. For example, machine learning, statistics, neural networks, database and data warehouse systems, genetic algorithms, fuzzy sets, visualization, etc. [6, 9, 13].

2.3 Data mining algorithms

A variety of algorithms, also known as methods, are proposed by many researchers to carry out data mining functions based on data mining techniques. For example, Apriori algorithm, Naïve Bayesian, k-Nearest Neighbour, k-Means, CLIQUE, STING, etc. [6, 14].

2.4 Data mining domains

Data mining can be used in set of domains. E.g. time-series data mining, web mining, temporal data mining, spatial data mining, tempo-spatial data mining, educational data mining, business, medical, science and engineering, etc. Each domain can have one or more applications of data mining [6, 15].

2.5 Data mining applications

It is a set of application areas where the one or more data mining function can be used. For example, financial data analysis, market-basket analysis, intrusion detection, fraud detection, recommender systems, cancer detection, etc. [9, 6, 15].

Out of the above mentioned categories of data mining research, only data mining tasks, techniques and real-life applications are surveyed and presented in this paper.

3 Data mining tasks (or functions)

A large database and/or data warehouse may have a variety of unknown patterns in it [16]. To extract this variety of unknown patterns, distinct type of data mining function, methods and techniques can be used [11, 12, 17]. Based on different types of patterns, data mining functions can be categorized into summarization, characterization and discrimination, classification, regression and trend analysis, clustering, outlier analysis, association, etc. [1, 2, 9, 13, 18]. Classified work of the aforesaid literature related to data mining tasks is listed in Table 1.

3.1 Summarization

Summarization results into a smaller set and presents a summary of the detailed data based on concept hierarchy. Usually, summarization is performed using aggregation which can be extended to different levels of abstraction and can be viewed from diverse angles. Various kinds of patterns can be extracted based on combinations of various levels of abstractions and different dimensions [13]. Data summarization is usually accomplished using attributed-oriented induction approach [69] and data cube approach [36, 70].

Data cube approach (also referred to as 'multidimensional databases', 'materialized views') materializes frequently queried expensive computations which involve group functions and then store the result as materialized views in a MDDB for decision support and knowledge discovery [13]. The attribute-oriented induction approach collects the related data in a database with the help of SQL-like DMQL and then a set of data generalization techniques [69] are applied for data generalization [13].

3.2 Characterization and discrimination

Characterization is basically summarization of data based concept hierarchy and generates characterization rules. On the other side, discrimination is used for identifying the varieties among various data sets. The output of the discrimination is generated in the form of discriminant rules [6, 23, 31].

3.3 Classification

Classification is the process to classify new observation based on the predetermined classes, i.e. supervised learning. A classification algorithm is used to forecast classes of the data [6]. A large collection of classification algorithms (or classifiers) have been proposed by the researchers [6, 47] so far. Some popular classification algorithms are summarized in Table 2. The classifiers based on genetic algorithms, rough set approach, fuzzy sets, semi-supervised learning and active learning have been also proposed by some researchers [6].

In addition to the popular classifiers mentioned in Table 2, many researchers have also presented and/or discussed a set of new classifiers such as classifier based on predictor features using supervised learning [43], a property-based classification [61] to adapt symbolic values, a unified classification model framework [50] for

Table 1 Classified work of reported literature related to data mining tasks

Reference	Data mining tasks						
	Summarization	Characterization and discrimination	Classification	Clustering	Association	Outlier analysis	Regression and trend analysis
Abuaiadah [19]				√			
Algergawy et al. [20]				√			
Angiulli et al. [21]						√	
Angiulli and Fassetti [22]						√	
Bhatnagar et al. [23]		√		√		√	
Bouguessa [24]				√		√	
Campello et al. [25]				√		√	
Carpineto et al. [26]				√			
Ceglar and Roddick [27]					√		
Chen et al. [13]	√	√	√	√	√		
Chen et al. [28]					√		
Chin-Yuan et al. [29]				√			
Das et al. [30]				√			
Dincer [31]		√		√			
Geng and Hamilton [32]	√		√		√		
Gupta and Chandra [33]				√			
Gupta and Chandra [34]				√			
HeaZ et al. [35]			√			√	
Hung et al. [36]					√		
Hung and Thu [37]					√		
Jain et al. [38]				√			
Jin et al. [39]				√			
Khandare and Alvi [40]				√			
Koh and Ravana [41]					√		
Kosina and Gama [42]			√				
Kotsiantis [43]			√				
Kumar et al. [44]			√			√	
Lee and Yun [45]					√		
Li and Zaki [46]			√				
Liao and Triantaphyllou [47]			√				
Mabroukeh and Ezeife [48]					√		
Mampaey and Vreeken [49]				√			
Menardi and Torelli [50]			√				
Mukhopadhyay et al. [51]				√			
Pei et al. [52]				√			

Table 1 continued

Reference	Data mining tasks						
	Summarization	Characterization and discrimination	Classification	Clustering	Association	Outlier analysis	Regression and trend analysis
Rafalak et al. [53]			√				
Reddy and Jana [54]				√			
Rustogi et al. [55]					√		
Shah-Hosseini [56]				√			
Silva et al. [57]				√			
Silva and Antunes [58]					√		
Sim et al. [59]				√			
Sohrabi and Roshani [60]					√		
Susan et al. [61]			√				
Tan et al. [62]				√	√		√
Tew et al. [63]				√	√		
Wang and Dong [64]				√			
Wang and Sun [65]			√	√			
Wang et al. [66]				√			
Zacharis [67]			√				
Zhang et al. [68]				√			

classification of skewed distribution of observations, adaptive very fast decision rules (AVFDR) [42], etc.

3.4 Clustering (or cluster analysis)

Clustering is used for partitioning or segmenting data objects (or observations) into subsets called as groups or clusters. The objects that are closed to each other are positioned in same group. Like classification, clustering classifies the similar data objects but unlike classification, the class labels are unknown (i.e. unsupervised learning) [6]. Cluster analysis is one of the most popular techniques which is not only used in data mining but also used in other domains such as statistics, image segmentation, pattern recognition, object recognition, information retrieval, bioinformatics, etc. [38].

