DOI:10.22413/ijatest/2022/v7/i1/1

Pattern Mining Applications: Intensive Case Study

Daphne. S[1] *and Anbuchelian. S*[2]

[1]*Research Scholar, Dept. of Ramanujan Computing Centre, Anna University, Chennai, India*
[2]*Assistant Professor(Sl. Gr), Dept. of Ramanujan Computing Centre, Anna University, Chennai, India*
E-Mail: daphnesam7@gmail.com

***Abstract**: Pattern Mining refers to the extraction of recurring relationships or patterns and knowledge from large amounts of raw data and sometimes termed as finding hidden in-formation in a database. The simplest form of frequent patterns mining is to identify the association and correlation rules using any efficient and scalable algorithm. Advanced forms of patterns can be mined using multilevel associations and multidimensional associations, high-dimensional patterns, quantitative association rules, rare patterns and negative patterns. Pattern mining helps in major data mining issues like clustering, outlier detection, and classification. Finding frequent patterns also supports in general-purpose ap-plications such as software bug analysis, web log analytics, biological data analysis etc. The key dimensions of the existing applications of pattern mining methods, their benefits, and limitations are presented in this paper. Further, we introduce and summarize the latest research techniques in pattern mining applications.*

Keywords*: Data Mining, Pattern Based Mining, Classification, Clustering, Semantic Annotation, Collaborative Filtering, Privacy-Preserving.*

I. Introduction

Data mining entails powerful information series and warehousing in addition to computer processing. Pattern mining is examining data styles in massive quantity of information with the help of more than one software. Pattern mining term can be used interchangeably with frequent pattern mining if no ambiguity exists. Since pattern mining includes rare and negative patterns whereas frequent pattern mining does not include it.

Han et al. [1], classified major data mining trends into 3 sub-categories: Statistical Data Mining, Foundations of Data Mining and Visual and Audio Data Mining. They have also described about the general road map for the pattern mining as 3 sub-categories: different kinds of pattern mining and rules; mining methods; extensions and applications.

Even though there have been many studies mainly addressing the 3 sub-categories of pattern mining, in this paper we particularly concentrate on all the major applications of pattern mining in depth. We also look at the comparative analysis of the latest research articles related to applications of pattern mining.

- Pattern-Based Classification
- Pattern-Based Clustering
- Pattern-Based Semantic Annotation
- Collaborative Filtering
- Privacy-Preserving

Section 2 presents the following five major applications of pattern mining and reviews comparative analysis of the algorithm based in these techniques. Finally, section 3 summarizes and concludes the comparative case study.

II. APPLICATIONS OF PATTERN MINING

The applications of pattern mining play an essential role in various domains like financial data mining, retail and telecommunication industries, science and technologies, intrusion detection and prevention, recommender systems etc. Moreover, we start off by introducing the pattern-based classification then followed by clustering, semantic annotation, collaborative filtering and finally privacy-preserving. In each section, we extend our discussion to compare latest research articles based on experimental evaluation, benefits and limitations.

A. Pattern-Based Classification

Classification can be used to categorize a banking transaction as fraudulent or genuine, predict performance in manufacturing product and medical diagnostics to predict which treatment is best suitable for a patient. Classification has 2-steps, starts by building a classification model using previous data then determine if model is acceptable and finally, accept the model to classify any data.

Figure 1. shows the major techniques of classifications: Decision tree classifiers, Bayesian classifiers, k-Nearest-Neighbour Classifiers and Support Vector Machine which are discussed as follows.

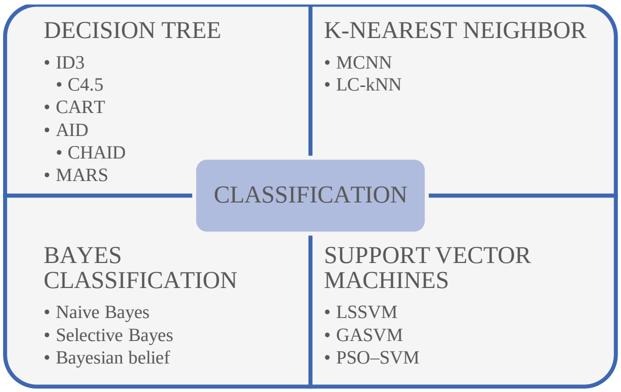

Fig. 1. Methods of pattern-based classification.

Decision tree classifiers is a collection of nodes represented in tree structure with topmost node as root, leaf nodes denote decisions. Decision tree classifiers learning step uses attribute selection measures to select the attribute which best splits the tuples into distinct classes. Common measures of attribute selection are ID3 (Iterative Dichotomiser), C4.5 (a successor of ID3), Classification and Regression Trees (CART), CHAID and MARS.

DOI:10.22413/ijatest/2022/v7/i1/1

Bayesian classifiers is the strongest statistical dependencies among the tuples using the statistical connections to predict class membership. Popular variations of Bayesian classifiers include Naive Bayes, Selective Bayes, Bayesian belief.