A large collection of clustering algorithms has been suggested by many researchers [29, 6, 62] in the last two decades. Some popular clustering algorithms are presented in Table 3. The clustering algorithms based on probabilistic model, fuzzy sets, expectation–maximization, correlation using PCA, graph have been also proposed by some researchers [6].

In addition to the popular clustering algorithms presented in Table 3, many researchers have also presented and/or discussed a set of new clustering algorithms such as parameter-free method using minimum description length [49], parallelized hierarchical clustering approach [66], gene expression data clustering approach based on z-score measure [30], fully automatic clustering algorithm for high dimensional categorical data [24], nature inspired swarm based Intelligent Water Drops—K-Means (IWD-KM) algorithm [56], Voronoi diagram based clustering algorithm [71] for artificial as well as biological data, bisect K-means clustering algorithm [19], domain knowledge based density-based clustering [39], algorithm for clustering large-scale data sets based on the unique combination of matrix decomposition and low-rank matrix approximation named as exemplar-based low-rank sparse matrix decomposition (EMD) [64], a three-phased cluster ensemble method based on discriminant analysis [23], etc. Campello et al. [25] presented a framework for density-based clustering. Khandare and Alvi [40] proposed an improved clustering algorithm by proposing a new method of cluster initialization. Gupta and Chandra [72] proposed an efficient approach based on the selection of well-separated data points as intial cluster centroids to improve the performance of k-means algorithm. New cluster initialization approaches using partitioning for k-means algorithm, called as P-k-means and M-P-k-means, are proposed in Gupta and Chandra [33, 73] respectively. Hypercube based cluster initialization method, called as HYBCIM is proposed in Gupta and Chandra [74]. HYBCIM, P-k-means and M-P-k-means algorithms give better results as compared to traditional k-means algorithm.

Table 2 Some Popular Classification Algorithms [6]

Category	Name of algorithm	Based on the concept
Decision tree classifiers (machine learning based)	ID3 (iterative dichotomiser)	Information gain
	CART (classification and regression trees)	Gini index
	C4.5 (a descendant of ID3)	Gain ratio
Bayesian classifiers (statistics based)	Naïve Bayesian (or Simple Bayesian) classifier	Bayes' theorem
	Bayesian belief networks	Bayes' theorem and probabilistic graphical model
Rule based classifiers (machine learning based)	IF–THEN rule using decision tree	Decision tree
	Sequential covering algorithms: AQ, CN2 and RIPPER	Entropy, information gain
Support vector machine classifier	Support vector machine	Linear optimal separating hyperplane
Classification using Backpropagation	Back propagation	Multilayer FF ANN
(neural network based)		
Classification using frequent patterns (associative classification based)	CBA (classification based on association)	Frequent Itemset Mining
	CMAR (classification based on predictive association rules)	Frequent itemset mining with rule pruning strategy
	CPAR (classification based on multiple association rules)	Frequent itemset mining with foil (first order inductive learner)
Lazy learners (machine learning based)	kNN (K-nearest neighbour)	Learning by analogy and euclidean distance
	CBR (case-based reasoning)	Database of problem solutions and background knowledge

Table 3 Some popular clustering algorithms [6]

Category	Name of algorithm	Based on the concept
Hierarchical	DIANA (divisive analysis)	Divisive method
	AGNES (agglomerative nesting)	Agglomerative method
	Chameleon	Dynamic modeling
	BIRCH (balanced iterative reducing and clustering using hierarchies)	Clustering feature tree
	Probabilistic hierarchical clustering	Probabilistic Model
Partitioning-based	k-Means	Centroid
	k-Medoids	Representative Object
	CLARA (clustering large applications)	Sampling
	CLARANS (clustering large applications based upon randomized search)	Randomized sampling
	PAM (partitioning around medoids)	Representative object
Grid-based	CLIQUE (clustering in quest)	Identification of monotonicity of dense cells with respect to dimensionality
	STING (statistical information grid)	Grid Cells containing statistical information
Density-based	DBSCAN (density-based spatial clustering of application of noise)	Connected regions with high density
	OPTICS (ordering points to identify the clustering structure)	Connected regions with high density characterized by global density parameters
	DENCLUE (density-based clustering)	Density distribution function

Clustering XML data is one of the straightforward problems in many latest data mining applications such as Web Mining, XML query processing, Bioinformatics, etc. The conventional data clustering methods are not appropriate for XML data clustering [20]. The traditional clustering techniques are not suitable for web search results clustering because it has specific requirements [26]. Clustering data streams is also a difficult process as it requires the ability to continuously cluster the streaming objects within given memory and time restrictions [57].

3.5 Outlier analysis

Data objects that differ in general behaviour of the data are called as outliers. The outliers are generally discarded by most of data mining methods as noise or exceptions. Sometimes, outliers may have more information in comparison to other data objects. Therefore outlier analysis is important for some application areas such as intrusion detection, fraud detection, anomaly detection, etc. [35]. Many data mining techniques generally use clustering to detect the outliers as a noise. The outlier detection methods can be classified as classification-based methods, statistical methods, clustering-based methods, supervised, semi-supervised and unsupervised methods, deviation-based methods and proximity-based methods [6].

Angiulli and Fassetti [21] stated that the background knowledge (or domain knowledge) can be used to detect the outliers easily. They proposed the solution as unsupervised but it can have relationship with supervised learning. Gradient outlier factor is investigated in Angiulli and Fassetti [21] for generalization and unification of statistical outliers. Campello et al. [25] presented a framework for density-based outlier detection. Two new algorithms inc-iVAT and dec-iVAT based on visual assessment of tendency for anomaly detection in data steam are introduced in [44].

3.6 Association analysis (or association mining)

Association analysis discovers associations (or links) among datasets and identifies data objects that can be realized collectively satisfying a minimum support and confidence thresholds. Identification of all frequent item sets followed by generation of strong association rules is accomplished in association mining [28, 6]. Association analysis includes mining frequent itemsets, subsequences and substructures [6]. Market-basket analysis is mainly using the association analysis. Apriori algorithm is widely used for association. Association analysis algorithms can be classified into classical algorithms, condensed representation algorithms, and incomplete set algorithms [27].