K-Nearest-Neighbor Classifiers also known as instance-based method, since all learning are based on instances. It uses distance-based comparison that intrinsically assign equal weight to each attribute. [1] This method combines attribute weighting and noise pruning of data tuples to avoid poor accuracy.

SVM is a linear and non-linear data classifier. SVM is an extension of perceptron with more equitable choice of separating hyperplane and less prone to overfitting. The common approaches of SVM are LSSVM, GASVM, PSO–SVM.

In summary, the comparative study of classification algorithms are in table 1 and studies comparing classification algorithms in latest research articles are discussed in table 2.

B. *Pattern-Based Clustering*

Clustering is an unsupervised learning where class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance. Given a hard and fast of statistics factors, clustering set rules is used to categories every statistics factor into a particular organization. In theory, statistics factors which can be with inside the equal organization ought to have comparable homes and/or features, whilst statistics factors in distinct corporations ought to have exceedingly numerous homes and/or features.

Hierarchical Clustering decomposes the set of objects data to form a dendrogram. The dendrogram is formed in two ways: bottom-up or agglomerative and top-down or divisive method. In agglomerative, each object is initially considered as a single-element cluster (leaf). It successively merges the objects or groups according to some measures like the distance between the two centres of two groups and this is done until all of the groups are merged into one. In Divisive, it begins with the root, in which all objects are included in a single cluster. In each successive iteration, a cluster is split into smaller clusters according to some measures until eventually each object is in one cluster, or until a termination condition holds.

In partitioning method, the set of objects data are partitioned into a predetermined number of clusters (k) in this process, so that within cluster variances are kept to a minimum. The most widely used centroid-based clustering algorithm is k-means clustering.

The set of objects data are grouped together in a density-dependent system based on the density of tightly packed points. Two common density-based algorithms are DBSCAN and OPTICS. In FEM-DBSCAN [2], Fisher expectation maximization (FEM) is used to adaptively divide the feature space into some subspaces in which no cluster is shared by adjacent subspaces. Following that, density-based spatial clustering (DBSCAN) is utilized for applications with noise to each partition, resulting in lower computational cost on each thread and improved learning of its parameters on each subspace.

Figure 2. shows the major techniques of Clustering: Hierarchical Method, Partitioning Method and Density-Based Methods are compared and discussed in table 3.

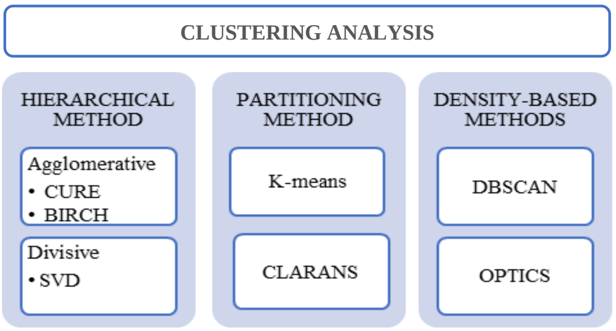

Fig. 2. Methods of pattern-based clustering.

The majority of real-time applications deal with high dimensional data. Cluster analysis and other data mining functions become increasingly complex as the number of dimensions grows. Thus, transforming the data to a graph and capture the data's complex geometry via graph topology, allows graph-theoretical principles and tools from complex network analysis to expose the data's structure. Neighbourhood based method (ε-ball, kNN and CkNN graphs) and Minimum spanning tree-based methods (PMST and RMST graphs) are the two categories of geometric based graphs clustering.

CkNN [3] is a graph-theoretical approach to data clustering that combines the construction of a graph from data with Markov Stability, a multiscale group detection system.

The data is clustered and clusters are established based on the likelihood of belonging to the same probability distribution in probabilistic-based clustering. Gaussian Mixture Model, Binomial distribution, fuzzy based some well-known methods of probabilistic-based clustering. In [4] a probability k-means clustering algorithm is used to segment decision makers with similar features into different sub-groups.

Grid based method for subspace clustering is a variant of conventional clustering that looks for clusters in various subspaces within a dataset. Some of the Bottom-up Subspace Search clustering techniques include SUBSCALE [5], CLIQUE, MAFIA, ENCLUS [6]. PROCLUS, ORCLUS, FINDIT [7] are some of the top-down subspace clustering algorithm.

MAFIA [6] is a modified CLIQUE algorithm that uses histograms to create variable-sized grids. The technique looks for non-minimally correlated subspaces with significant clustering, which are of interest. The SUBSCALE [5] clustering algorithm finds non-trivial subspace clusters for a k-dimensional data set with minimal cost and just k database scans. This algorithm is extremely parallelizable and scales well with the dataset's dimensionality.

Hence, we have depicted advanced or complex data types clustering algorithms in figure 3. Moreover, studies comparing the implementation of clustering algorithms in latest research articles are discussed in table 4.

DOI:10.22413/ijatest/2022/v7/i1/1

TABLE 1. COMPARATIVE ANALYSIS OF PATTERN-BASED CLASSIFICATION ALGORITHMS.

Classification Algorithm		**Type**	**Highlighted features**	**Goal**	**Problems addressed**	**Benefits**	**Limitations**
Decision Tree	ID3 [8]	Supervised learning	Uses *Entropy function* & *Information gain*	Create a training model to make predictions by learning simple decision rules inferred from prior data.	- Regression and classification problems - Attribute’s selection	-Rules can be generated and they are easy to interpret and understand. -It is scalable for large database.	-Overfitting in Decision Trees -Handling missing data is difficult.
	C4.5 [9]		C4.5 uses *Gain ratio*				
	CART [10]		Uses *Gini index*. Higher the Gini index higher the homogeneity.				
	CHAID [11]		Uses multi-level splits points.				
	MARS [12]		Uses hinge functions; Can handle both continuous and categorical data.				
K-Nearest Neighbour	ICNN [13]	Supervised learning	Histogram of gradient (HOG) technique	Computes the distance between each test sample and training samples in the dataset and return nearest neighbours.	- Linear time complexity over the sample size (i.e., High cost)	Efficient Data Pruning	- Heavy computations - Sensitive to Outliers
	LC-kNN [14]		Different shape data clusters and effective metrics for accurate classification.				
Bayes Classification	Naive Bayesian [15]	Supervised learning	Class-Conditional Independence; Max posteriori probability	A directed acyclic graph in which each edge corresponds to a conditional dependency, and each node corresponds to a unique random variable.	-Lack of probability data - Factor correlation in quantitative assessment	-Ability to incorporate prior information.	Handling uncertainty
	Bayesian Belief [16]		Joint-Conditional Probability Distribution				
	Feature Correlated Naïve Bayes [17]		Conditional Probability Table				
SVM	LSSVM [18]	Supervised learning	-Equality constraints for linear programming; -Squared loss function from error variable	Finding a hyperplane that best separates the features into different domains.	- Unbalanced data sets -Data are linearly separable /Inseparable	- Support vectors - Short description of the learned model	Slow training process
	PSO–SVM [19]		Safety risk predictions				

DOI:10.22413/ijatest/2022/v7/i1/1

TABLE 2. COMPARATIVE ANALYSIS OF RESEARCH ARTICLES IN PATTERN-BASED CLASSIFICATION ALGORITHMS.

Articles based on Classification Algorithm	Datasets	Classification	Implementation Tools & Metrics	Evaluation Results	Benefits	Limitations
Two-Step Classification with SVD Preprocessing of Distributed Massive Datasets in Apache Spark [20]	Higgs-binomial classification & PAMAP - multinomial classification	Two-step classifier implemented on several classification algorithms.	Apache Spark; Apache MLlib; PySpark; SparkQL; Preprocessing - Singular Value Decomposition (SVD); 10V data; F1 Metrics.	*Higgs Dataset*- logistic regression classifier showed very low computational time. *PAMAP dataset*- Decision tree classifier was the best option due better metric scores.	-Proposed architecture outperforms the individual classifiers - Capable of limiting vulnerability from cyber attacks	-Overfitting procedure - Randomly split the dataset
MC4.5 decision tree algorithm: an improved use of continuous attributes [21]	30 datasets	Decision tree	MC4.5; statistical mean; continuous attributes; Information gain;	Proposed algorithm generates smaller decision trees with better accuracy comparing to the C4.5 algorithm.	Minimization of information loss and reduction in time complexity.	N/A
Evaluation of tree-base data mining algorithms in land used/land cover mapping in a semi-arid environment [22]	A number of training and testing dataset for image classification with four class region pixels.	Tree-base data mining;	WEKA 3.8; R programming language; Land Use Land Cover (LULC) mapping; Reduced Error Pruning Tree (REP Tree); semi-arid environment.	LAD Tree, J48 Consolidated, and BF Tree classifiers proves better where there are few training datasets.	- A 10-fold cross-validation technique is used to prevent overfitting issue.	N/A
Combining deep generative and discriminative models for Bayesian semi-supervised learning [23]	Image recognition datasets-Half-Moons; SecStr; MNIST; fashion-MNIST;	Bayesian semi-supervised learning	Variational autoencoders - class of DGMs; Deep blended models are investigated with MAP estimates and approximate Bayesian inference.	-Blended version of M2 model achieves convergence & is faster than M2. -Accuracy is good with lesser Expected calibration error (ECE) for semi-supervised setting.	-Used in downstream decision-making task. -Achieving approximate Bayesian inference.	Accuracy is affected for fully supervised model
CNN-SVM Learning Approach Based Human Activity Recognition [24]	MSR Daily Activity 3D dataset	CNN-SVM Classification	Matlab 2018, NVIDIA GeForce 960M GPU, 64 GB & Intel Core i7-6700 HQ (2.60 GHz) processor.	- Accuracy 99.92% - Best Performance metrics	Pretrained ResNet model with Multi-Layer Perception (MLP)	Small pretrained dataset applied.