Some popular association mining algorithms are summarized in Table 4. The association mining algorithms for multilevel association, multidimensional association, quantitative association, rare (or infrequent) patterns, constraint-based association, etc. have been also proposed by some researchers [6].

Multi-relational Data Mining (MRDM) is the process to look for multiple tables based patterns [58]. Sampling methods using disjunctive normal form (DNF) have been developed by Li and Zaki [46]. Rare or infrequent pattern mining is becoming popular nowadays in some application areas [41]. A correct and efficient algorithm for uncertain frequent patterns mining using minimum data structure is investigated by Lee and Yun [45]. A new less time consuming algorithm based on cellular learning automata (CLA) for mining frequent itemsets is presented in Sohrabi and Roshani [60].

Data mining can also extract uninteresting patterns/rules. Therefore, pattern evaluation (measuring interestingness) is required to filter out the only interesting patterns. Geng and Hamilton [32] presented nine specific criteria for measuring the interestingness of the mined rules and summaries. Further, these nine criteria have been categorized into three categories (1) subjective, (2) objective and (3) semantic-based. Tew et al. [63] suggested a technique to detect equivalences among interestingness measures using rule-ranking behaviour-based clustering for association rule mining. Hung et al. [36] presented a method WSWFP-stream based on FPGrowth method for mining frequent itemsets with weights for data stream. A new algorithm called MFIWDSIM based on weights using Inverted Matrix for mining frequent itemsets is proposed in [37]. Rustogi et al. [55] presented a improved parallel Apriori algorithm for multi-core.

3.7 Regression and trend analysis (or evolution analysis)

Regression predicts the value of attribute based on regression technique(s) over time. The future values of variables are predicted with the help of historical time series plot [6].

Trend analysis (also called as evolution analysis) discovers interesting patterns in the evolution history of the objects. Identification of patterns in an object's evolution and matching of the objects' changing trends are the two major aspects of trend analysis [28]. Trends of the objects, whose behaviour evolves over time, can be described using trend analysis and regression models. Trend analysis exposes time-varying trends of the data objects within the dataset. The association analysis can also be used for evolution analysis [62].

4 Data mining techniques

As data mining is a multi-disciplinary field, variety of techniques or approaches are adopted in data mining from number of domains which includes statistics, machine

Table 4 Some popular association algorithms [6]

Category	Name of algorithm	Based on the concept
Apriori-like	Apriori	Confined candidate generation
Frequent pattern growth-based	FP-growth	Conditional pattern base without candidate generation
Vertical data format-based	Eclat (equivalence class transformation)	Data transformation and candidate generation

learning, neural networks, database systems, genetic algorithms, fuzzy sets, visualization, etc. [28, 9, 6]. Classified work of the aforesaid literature related to data mining techniques is listed in Table 5.

4.1 Statistical approaches

Sometimes the terms ‘statistics’ or ‘statistical techniques’ are used as alias for data mining. But, statistics was coined before the term ‘data mining’. Statistics is data driven and is used to discover patterns and to build predictive (in statistics also called as regression) models. Due to its data driven approach, statistics is also used as one of the major technique for data mining [86]. In other words, data mining has an inherent connection with statistics [6]. Many statistical analysis tools including Bayesian network, correlation analysis, factor analysis, discriminant analysis, cluster analysis, regression analysis, etc. are widely used for data mining [7, 18, 87]. Usually, most of the statistical models are built from training data set. A variety of rules and patterns are then drawn from the model. Most of the data mining tasks are performed using one or more statistical approaches [18].

The statistical methods commonly used in data mining are described as follows [7, 18, 87]:

- Bayesian network: It represents the casual relationship among the variables, calculated through Bayesian probability theorem [88].
- Correlation: The relationship between two more variables/facts/dimensions can be determined using correlation [89].
- Regression: It is a derivation of a function to map a set of variables of various objects to an output variable [90].
- Cluster analysis: It groups the objects based on similarity measures so that objects that are similar to each other are located within same cluster [91].
- Discriminant analysis: It assigns data objects to one or more groups based on discriminant function [92].
- Factor analysis: It is used to understand and find out the main causes for the correlations and to identify the important ones [93].

4.2 Machine learning

Machine learning deals with the study of determining that how machines and humans can learn from data. Due to the importance of machine learning in data mining, a large number of the data mining algorithms have their roots in machine learning [83].

Machine learning increases levels of automation in the knowledge discovery in databases process to improve accuracy and efficiency. The systems produced by machine learning can be used regularly in the industry or education sector. In some of the applications, the machine learning methods gives performance better than the methods without learning [94, 85].

Inductive and deductive are two categories of machine learning. Deductive learning deals with facts and knowledge that existed over the time and then generates new knowledge from the old knowledge. In inductive learning, examples are generalized instead of starting with existing knowledge.

Meta-learning combines a number of detached learning processes in an intellectual fashion [95]. A meta-learning architecture exhibit two key behaviours: (1) an accurate final classification system (or final outcome), and (2) it must be fast, relative to an individual sequential learning algorithm [95]. For mining DSS solutions, RSA (rough set analysis) and DNA (dependency network analysis) have been suggested by Gengshen and Guenther [80].

The increasing popularity of Internet leads to the increase in network attacks. Therefore, intrusion detection (ID) is becoming the one of the key research areas for network security that. It is used to identify uncommon access or attacks to secure networks. Machine learning is also used in intrusion detection systems (IDS). IDSs monitor computers in case of security violations and trigger alerts to report any violation [96]. These reported alerts are given to an analyst for evaluation and initiation of an appropriate action. Two approaches based on reduction of the number of false positives in intrusion detection are proposed in Chih-Fong et al. [96].

Table 5 Classified work of reported literature related to data mining techniques

Reference	Data Mining Techniques						
	Statistics	Machine learning	Neural networks	Database system and data warehouse	Genetic algorithms	Fuzzy set and logic	Visualization
Abuaiadah [19]	√						
Angiulli et al. [21]		√	√				
Angiulli and Fassetti [22]	√						
Bhatnagar et al. [23]	√						
Bouguessa [24]	√						
Campello et al. [25]	√						√
Carpineto et al. [26]				√			
Chen et al. [13]				√			
Chen et al. [28]	√						
Chin-Yuan et al. [29]		√	√				
David et al. [75]		√					
Das et al. [30]	√						
Edward and Olgierd [76]		√					
Esling and Agon [77]	√			√			
Eyke [78]		√				√	
Eyke [79]		√				√	
Friedman [7]	√						
Gengshen and Guenther [80]		√					
Jain et al. [38]		√	√		√	√	
Jin et al. [39]	√						
Kate et al. [81]			√				
Koh and Ravana [41]	√	√	√				
Kosina and Gama [42]		√					
Kotsiantis [43]		√	√				
Kumar et al. [44]		√					√
Lee and Yun [45]		√		√			√
Mabroukeh and Ezeife [48]				√			
Mampaey and Vreeken [49]	√			√			
Menardi and Torelli [50]	√	√					
Mukhopadhyay et al. [51]					√		
Mu-Jung et al. [82]			√			√	
Padhraic [83]	√						
Pei et al. [52]		√					
Philip and Salvatore [95]		√					
Rafalak et al. [53]	√						
Saeed and Ali [84]			√				
Silva et al. [57]				√			
Sim et al. [59]	√						
Singh et al. [85]		√					

Table 5 continued

Reference	Data Mining Techniques						
	Statistics	Machine learning	Neural networks	Database system and data warehouse	Genetic algorithms	Fuzzy set and logic	Visualization
Sohrabi and Roshani [60]		√					
Susan et al. [61]		√					
Tan et al. [62]					√		
Tew et al. [63]	√						
Wang and Dong [64]	√	√					
Wang and Sun [65]	√		√				
Zhang et al. [68]		√				√	

4.3 Neural network

A neural network is a network or circuit of biological neurons. It has the capability to learn by examples which makes them flexible and powerful. Artificial neural network (ANN) is composed of artificial neurons or nodes and electrical signalling similar to the biological neural networks [97]. In ANN, knowledge is represented as a layered set of interconnected processors (also called as neurons). Different types of neural network models are also used to solve business problems as well as also play a vital role as a modern operations research tool [81].

Classification based on ANN to examine an effective forecast of future values is discussed in David et al. [75]. Saeed and Ali [84] proposed new privacy-preserving protocols for partitioned data based on extreme learning machine (ELM) and back-propagation (BP) algorithms.

Contemporary ANN approaches can also be used in spatial environmental data analysis. Valorisation and representativity of data is discussed in Kanevski et al. [98]. A hybrid model, based on support vector regression and multilayer perceptron ML algorithms, called as machine learning residuals sequential simulations (MLRSS) has also been presented in Kanevski et al. [98].

4.4 Database systems and data warehouses

Database-oriented and data warehouse-oriented approaches are not based on best model but uses existing data model to utilize the characteristics of the existing data [16]. To achieve scalability and great effectiveness of data mining tasks, that need to handle large data sets, the database technologies can be used for data mining. The systematic data analysis capabilities have been embedded in the recent commercial database systems also [1]. The iterative database scanning for the attribute focusing, attribute-oriented induction and frequent item sets are the major methods of this approach [28]. Multi-dimensionality nature of data structure in data warehouse also promotes multidimensional data mining [6].

4.5 Genetic algorithms

Genetic algorithms are based on concept of natural biological evaluation, i.e. processes of selection, reproduction, mutation, and survival of the fittest. Just like nature does, genetic algorithms can provide a better solution by combining the DNA of living beings [99]. But, in genetic algorithms the solutions are difficult to explain and no statistical measure exists to enable the user to understand why the particular solution has been reached [87].

4.6 Fuzzy sets

The concept of fuzzy sets theory was founded by Lotfi Zadeh. Fuzzy set defines the degree of membership based on the possibility value calculated with the help of membership function. It is widely used in classification and cluster analysis [6]. Fuzzy set theory is building potential contributions to the various applications of data mining, machine learning, and related fields [78, 79].

A knowledge discovery model based on integration of modification of the fuzzy transaction data-mining algorithm (MFTDA) and adaptive-network-based fuzzy inference systems (ANFIS) has been described in Mu-Jung et al. [82]. A machine learning approach combining fuzzy modelling which returns set of fuzzy rules was proposed by Edward and Olgierd [76].

4.7 Visualization

Visualization is a very useful data mining technique to identify and represent patterns in data sets. In visualization, data are translated into objects such as points, lines, and areas, etc. which are displayed in 2- or 3-dimensional space. By visual examination, the interesting patterns can be interactively explored by the users [18, 86]. Campello et al. [25] presented a framework for density-based estimates for visualization.

5 Real-life applications of data mining

Due to the power of data mining for data analytics, data mining uses in a wide range of real-life applications across variety of domains [100–102]. One or more data mining tasks, techniques and methods are applied in these applications [6, 15]. The various real-life applications of data mining are presented in the following sub-sections.

5.1 Telecommunication sector

Data mining is used by telecom/mobile service providers to formulate and design strategies for (i) marketing campaign, (ii) customer retention, (iii) packages for customers based on customer segmentation, (iv) optimum utilization of communication infrastructure, etc. By using classification and clustering, the mobile service providers can formulation strategies for their marketing campaign to promote direct marketing. With the help of clustering followed by classification, the customers can be segmented into various groups to predict the moving customers. The specific marketing strategies and packages can be formulated and designed for moving customers so that they can be retained with service provider. Based the identified customer groups, the specific packages can also be formulated based on the needs/requirements of these various customer groups. For designing packages the association analysis can also be used. The network usage pattern can be analyzed using data mining to identify the under-utilized and over-utilized network infrastructure so that the overall infrastructure can be optimally utilized and/or enhanced as per requirement [6, 15, 103–106].

5.2 Retail sector

Retail sector and super-market owners can be benefitted by data mining. With the help of data mining they can predict (i) buying behavior of the customers, (ii) market-basket analysis, (iii) choice of the customers, (iv) placement of products on shelves, (v) introduction of effective offers/-coupons/discounts, (vi) customer segmentation, etc. To discover the buying behavior of the customers and market-basket analysis the association analysis is used. Using association, frequent itemsets based on given support and confidence level can be discovered from the sales data so that these frequent itemsets can be placed nearby so that their sales can be increased. The marketing campaigns can be designed using RFM (recency, frequency, and monetary) grouping. By analyzing sales data using clustering, the best location (i.e. shelves) for the placement of products and best optimal offers can be discovered so that the sales can be increased. The sales data can also be analyzed to discover the various segments of the customers using clustering and/or classification. The different marketing campaigns and promotions/offers can be customized for discovered segments of customers. The customer who buys very less frequently but spends a lot shall be treated differently from the customer who buys very frequently but of fewer amounts [6, 15, 103, 105, 107].