DOI:10.22413/ijatest/2022/v7/i1/1

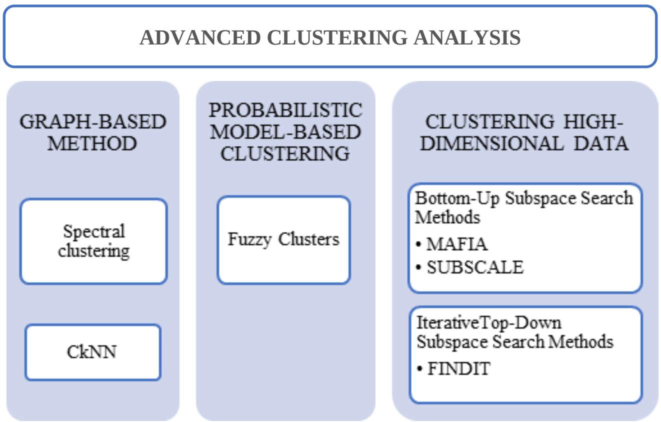
ADVANCED CLUSTERING ANALYSIS
GRAPH-BASED METHOD
Spectral clustering
CkNN
PROBABILISTIC MODEL-BASED CLUSTERING
Fuzzy Clusters
CLUSTERING HIGH-DIMENSIONAL DATA
Bottom-Up Subspace Search Methods
• MAFIA
• SUBSCALE
IterativeTop-Down Subspace Search Methods
• FINDIT

Fig. 3. Advanced Methods of pattern-based clustering.

C. *Pattern-Based Semantic Annotation*

Frequent pattern mining usually is associate intermediate step for improved information understanding and additional powerful information analysis. For instance, it is used as a feature extraction step for classification, that is usually named as pattern-based classification. Pattern-based clustering, on the other hand, has demonstrated its superiority in clustering high-dimensional data. Patterns can be used for semantic annotation or contextual analysis to improve data understanding. Pattern analysis is also used in recommender systems, which suggest information items (e.g., books, web pages etc.) that are likely to be of interest to the user based on the patterns of other users. Different analysis tasks may necessitate mining instead of variety of patterns.

Semantic annotation, similar to dictionary annotations, provide semantic information about the textual terms like "frequent, pattern". This data consists of context indicators (e.g., terms indicating the context of that pattern), the most representative data transactions (e.g., fragments or sentences containing the term), and the most semantically similar patterns (e.g., "maximal, pattern" is semantically similar to "frequent, pattern") [1]. The annotations provide a view of the pattern's context from various angles, which aids in comprehension.

Here the challenge is to determine which types of semantic annotation data should be extracted from available sources and how to filter it appropriately to match the moving object at hand. Zhang et al. [25] , has used automatic semantic annotation method for IoT data resources. They convey additional information, in this case related to the activity (e.g., work, shopping) or means of transportation (on foot, on bus, etc.).

However, in any application, a user frequently has some prior knowledge that can provide useful guidance for context modelling and semantic annotation of a given pattern, potentially leading to much more useful annotations. To facilitate the search, composition, and reuse of web services, their descriptions can be semantically annotated using ontology definitions, leading to semantic web services. Calache & Farias [26], has used semantic annotation using SAWSDL framework for represent semantic web services in graphical notations. Furthermore, the collaborative approach to semantic annotation has been advantageous.

D. *Collaborative Filtering*

Today's online shopping are packed with millions of goods and services. Data mining techniques are used by collaborative recommender systems to make personalized product recommendations based on the opinions of other customers. The collaborative approach (also known as the collaborative filtering approach) may take into account a user's social environment. These recommender systems assist consumers by recommending products that are likely to be of interest to the user, such as electronic items, books, food other services.

To find similarities between items and customer preferences, recommender systems employ a wide range of techniques from information retrieval, statistics, machine learning, and data mining. As part of their marketing strategy, Amazon is a pioneer in using collaborative recommender systems which provides "a personalized store for every customer". [1] Collaborative recommender systems can be memory (or heuristic) based or model based. Memory-based methods make rating predictions using heuristics based on the entire collection of items previously rated by users. Pearson's correlation coefficient, cosine similarity, and the k-nearest-neighbor approach are the most popular heuristics approaches. Model-based collaborative recommender systems learn a model from a set of ratings, that would be used to make rating predictions. The most popular models are probabilistic models, clustering, and Bayesian networks.