5.3 Financial data analysis

The financial data in financial industry and banking facilitates systematic data analysis and data mining. Data mining for financial data analysis can be used for (i) loan payment prediction, (ii) customer credit policy analysis, (iii) customer segmentation for targeted marketing, (iv) detection of money laundering and other financial crimes, etc.

With the help of attribute ranking and attribute selection, methods of data mining, the customer payment history can be analyzed to discover (i) credit history, (ii) payment to income ratio, (iii) the term of the loan, etc. of the customers. This prediction will help the banks/financial inistitutions to decide their loan granting policy and to grant loans to the customers as per their score. Now these days, the banks and financial institutions checks the CIBIL score, which is based on data mining, of the customers before granting the loans to them [6, 15, 103–106].

5.4 Healthcare sector

Recently, the data mining is widely used in healthcare sector to (i) identify and analyze chronic diseases, (ii) to identify and discover symptoms, possible causes and medicines for effective treatments, (ii) track high-risk regions prone to the spread of disease, (iii) design programs to reduce the spread of disease, (iv) identify regions of patients, etc. In healthcare sector, the imaging/lab test data/reports are analyzed using data mining tasks such as clustering, classification, association and outlier detection. These tasks are used to identify/discover/predict the chronic diseases; their symptoms, possible causes and medicines so that these diseases can be treated effectively. The analysis can be further extended to identify and track the high-risk regions which are prone to the spread of disease. Based on the analysis, the campaigns can be designed for the regions to make people aware of the disease and their precautions. Using data mining, continuous comparison of symptoms, causes, and medicines, data

Table 6 Relationship between Data Mining Tasks and Data Mining Techniques

	Data mining tasks						
Data mining techniques	Summarization	Characterization and discrimination	Classification	Clustering	Association	Outlier analysis	Regression and trend analysis
Statistics	√	√	√	√	√	√	√
Machine learning		√	√	√	√	√	√
Neural networks		√	√	√	√	√	√
Database system	√	√			√	√	
Genetic algorithms			√	√	√	√	
Fuzzy set and logic		√	√	√	√	√	
Visualization		√	√	√		√	√

analysis can be performed to make effective treatments and the associated side-effects [6, 15, 103–106].

5.5 Fraud detection and crime prevention

The outliers can also be discovered using data mining from the vast amount of data. The outliers can be identified by discovering the infrequent patterns in the data. The infrequent patterns are generally belongs to fraudulent/criminal activity. Hence, with the help of outlier detection and/or infrequent pattern mining, the possible frauds can be identified and predicted so that the occurrence of crimes can be prevented [6, 15, 103, 107].

5.6 Customer relationship management (CRM)

Good customer relationships can be made by appealing more appropriate customers and better retention. Data mining can reinforce CRM by the identification and prediction of (i) database marketing, (ii) customer acquisition and customer retention campaigns, etc. [6, 15, 103, 107].

5.7 Recommender systems

Recommender systems give stakeholders with varied recommendations that may be of interest to the users using data mining. Recommender systems examine the user transactions, user profiles, keywords, common features among items to estimate an item for the user. Many data mining techniques such as machine learning, statistics, information retrieval, etc. are used in recommender systems. For example, in marketing, recommender system may recommend items which are either similar to the items queried by the user in the past or by looking at the other customer preferences which have similar taste as the user [103].

5.8 Online marketing/E-commerce

Various big brands/vendors of online marketing and e-commerce are also using the data mining to enhance their business. For examples: (i) E-commerce vendors discover the lowest price of the product using text mining on the web, (ii) large fast food chain vendors studies the ordering pattern of customers, waiting times, size of orders, etc. using big data mining to enhance their customer experiences, (iii) online media service providers also uses data mining to find out how to make a series or a movie popular among the customers [103].

6 Relationship between data mining tasks and data mining techniques

Data mining task is carried out with one or more data mining techniques. In data mining technique, one or more data mining methods can be applied. Table 6 represents the data mining tasks are carried out based on which major technique(s).

7 Summary and conclusion

There is need to evolve data mining to efficiently analyze the huge volume of data as well as to discover knowledge from it. The application domains of data mining are also increasing regularly. Hence, it is required to find the uniform methods/algorithms which can be implemented on large variety of applications without or with a few changes.

Table 7 Challenges of data mining research

Challenges	Addressed by
Development of unifying theory of data mining	Jackson [87]; Padhy et al. [109]
Use of intelligent interfaces and intelligent agents for generosity	Padhy et al. [109]
Development of adaptive, fault-tolerant and extendable system	Sawant et al. [71]
Ability to continually change and provide new understanding	Liao et al. [97]
Integration of qualitative and quantitative methods	Liao [47]
Distance metric learning for big data	Wang and Sun [65]
Formulation of generic distance learning metric	Wang and Sun [65]
Improving efficiency and scalability of data mining algorithms	Han et al. [6]
Dealing with diverse data types	Han et al. [6]; Silva et al. [57]
User interaction	Han et al. [6]; Esling and Agon [77]; Geng and Hamilton [32]
Data cube-oriented multidimensional data mining	Han et al. [6]
Meta-learning to automatically select or combine appropriate measures	Geng and Hamilton [32]
Convergence and hybrid approaches	Esling and Agon [77]
Parameter-free data mining	Esling and Agon [77]
Exhaustive benchmarking	Esling and Agon [77]
Adaptive mining algorithm dynamics	Esling and Agon [77]
Disassociation, privacy and incremental mining	Ceglar and Roddick [27]
Interactive and iterative mining	Ceglar and Roddick [27]
XML data clustering	Algergawy et al. [20]
Trade-off between scalability and quality of clustering algorithms	Algergawy et al. [20]
Detecting evolution of data distribution	Silva et al. [57]
The coherence of the cluster structure	Carpineto et al. [26]
Advanced visualization techniques to provide better overviews with clustered results	Carpineto et al. [26]
Reduction of learning/training time	Gibert et al. [110]
Alternative measures for cluster quality in the unsupervised and semi-supervised learning	Campello et al. [25]
Rare pattern mining on data streams	Koh and Ravana [41]
Scalable real time rare pattern mining	Koh and Ravana [41]
Rare pattern mining in Probabilistic Datasets	Koh and Ravana [41]
Dynamic and representative pattern mining on data streams for uncertain frequent patterns	Lee and Yun [45]

Most of the data mining systems employ a combination of methods to handle various types of data, data mining tasks and application areas [18].

A number of challenges of the data mining research have been stated by many researchers [108]. Some of these are presented in Table 7 and require more research attention.

Various data mining tasks and techniques help different companies to (i) gain knowledge, and (ii) increase their profitability by making amendments in procedures and operations. Data mining helps businesses in decision making through analysis of hidden patterns and trends [103].

Finally, it is concluded that (1) unification, scalability and optimization of data mining algorithms/methods, (2) cube-oriented multidimensional data mining, and (3) scalable real-time mining are the areas of data mining which also require more attention from researchers.

References

1. Fayadd U, Piatesky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AAAI Press/The MIT Press, Massachusetts Institute of Technology. ISBN 0–262 56097–6 Fayap
2. Fayadd U, Piatesky-Shapiro G, Smyth P (1996) Knowledge discovery and data mining: towards a unifying framework. In: Proceedings of the 2nd ACM international conference on knowledge discovery and data mining (KDD), Portland, pp 82–88

3. Heikki M (1996) Data mining: machine learning, statistics, and databases. In: SSDBM ’96: proceedings of the eighth international conference on scientific and statistical database management, June 1996, pp 2–9
4. Arora RK, Gupta MK (2017) e-Governance using data warehousing and data mining. Int J Comput Appl 169(8):28–31
5. Morik K, Bhaduri K, Kargupta H (2011) Introduction to data mining for sustainability. Data Min Knowl Discov 24(2):311–324
6. Han J, Kamber M, Pei J (2012) Data mining concepts and techniques, 3rd edn. Elsevier, Netherlands
7. Friedman JH (1997) Data mining and statistics: What is the connection? in: Keynote Speech of the 29th Symposium on the Interface: Computing Science and Statistics, Houston, TX, 1997
8. Turban E, Aronson JE, Liang TP, Sharda R (2007) Decision support and business intelligence systems. 8th edn, Pearson Education, UK
9. Gheware SD, Kejkar AS, Tondare SM (2014) Data mining: tasks, tools, techniques and applications. Int J Adv Res Comput Commun Eng 3(10):8095–8098
10. Kiranmai B, Damodaram A (2014) A review on evaluation measures for data mining tasks. Int J Eng Comput Sci 3(7):7217–7220
11. Sharma M (2014) Data mining: a literature survey. Int J Emerg Res Manag Technol 3(2):1–4
12. Venkatadri M, Reddy LC (2011) A review on data mining from past to the future. Int J Comput Appl 15(7):19–22
13. Chen M, Han J, Yu PS (1996) Data mining: an overview from a database perspective. IEEE Trans Knowl Data Eng 8(6):866–883
14. Gupta MK, Chandra P (2019) A comparative study of clustering algorithms. In: Proceedings of the 13th INDIACom-2019; IEEE Conference ID: 461816; 6th International Conference on “Computing for Sustainable Global Development”
15. Ponniah P (2001) Data warehousing fundamentals. Wiley, USA
16. Chandra P, Gupta MK (2018) Comprehensive survey on data warehousing research. Int J Inform Technol 10(2):217–224
17. Weiss SH, Indurkhya N (1998) Predictive data mining: a practical guide. Morgan Kaufmann Publishers, San Francisco
18. Fu Y (1997) Data mining: tasks, techniques, and applications. IEEE Potentials 16(4):18–20
19. Abuaiadah D (2015) Using bisect k-means clustering technique in the analysis of arabic documents. ACM Trans Asian Low-Resour Lang Inf Process 15(3):1–17
20. Algergawy A, Mesiti M, Nayak R, Saake G (2011) XML data clustering: an overview. ACM Comput Surv 43(4):1–25
21. Angiulli F, Fassetti F (2013) Exploiting domain knowledge to detect outliers. Data Min Knowl Discov 28(2):519–568
22. Angiulli F, Fassetti F (2016) Toward generalizing the unification with statistical outliers: the gradient outlier factor measure. ACM Trans Knowl Discov Data 10(3):1–26
23. Bhatnagar V, Ahuja S, Kaur S (2015) Discriminant analysis-based cluster ensemble. Int J Data Min Modell Manag 7(2):83–107
24. Bouguessa M (2013) Clustering categorical data in projected spaces. Data Min Knowl Discov 29(1):3–38
25. Campello RJGB, Moulavi D, Zimek A, Sander J (2015) Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Trans Knowl Discov Data 10(1):1–51
26. Carpineto C, Osinski S, Romano G, Weiss D (2009) A survey of web clustering engines. ACM Comput. Surv. 41(3):1–38
27. Ceglar A, Roddick JF (2006) Association mining. ACM Comput Surv 38(2):1–42
28. Chen YL, Weng CH (2009) Mining fuzzy association rules from questionnaire data. Knowl Based Syst 22(1):46–56
29. Fan Chin-Yuan, Fan Pei-Shu, Chan Te-Yi, Chang Shu-Hao (2012) Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. Expert Syst Appl 39:8844–8851
30. Das R, Kalita J, Bhattacharya (2011) A pattern matching approach for clustering gene expression data. Int J Data Min Model Manag 3(2):130–149
31. Dincer E (2006) The k-means algorithm in data mining and an application in medicine. Kocaeli Univesity, Kocaeli
32. Geng L, Hamilton HJ (2006) Interestingness measures for data mining: a survey. ACM Comput Surv 38(3):1–32
33. Gupta MK, Chandra P (2019) P-k-means: k-means using partition based cluster initialization method. In: Proceedings of the international conference on advancements in computing and management (ICACM 2019), Elsevier SSRN, pp 567–573
34. Gupta MK, Chandra P (2019) An empirical evaluation of k-means clustering algorithm using different distance/similarity metrics. In: Proceedings of the international conference on emerging trends in information technology (ICETIT-2019), emerging trends in information technology, LNEE 605 pp 884–892 DOI: https://doi.org/10.1007/978-3-030-30577-2_79
35. Hea Z, Xua X, Huangb JZ, Denga S (2004) Mining class outliers: concepts, algorithms and applications in CRM. Expert Syst Appl 27(4):681e97
36. Hung LN, Thu TNT, Nguyen GC (2015) An efficient algorithm in mining frequent itemsets with weights over data stream using tree data structure. IJ Intell Syst Appl 12:23–31
37. Hung LN, Thu TNT (2016) Mining frequent itemsets with weights over data stream using inverted matrix. IJ Inf Technol Comput Sci 10:63–71
38. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput. Surv 31(3):1–60
39. Jin H, Wang S, Zhou Q, Li Y (2014) An improved method for density-based clustering. Int J Data Min Model Manag 6(4):347–368
40. Khandare A, Alvi AS (2017) Performance analysis of improved clustering algorithm on real and synthetic data. IJ Comput Netw Inf Secur 10:57–65
41. Koh YS, Ravana SD (2016) Unsupervised rare pattern mining: a survey. ACM Trans Knowl Discov Data 10(4):1–29
42. Kosina P, Gama J (2015) Very fast decision rules for classification in data streams. Data Min Knowl Discov 29(1):168–202
43. Kotsiantis SB (2007) Supervised machine learning: a review of classification techniques. Informatica 31:249–268
44. Kumar D, Bezdek JC, Rajasegarar S, Palaniswami M, Leckie C, Chan J, Gubbi J (2016) Adaptive cluster tendency visualization and anomaly detection for streaming data. ACM Trans Knowl Discov Data 11(2):1–24
45. Lee G, Yun U (2017) A new efficient approach for mining uncertain frequent patterns using minimum data structure without false positives. Future Gener Comput Syst 68:89–110
46. Li G, Zaki MJ (2015) Sampling frequent and minimal boolean patterns: theory and application in classification. Data Min Knowl Discov 30(1):181–225. https://doi.org/10.1007/s10618-015-0409-y
47. Liao TW, Triantaphyllou E (2007) Recent advances in data mining of enterprise data: algorithms and applications. World Scientific Publishing, Singapore, pp 111–145
48. Mabroukeh NR, Ezeife CI (2010) A taxonomy of sequential pattern mining algorithms. ACM Comput Surv 43:1
49. Mampaey M, Vreeken J (2011) Summarizing categorical data by clustering attributes. Data Min Knowl Discov 26(1):130–173
50. Menardi G, Torelli N (2012) Training and assessing classification rules with imbalanced data. Data Min Knowl Discov 28(1):4–28. https://doi.org/10.1007/s10618-012-0295-5