Scalability and providing quality recommendations to consumers are major challenges for recommender systems. Cold start and relevant data history sparsity continue to be major issues for collaborative filtering. Xia et al. [27], have used GRU_Attention Neural collaborative filtering, (GANCF) recommender method that incorporates auxiliary information from the item and user. The attention mechanism learns the weights of words from the context feature and chooses informative words based on those weights. These combined features are then used to extract text features while also taking into account the order and context of words and these features could also be utilized for rating prediction.

Vahidi Farashah et al. [28], have used Collaborative Recommender System based on the Hybrid Similarity Criterion, in which similarities between the new user and the users in the selected category are calculated based on a threshold (lambda). Based on the adjacency matrix, the higher-rated movie services are then recommended to the new user. After that Pro-FriendLink algorithm connects to users and ultimately recommends users with higher credibility in the social network to the target user.

E. *Privacy-Preserving*

Thanks to recent technological advancements, data mining techniques are now being used in our daily lives without our knowledge. This imposes the need to implement privacy-preserving on the data published in order to avoid potential violation of individual privacy and incorporate data protection rights as well.

DOI:10.22413/ijatest/2022/v7/i1/1

TABLE 3. COMPARATIVE ANALYSIS OF PATTERN-BASED CLUSTERING ALGORITHMS.

Clustering Algorithm		**Type**	**Highlighted features**	**Goal**	**Problems addressed**	**Benefits**	**Limitations**
Hierarchical Clustering	CURE [29]	Unsupervised learning	Used for problems which involves point linkages	Recursive partitioning of a dataset into clusters obtained by grouping tree-based two-dimensional plot.	- Image segmentation - City planning	- Simple Concept; - Less sensitive to noise.	- Cluster merging/splitting is permanent. - Difficult in handling convex shaped clusters. - Divisive methods can be computational hard.
	BIRCH [30]		Computes the optimal threshold even without the global clustering phase. Thus, removes supercluster splitting with flat tress.				
	SVD [31]		First-order (Euler-like) and second-order (Runge-like) numerical learning methods				
Partitioning Method	K-means [32] [33]	Unsupervised learning	Vertical fragmentation and allocation.	Distance-based; Each object must belong to exactly one group.	- Data locality maximization. - Communication costs reduction are met.	- Computationally faster method. - Scalable; - Faster for low dimensional data.	- Not handle non-globular data of different size and densities. - Sensitive to outliers & noise - Restricted to data which has the notion of centroid.
	CLARANS [7]		Two parameters: the maximum number of neighbours examined (maxneighbor) and the number of local minima obtained (numlocal). Handles polygon objects efficiently				
Density-Based Clustering	DBSCAN [34]	Unsupervised learning	Starts with an arbitrary instance in data set and retrieves all instances of data set with reference to epsilon (Eps) and minimum points.	Separates data points into three classes: - Hub points - Edge points - Noise points.	- Crystallography of x-ray. - Anomaly detection in temperature data.	- Can discover arbitrarily shaped clusters. - Find cluster completely surrounded by different clusters.	- Datasets with altering densities are tricky. - Sampling affects density measures.
	OPTICS [35]		Word-Ordering Index and Order Merging is preferable when it is hard to set the density parameters of DBSCAN.				

DOI:10.22413/ijatest/2022/v7/i1/1

TABLE 4. COMPARATIVE ANALYSIS OF RESEARCH ARTICLES IN PATTERN-BASED CLUSTERING ALGORITHMS.

Articles based on Clustering Algorithm	Datasets	Clustering	Implementation Tools & Metrics	Evaluation Results	Benefits	Limitations
Land Use Land Cover map segmentation using Remote Sensing: A Case study of Ajoy river watershed, India [36]	Salinas - A hyper spectral and Ajoy river catchment imagery.	Hybrid 2-Dimensional Cellular-Automata model based on K-means segmentation (KCA)	MATLAB R2014a; Quantitative evaluation and Statistical analysis (Median values, P-values and H-values).	Quantitative evaluation: Internal indices: Smallest Davies-Bouldin (DB) index value and highest Dunn index value; External indices - higher scores: Rand, Adjusted Rand, Jaccard and Fowlkes-Mallows (FM). Statistical significance test: Wilcoxon's rank sum	- Improvement over detection of mixed land cover regions in the satellite image. - Optimized quantitative internal indices and external indices.	Segmentation method is done with available true class labels for hyper-spectral.
A Clustering Algorithm to Improve the Scan Statistic in Sensor Detection Systems [37]	Three sensors system; Sensor form a regular square grid; Sensors at random locations	Voronoi diagram partitions in K convex polygonal regions.	Wireless sensor network, signal and noise models, cluster set, fusion center, ZigBee devices and computational tools.	Monte Carlo simulations; ZigBee Protocol	- Clustering set improves the detection performance in scan statistics.	A suboptimal cluster sets is considered instead of optimal cluster set for a set of sensor locations.
KPIs-Based Clustering and Visualization of HPC Jobs: A Feature Reduction Approach [38]	Centro de Supercomputación de Galicia (CESGA): 195 running nodes, 11 KPIs, 9006 jobs executed in a period of ten months.	K-Means clustering - Time series data with twofold feature selection (literature and variance based)	CESGA Big Data platform to extract Key Performance Indicators (KPIs) data. Hadoop ecosystem with HDFS, Apache HBase; Hive; Apache Spark.	Principal Component Analysis (PCA) techniques; Silhouette scores to evaluate the quality of the clustering.	- Clustering model information would help in minimizing the infrastructure cost.	Visualization of clustering using only three more influential features.
A Novel Approach for Increased Convolutional Neural Network Performance in Gastric-cancer Classification using Endoscopic Images [39]	CIFAR-10	Simple Linear Iterative Clustering (SLIC) superpixel	MATLAB 2019a; Google's Auto Augment for augmentation; Xception Network - CNN model; Performance Metrics: Receiver Operating Characteristic (ROC) curve	The area under the ROC curve was 0.96 for augmentation and segmentation, which is 0.06 higher than that non augmentation image.	- Fully automated without the manual specification of region-of-interests for the test.	Consider the impact of the number of segmentations and the classification threshold value in segmented areas such as cancer.

DOI:10.22413/ijatest/2022/v7/i1/1

Privacy-preserving methods can be categories into the following four types based on the data lifecycle phases [40] [1].

- *Data collection: Randomization methods*

At data collection, if an untrustworthy collector adversary gathers and improperly use private data from individuals. Then randomization is used to transform the original data so as to prevent privacy disclosure. To mask some attribute values of records, this process introduces noise to the data.

- *Data publishing: k-anonymity, l-diversity, t-closeness and ϵ -differential privacy method.*

Entities may seek to publish their data publicly for further analysis. However, malicious intruders can perform cyberattacks to de-anonymize or target record owners for malevolent purpose. Privacy models are implemented to the dataset prior to the release as to effectively anonymize records. Individual records are altered in this manner so that they can no longer be uniquely identified.

- *Data distribution:* Set of secure protocols (Like oblivious transfer protocol, homomorphic encryption) for primitive operations like secure sum, set intersection, scalar product etc.

Large data sets could be partitioned and distributed either across multiple sites or by their attributes. Multiple entities may try to mine global insights in the form of aggregate statistics from all partitioned data without revealing local information to the other entities. Malicious adversaries participating in distributed computing may attempt to learn more from the shared information and, as a result, deviate from the protocols. To prevent the disclosure of local datasets in this scenario, secure multiparty protocols are used.

- *Output of the data mining: Association rule hiding and Downgrading classifier effectiveness*.

The results of data mining techniques are frequently made available via applications or interfaces. A malicious user can query these systems to learn sensitive information about the underlying data. To prevent disclosure, either the data or the application is altered in these cases by either altering mining results, such as hiding some association rules or slightly altering some classification models, or by removing some classification models entirely.

Rodríguez-Hoyos et al. [41], uses Maximum Distance to Average Vector (MDAV) additional to the k-anonymous microaggregate database to avoid privacy risks at the same time archive less computation time than the original micro aggregation mechanism.

III. CONCLUSION

Data Mining is widely by many researchers in a variety of fields in the information age. In view of extensive research, numerous extensions of the problem scope, and broad application studies, frequent pattern mining has progressed far beyond the basics. The traditional research methods comparison of applications of frequent pattern mining is presented in this paper. In addition to that, this paper fuses the comparative analysis of the recent research trends to view their benefits and limitations. Among the five major applications of pattern mining pattern-based classification and clustering are given more importance since its data pre-processing scope is of higher value. Semantic Annotation are dictionary-like annotations. Collaborative filtering systems are widely used in recommendation systems. While performing successful data mining, the ideology is to maintain data sensitivity and protect people's privacy.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their valuable comments. This work is supported by Anna Centenary Research Fellowship.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, “Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems),” 2011.

[2] U. Kazemi and R. Boostani, “FEM-DBSCAN: An Efficient Density-Based Clustering Approach,” *Iran. J. Sci. Technol. - Trans. Electr. Eng.*, pp. 1–14, Jan. 2021, doi: 10.1007/s40998-020-00396-4.

[3] Z. Liu and M. Barahona, “Graph-based data clustering via multiscale community detection,” *Appl. Netw. Sci.*, vol. 5, no. 1, pp. 1–20, Dec. 2020, doi: 10.1007/s41109-019-0248-7.

[4] Q. Liu, H. Wu, and Z. Xu, “Consensus model based on probability K-means clustering algorithm for large scale group decision making,” *Int. J. Mach. Learn. Cybern.*, pp. 1–18, Jan. 2021, doi: 10.1007/s13042-020-01258-5.

[5] A. Kaur and A. Datta, “A novel algorithm for fast and scalable subspace clustering of high-dimensional data,” *J. Big Data*, vol. 2, no. 1, p. 17, Dec. 2015, doi: 10.1186/s40537-015-0027-y.