51. Mukhopadhyay A, Maulik U, Bandyopadhyay S (2015) A survey of multiobjective evolutionary clustering. ACM Comput Surv 47(4):1–46
52. Pei Y, Fern XZ, Tjahja TV, Rosales R (2016) 'Comparing clustering with pairwise and relative constraints: a unified framework. ACM Trans Knowl Discov Data 11:2
53. Rafalak M, Deja M, Wierzbicki A, Nielek R, Kakol M (2016) Web content classification using distributions of subjective quality evaluations. ACM Trans Web 10:4
54. Reddy D, Jana PK (2014) A new clustering algorithm based on Voronoi diagram. Int J Data Min Model Manag 6(1):49–64
55. Rustogi S, Sharma M, Morwal S (2017) Improved Parallel Apriori Algorithm for Multi-cores. IJ Inf Technol Comput Sci 4:18–23
56. Shah-Hosseini H (2013) Improving K-means clustering algorithm with the intelligent water drops (IWD) algorithm. Int J Data Min Model Manag 5(4):301–317
57. Silva JA, Faria ER, Barros RC, Hruschka ER, de Carvalho ACPLF, Gama J (2013) Data stream clustering: a survey. ACM Comput Surv 46(1):1–31
58. Silva A, Antunes C (2014) Multi-relational pattern mining over data streams. Data Min Knowl Discov 29(6):1783–1814. https://doi.org/10.1007/s10618-014-0394-6
59. Sim K, Gopalkrishnan V, Zimek A, Cong G (2012) A survey on enhanced subspace clustering. Data Min Knowl Discov 26(2):332–397
60. Sohrabi MK, Roshani R (2017) Frequent itemset mining using cellular learning automata. Comput Hum Behav 68:244–253
61. Craw Susan, Wiratunga Nirmalie, Rowe Ray C (2006) Learning adaptation knowledge to improve case-based reasoning. Artif Intell 170:1175–1192
62. Tan KC, Teoh EJ, Yua Q, Goh KC (2009) A hybrid evolutionary algorithm for attribute selection in data mining. Expert Syst Appl 36(4):8616–8630
63. Tew C, Giraud-Carrier C, Tanner K, Burton S (2013) Behavior-based clustering and analysis of interestingness measures for association rule mining. Data Min Knowl Discov 28(4):1004–1045
64. Wang L, Dong M (2015) Exemplar-based low-rank matrix decomposition for data clustering. Data Min Knowl Discov 29:324–357
65. Wang F, Sun J (2014) Survey on distance metric learning and dimensionality reduction in data mining. Data Min Knowl Discov 29:534–564
66. Wang B, Rahal I, Dong A (2011) Parallel hierarchical clustering using weighted confidence affinity. Int J Data Min Model Manag 3(2):110–129
67. Zacharis NZ (2018) Classification and regression trees (CART) for predictive modeling in blended learning. IJ Intell Syst Appl 3:1–9
68. Zhang W, Li R, Feng D, Chernikov A, Chrisochoides N, Osgood C, Ji S (2015) Evolutionary soft co-clustering: formulations, algorithms, and applications. Data Min Knowl Discov 29:765–791
69. Han J, Fu Y (1996) Exploration of the power of attribute-oriented induction in data mining. Adv Knowl Discov Data Min. AAAI/MIT Press, pp 399-421
70. Gupta A, Mumick IS (1995) Maintenance of materialized views: problems, techniques, and applications. IEEE Data Eng Bull 18(2):3
71. Sawant V, Shah K (2013) A review of distributed data mining using agents. Int J Adv Technol Eng Res 3(5):27–33
72. Gupta MK, Chandra P (2019) An efficient approach for selection of initial cluster centroids for k-means clustering algorithm. In: Proceedings international conference on recent developments in science engineering and technology (REDSET-2019), November 15–16 2019
73. Gupta MK, Chandra P (2019) MP-K-means: modified partition based cluster initialization method for k-means algorithm. Int J Recent Technol Eng 8(4):1140–1148
74. Gupta MK, Chandra P (2019) HYBCIM: hypercube based cluster initialization method for k-means. IJ Innov Technol Explor Eng 8(10):3584–3587. https://doi.org/10.35940/ijitee.j9774.0881019
75. Enke David, Thawornwong Suraphan (2005) The use of data mining and neural networks for forecasting stock market returns. Expert Syst Appl 29:927–940
76. Mezyk Edward, Unold Olgierd (2011) Machine learning approach to model sport training. Comput Hum Behav 27:1499–1506
77. Esling P, Agon C (2012) Time-series data mining. ACM Comput Surv 45(1):1–34
78. Hüllermeier Eyke (2005) Fuzzy methods in machine learning and data mining: status and prospects. Fuzzy Sets Syst 156:387–406
79. Hullermeier Eyke (2011) Fuzzy sets in machine learning and data mining. Appl Soft Comput 11:1493–1505
80. Gengshen Du, Ruhe Guenther (2014) Two machine-learning techniques for mining solutions of the ReleasePlanner™ decision support system. Inf Sci 259:474–489
81. Smith Kate A, Gupta Jatinder ND (2000) Neural networks in business: techniques and applications for the operations researcher. Comput Oper Res 27:1023–1044
82. Huang Mu-Jung, Tsou Yee-Lin, Lee Show-Chin (2006) Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge. Knowl Based Syst 19:396–403
83. Padhraic S (2000) Data mining: analysis on grand scale. Stat Method Med Res 9(4):309–327. https://doi.org/10.1191/096228000701555181
84. Saeed S, Ali M (2012) Privacy-preserving back-propagation and extreme learning machine algorithms. Data Knowl Eng 79–80:40–61
85. Singh Y, Bhatia PK, Sangwan OP (2007) A review of studies on machine learning techniques. Int J Comput Sci Secur 1(1):70–84
86. Yahia ME, El-taher ME (2010) A new approach for evaluation of data mining techniques. Int J Comput Sci Issues 7(5):181–186
87. Jackson J (2002) Data mining: a conceptual overview. Commun Assoc Inf Syst 8:267–296
88. Heckerman D (1998) A tutorial on learning with Bayesian networks. Learning in graphical models. Springer, Netherlands, pp 301–354
89. Politano PM, Walton RO (2017) Statistics & research methodol. Lulu. com
90. Wetherill GB (1987) Regression analysis with application. Chapman & Hall Ltd, UK
91. Anderberg MR (2014) Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks, vol 19. Academic Press, USA
92. Mihoci A (2017) Modelling limit order book volume covariance structures. In: Hokimoto T (ed) Advances in statistical methodologies and their application to real problems. IntechOpen, Croatia. https://doi.org/10.5772/66152
93. Thompson B (2004) Exploratory and confirmatory factor analysis: understanding concepts and applications. American Psychological Association, Washington, DC (ISBN:1-59147-093-5)
94. Kuzey C, Uyar A, Delen (2014) The impact of multinationality on firm value: a comparative analysis of machine learning techniques. Decis Support Syst 59:127–142
95. Chan Philip K, Salvatore JS (1997) On the accuracy of meta-learning for scalable data mining. J Intell Inf Syst 8:5–28
96. Tsai Chih-Fong, Hsu Yu-Feng, Lin Chia-Ying, Lin Wei-Yang (2009) Intrusion detection by machine learning: a review. Expert Syst Appl 36:11994–12000