[6] C.-H. Cheng, A. W. Fu, and Y. Zhang, “Entropy-based subspace clustering for mining numerical data,” 1999, pp. 84–93, doi: 10.1145/312129.312199.

[7] L. Parsons, E. Haque, and H. Liu, “Subspace clustering for high dimensional data,” *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 90–105, Jun. 2004, doi: 10.1145/1007730.1007731.

[8] W. Xiaohu, W. Lele, and L. Nianfeng, “An Application of Decision Tree Based on ID3,” *Phys. Procedia*, vol. 25, pp. 1017–1021, Jan. 2012, doi: 10.1016/j.phpro.2012.03.193.

[9] X. Wu *et al.*, “Top 10 algorithms in data mining,” *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.

[10] N. Z. Zacharis, “Classification and regression trees (CART) for predictive modeling in blended learning,” *IJ Intell Syst Appl*, vol. 3, 2018.

[11] F. M. Díaz-Pérez and M. Bethencourt-Cejas, “CHAID algorithm as an appropriate analytical method for tourism market segmentation,” *J. Destin. Mark. Manag.*, vol. 5, no. 3, pp. 275–282, Sep. 2016, doi: 10.1016/j.jdmm.2016.01.006.

[12] E. Kartal Koc and H. Bozdogan, “Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function,” *Mach. Learn.*, vol. 101, no. 1–3, pp. 35–58, Oct. 2015, doi: 10.1007/s10994-014-5440-5.

[13] B. Satish and K. P. Supreethi, “An independent condensed nearest neighbor classification technique for medical image retrieval,” *J. Ambient Intell. Humaniz. Comput.*, pp. 1–11, Mar. 2021, doi: 10.1007/s12652-021-03028-9.

[14] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi, and S. Shamshirband, “A new k-nearest neighbors classifier for big data based on efficient data pruning,” *Mathematics*, vol. 8, no. 2, Feb. 2020, doi: 10.3390/math8020286.

[15] A. Prasetya Wibawa *et al.*, “Naïve Bayes Classifier for Journal Quartile Classification,” *Int. J. Recent Contrib. from*

Eng. Sci. IT, vol. 7, no. 2, pp. 91–99, Jun. 2019, doi: 10.3991/ijes.v7i2.10659.

[16] A. Radl, M. Lexer, and H. Vacik, "A Bayesian Belief Network Approach to Predict Damages Caused by Disturbance Agents," *Forests*, vol. 9, no. 1, p. 15, Dec. 2017, doi: 10.3390/f9010015.

[17] N. A. Mansour, A. I. Saleh, M. Badawy, and H. A. Ali, "Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy," *J. Ambient Intell. Humaniz. Comput.*, vol. 1, p. 3, Jan. 2021, doi: 10.1007/s12652-020-02883-2.

[18] Y. Lu, Q. Yin, H. Li, H. Sun, Y. Yang, and M. Hou, "The LS-SVM algorithms for boundary value problems of high-order ordinary differential equations," *Adv. Differ. Equations*, vol. 2019, no. 1, p. 195, Dec. 2019, doi: 10.1186/s13662-019-2131-3.

[19] P. Liu, M. Xie, J. Bian, H. Li, and L. Song, "A Hybrid PSO–SVM Model Based on Safety Risk Prediction for the Design Process in Metro Station Construction," *Int. J. Environ. Res. Public Health*, vol. 17, no. 5, p. 1714, Mar. 2020, doi: 10.3390/ijerph17051714.

[20] A. Alexopoulos, G. Drakopoulos, A. Kanavos, P. Mylonas, and G. Vonitsanos, "Two-step classification with SVD preprocessing of distributed massive datasets in apache spark," *Algorithms*, vol. 13, no. 3, Mar. 2020, doi: 10.3390/a13030071.

[21] A. Cherfi, K. Nouira, and A. Ferchichi, "MC4.5 decision tree algorithm: an improved use of continuous attributes," *Int. J. Comput. Intell. Stud.*, vol. 9, no. 1/2, p. 4, 2020, doi: 10.1504/ijcistudies.2020.106485.

[22] H. Moayedi, A. Jamali, M. B. A. Gibril, L. Kok Foong, and M. Bahiraei, "Evaluation of tree-base data mining algorithms in land used/land cover mapping in a semi-arid environment through Landsat 8 OLI image; Shiraz, Iran," *Geomatics, Nat. Hazards Risk*, vol. 11, no. 1, pp. 724–741, Jan. 2020, doi: 10.1080/19475705.2020.1745902.

[23] J. Gordon and J. M. Hernández-Lobato, "Combining deep generative and discriminative models for Bayesian semi-supervised learning," *Pattern Recognit.*, vol. 100, p. 107156, Apr. 2020, doi: 10.1016/j.patcog.2019.107156.

[24] H. Basly, W. Ouarda, F. E. Sayadi, B. Ouni, and A. M. Alimi, "CNN-SVM Learning Approach Based Human Activity Recognition BT - Image and Signal Processing," in *Image and Signal Processing*, 2020, pp. 271–281.

[25] M. Zhang, L. Han, L. Yuan, and N. Chen, "Ontology-based Automatic Semantic Annotation Method for IoT Data Resources," in *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, 2020, pp. 661–667, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics50389.2020.00115.

[26] M. D. L. Calache and C. R. G. D. Farias, "Graphical and Collaborative Annotation Support for Semantic Web Services," in *Proceedings - 2020 IEEE International Conference on Software Architecture Companion, ICSA-C 2020*, Mar. 2020, pp. 210–217, doi: 10.1109/ICSA-C50368.2020.00044.

[27] H. Xia, Y. Luo, and Y. Liu, "Attention neural collaboration filtering based on GRU for recommender systems," *Complex Intell. Syst.*, vol. 1, p. 3, Jan. 2021, doi: 10.1007/s40747-021-00274-4.

[28] M. Vahidi Farashah, A. Etebarian, R. Azmi, and R. Ebrahimzadeh Dastjerdi, "A hybrid recommender system based-on link prediction for movie baskets analysis," *J. Big Data*, vol. 8, no. 1, pp. 1–24, Dec. 2021, doi: 10.1186/s40537-021-00422-0.

[29] A. Tripathi and K. Panwar, "Modified CURE algorithm with enhancement to identify number of clusters," *Int. J. Artif. Intell. Soft Comput.*, vol. 5, no. 3, p. 226, Jan. 2016, doi: 10.1504/ijaisc.2016.078517.

[30] B. Lorbeer, A. Kosareva, B. Deva, D. Softić, P. Ruppel, and A. Küpper, "Variations on the Clustering Algorithm BIRCH," *Big Data Res.*, vol. 11, pp. 44–53, Mar. 2018, doi: 10.1016/j.bdr.2017.09.002.

[31] S. Fiori, L. Del Rossi, M. Gigli, and A. Saccuti, "First Order and Second Order Learning Algorithms on the Special Orthogonal Group to Compute the SVD of Data Matrices," *Electronics*, vol. 9, no. 2, p. 334, Feb. 2020, doi: 10.3390/electronics9020334.

[32] A. Amer, "On K-means clustering-based approach for DDBSs design," *J. Big Data*, vol. 7, no. 1, p. 31, Dec. 2020, doi: 10.1186/s40537-020-00306-9.

[33] S. Wang *et al.*, "K-Means Clustering With Incomplete Data," *IEEE Access*, vol. 7, pp. 69162–69171, 2019, doi: 10.1109/ACCESS.2019.2910287.

[34] X. Hu, L. Liu, N. Qiu, D. Yang, and M. Li, "A MapReduce-based improvement algorithm for DBSCAN," *J. Algorithm. Comput. Technol.*, vol. 12, no. 1, pp. 53–61, Mar. 2018, doi: 10.1177/1748301817735665.

[35] D. Wu *et al.*, "Density-Based Top-K Spatial Textual Clusters Retrieval," *IEEE Trans. Knowl. Data Eng.*, 2021, doi: 10.1109/TKDE.2021.3049785.

[36] K. Mahata, R. Das, S. Das, and A. Sarkar, "Land Use Land Cover map segmentation using Remote Sensing: A Case study of Ajoy river watershed, India," *J. Intell. Syst.*, vol. 30, no. 1, pp. 273–286, Jan. 2021, doi: 10.1515/jisys-2019-0155.

[37] B. J. B. Fonseca, "A Clustering Algorithm to Improve the Scan Statistic in Sensor Detection Systems," *IEEE Access*, vol. 8, pp. 10373–10389, 2020, doi: 10.1109/ACCESS.2020.2965563.

[38] M. S. Halawa, R. P. D. Redondo, and A. F. Vilas, "KPIs-Based Clustering and Visualization of HPC Jobs: A Feature Reduction Approach," *IEEE Access*, vol. 9, pp. 25522–25543, 2021, doi: 10.1109/ACCESS.2021.3057427.

[39] S. Lee, H. C. Cho, and H. Cho, "A Novel Approach for Increased Convolutional Neural Network Performance in Gastric-cancer Classification using Endoscopic Images," *IEEE Access*, pp. 1–1, 2021, doi: 10.1109/ACCESS.2021.3069747.

[40] R. Mendes and J. P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications," *IEEE Access*, vol. 5, pp. 10562–10582, Jun. 2017, doi: 10.1109/ACCESS.2017.2706947.

[41] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, A. M. Mezher, J. Parra-Arnau, and J. Forné, "The Fast Maximum Distance to Average Vector (F-MDAV): An algorithm for k-anonymous microaggregation in big data," *Eng. Appl. Artif. Intell.*, vol. 90, p. 103531, Apr. 2020, doi: 10.1016/j.engappai.2020.103531.