97. Liao SH, Chu PH, Hsiao PY (2012) Data mining techniques and applications—a decade review from 2000 to 2011. Expert Syst Appl 39:11303–11311
98. Kanevski M, Parkin R, Pozdnukhov A, Timonin V, Maignan M, Demyanov V, Canu S (2004) Environmental data mining and modelling based on machine learning algorithms and geostatistics. Environ Model Softw 19:845–855
99. Jain N, Srivastava V (2013) Data mining techniques: a survey paper. Int J Res Eng Technol 2(11):116–119
100. Baker RSJ (2010) Data mining for education. In: McGaw B, Peterson P, Baker E (eds) International encyclopedia of education, 3rd edn. Elsevier, Oxford, UK
101. Lew A, Mauch H (2006) Introduction to data mining and its applications. Springer, Berlin
102. Mukherjee S, Shaw R, Haldar N, Changdar S (2015) A survey of data mining applications and techniques. Int J Comput Sci Inf Technol 6(5):4663–4666
103. Data mining examples: most common applications of data mining (2019). https://www.softwaretestinghelp.com/data-mining-examples/. Accessed 27 Dec 2019
104. Devi SVSG (2013) Applications and trends in data mining. Orient J Comput Sci Technol 6(4):413–419
105. Data mining—applications & trends. https://www.tutorialspoint.com/data_mining/dm_applications_trends.htm
106. Keleş MK (2017) An overview: the impact of data mining applications on various sectors. Tech J 11(3):128–132
107. Top 14 useful applications for data mining. https://bigdata-madesimple.com/14-useful-applications-of-data-mining/. Accessed 20 Aug 2014
108. Yang Q, Wu X (2006) 10 challenging problems in data mining research. Int J Inf Technol Decis Making 5(4):597–604
109. Padhy N, Mishra P, Panigrahi R (2012) A survey of data mining applications and future scope. Int J Comput Sci Eng Inf Technol 2(3):43–58
110. Gibert K, Sanchez-Marre M, Codina V (2010) Choosing the right data mining technique: classification of methods and intelligent recommendation. In: International Congress on Environment Modelling and Software Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada