International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(4-1), pp.234-243
DOI: http://dx.doi.org/10.21172/1.841.40
e-ISSN:2278-621X

A STUDY ON DATA MINING BEHAVIOR, CHALLENGES AND METHODS

Dhiraj Khurana(Correspondence Author)[1] & Sunita Dhingra[2]

Abstract: Information processing is having the greater significance for business organizations to take the management and financial decisions. The complete KDD process can be applied over raw input dataset to identify the analytical observations. Various feature selection methods, clustering and classification algorithms are proposed by the researchers to improve the accuracy. In this paper, a descriptive study of advance work carried out in various associated work stages of KDD process. The paper has recognized the contribution of researchers with effective observation on methodology, scope and significance of their contribution. The paper has explored the various functional stages of data processing including data cleaning, feature generation, clustering, classification and outlier detection. Each stage is here described with descriptive approach, significance, application domain and the relative scope.

Keywords: Data Mining, Classification, Segmentation, Data Cleaning, Features

1. INTRODUCTION

Data mining [1][2] is about to process the acquired information and extract the hidden and valuable aspects from it. The extracted raw data goes under various processing stages to transform the data to knowledge desired knowledge form. The impurities reduction, attribute evaluation and pattern generation is done to observe the significant information available over the dataset. This transformation is also done based on the application and objective of the designed model. After generating the normalized data form, various mining tools and techniques can be applied to build the model and to acquire the analytical derivation. Today, data mining is having the involvement in each business application, human activity and behavior to discover the knowledge patterns and to acquire the application specific decisions. For most of the applications, the single data mining method is not sufficient to retrieve the desired outcome. Each processing stage of knowledge discovery requires to go through from various algorithmic methods to qualify the work stage.

In more innovative form, the knowledge discovery process is controlled under the realization of domain requirements, industrial needs and the integrated procedural engagements. These mining processes are more directed to solve real word problem and directed towards the delivery. This kind of controlled and directed mining processed is called actionable knowledge discovery and delivery (AKD) [3]. This learning model or framework can be developed to handle the issues and requirements at micro and macro level. The data centered pattern mining framework can be conducted to generate the actionable and appropriate knowledge discovery. Various business problems, real life problems can be predicted and solved using prior interpretations at domain level, application level and the requirement level. The dynamic business process modeling can applied to avoid the uncertain situations and the financial losses. The incorporated ubiquitous intelligent method also able to under the dynamic challenges and requirement in any business model and able to facilitate the human driven interactions. This kind of unified model increases the capability of self learning and generation of new pattern in a running system. The user behavior evaluation, process driven analysis are combined by these models to take the situational decisions. The creative and imaginary thinking is the main strength and requirements of such integrated models.

1.1. Challenges

The data mining model with multiple integrated work stages suffers from various challenges targeting each stage. These challenges are at requirement level, domain understanding based, problem formulation based, data level and the process level. Some of the common data processing and knowledge discovery challenges [3][5][6] are listed here under

1.1.1. Heterogeneity

In the complex real time environment, each application domain contains different form of data such as numeric, textual, multimedia, spatial, temporal etc. In many data mining applications, it is required to process these multiple data forms in parallel to take more generic and realistic decisions. To maintain these data forms multiple database systems or storage

[1] *Department of Computer Science and Engineering, University Institute of Engineering and Technology, Maharshi Dayanand University, Rohtak (Haryana) India*

[2] *Department of Computer Science and Engineering, University Institute of Engineering and Technology, Maharshi Dayanand University, Rohtak (Haryana) India*

architectures are also required such as relational database, spatial database and multiple database systems. Based on the data requirement, the feature generation methods and the knowledge discovery methods can be applied and combined to generate the single combined decision.

1.1.2. Uncertainty

The real time connectivity of central database system to the user environment modify the database by including new information to it. It not only increases the volume of data, but can generate the new features. These kind of data adaptive uncertainty is required to measure to reduce the prediction errors. A continuous monitoring method is required to recognize the type of changes occurring in the database. Such as for the recommender system applications, it is required to monitor the change in user interest. The temporal and the spatial information aspects are also integrated with mining applications to reduce the impact of uncertainty.

1.1.3. Security

Each instance or feature database is not shared with all users. Instead, some of the information can be role specific or user specific. This kind of private information requires following up the security concerns. The mining methods can be applied to prevent the authorized access of this information. The access control measures at different applications and process stages can be applied to secure this information from illegal access. The violation of this security is also a critical challenge for data management in global environment. Different forms and angles of security measures can be applied to provide more secure information processing in open environment.

1.1.4. Response Time

Many of the data mining applications are directed connected to real time user response. In such applications, the instant data retrieval, processing and submission is done to update the database. In such applications, the algorithmic response time is one such crucial requirement so that the latest information will be processed by the other distributed users. The fraudulent transaction analysis is one such application to identify the target user and to block his all other parallel transactions. The data mining methods are today capable to handle such situation upto a larger extent. But, even though, they also suffers from the problem of scalability, memory and the distributed computing. Various inherent sequential, distributed and parallel computing methods can be applied to provide the contextual analytical decisions in minimum process time[7].

1.1.5. Memory

As the volume of the database increases, the larger memory is required to track the input-output operations and to maintain the log files. The operational growth over the system can be achieved by increasing the size of processing memory and log data. The database information is required to load in the temporary memory before processing. A larger difference exist between the overall database size and the processing memory size. Some scheduling methods and the faster page replacement methods are required to provide better utilization of limited memory[7].

1.1.6. Data Noise

The data, collected from various sources and real time environment, can have missing and incorrect values called noise. The processing of such noisy data can result the incorrect results with higher error rate. The identification of these impurities and removal of data noise is the main requirement to transform the dataset to normalized dataset. In the simpler form, the deletion can be applied to eliminate the impurities. For the intelligent rectification of data noise, various machine learning and constraint specific methods are also available. The frequency driven, association analysis and weight based methods are available to rank the dataset features and to locate the impurities or outlier over the dataset. The rectification of dataset from these impurities ensures the quality upgradation of any dataset operation[8].

In this paper, an exploration to various relative aspects to data mining is explored. The scope and importance of different data mining methods in different forms and applications are identified in this paper. The latest work contribution of earlier researchers on different application areas through different data mining methods is described in this paper. In this section, a detailed exploration of knowledge discovery requirements and scope is described. Various associated challenges to the data mining methods are also explored in this section. In section II,

2. MACHINE LEARNING

All the data mining processes are combined in the form of framework to follow up specific application objective. Machine learning [8][9][10] framework incorporate these classification, clustering, feature generation and preprocessing techniques to provide the problem solution. Each machine learning framework is focused on one or more algorithms which are supported by other data mining methods. Machine learning framework is self explanatory automated model that can process raw data to generate the predictive patterns. The machine learning models having the capability to update itself based on real time experience. At the earlier stage, the model begins with some definite constraints and structure which is later on updated based on environmental configuration, new data inflow to the system and with generated decision patterns. Figure 1 is showing the machine learning model for generic applications. The figure shows that the machine learning model has three man interconnected stages called preprocessing, learning and self evaluation stages. The model accepts the raw input and controls it under the user, domain and system constraints. The preprocessing stage basically rectifies the input data by removal of the inclusive impurities as well as transforms the dataset to featured form. The learning stage is defined to recognize the hidden pattern based on the objective requirements. In the final stage, the evaluation measures are implied to utilize these generated patterns in effective way. This process is repeated in a cyclic way and the outcome is derived. Each outcome is also taken by the system to improve the accuracy and characterization of the model.

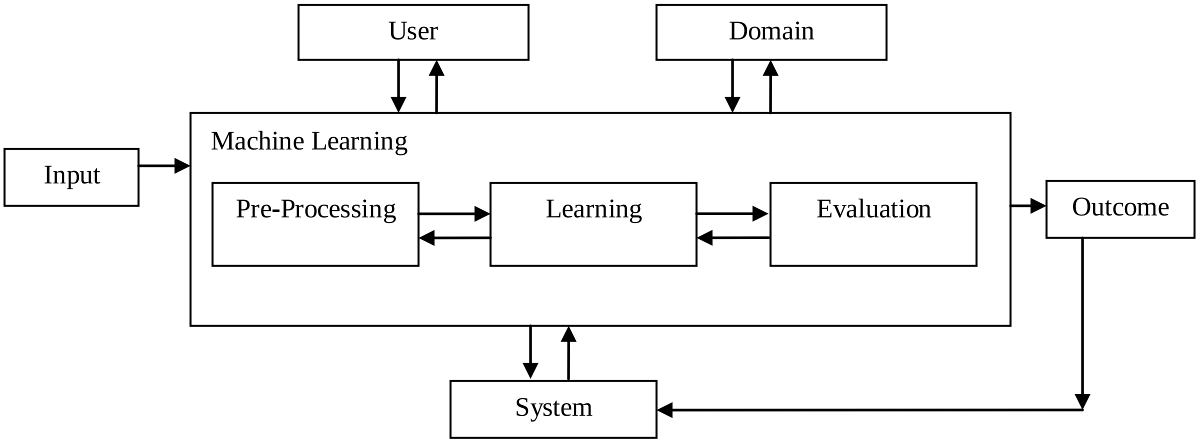

Figure 1: Machine Learning Framework

The machine learning [8][9] algorithms are able to categorize, categorize, optimize and predict the outcome for the available dataset. The algorithmic formulation can be applied as independent model, in sequential work stages or in parallel. The model can be defined under the data level and parameter specific parallelism. The structured form of this model can be defined to reduce the effort and to improve the accuracy of the real time system.

3. DATA CLEANING

The real life datasets are generally collected from distributed multiple sources and form a centralized wider database. When each distributed data source generates and maintains data independently, the structures of database management can different aspects. These dataset can be different in terms of number of attributes, attribute names and attribute types. Because of this the centralized dataset suffers from various impurities[16][17] including the missing values, data integration problem, dummy values, contradicting values, duplicate values etc. The existence of these impurities in the dataset can affect the accuracy of database system. To improve the quality of data mining and machine learning methods, it is required to clean the dataset by removing these impurities. In preprocessing stage, a wider cleaning method is required by considering the noise or impurity type. Data cleaning is itself a scientific and structured task. This structured task is composed using multiple cyclic stages called analysis, transformation and verification.

3.1. Analysis

To rectify the dataset, the first requirement is to identify the type of error exist over the dataset. To identify these error some physical, semantic and ontology specific assessment and mapping can be applied. Structured metadata specific methods can be applied to analyze the data properties and to validate the existence of these attributes, instances and terms. Based on the prior observation, the categorization of each inclusive term is done. There are number of defined methods to validate the attribute significance such as Gain Ratio, Correlation Value, Information gain method etc. Based on this significant analysis, the attributes weights can be evaluated. In same way, the instance level analysis can be obtained by generating the rules or the pattern mapper. The associatively at attribute level for each instance can be verified to recognize the scope of the instances.

3.2. Transformation/Rectification

The transformation stage is basically to standardize or normalizes the dataset by following some definite structure or schema. By considering all the partial structures, a normalized structure is formed to integrate all the valuable information. The integration stage is able to handle the attribute mismatch, duplicate attributes, conflicting attribute problem. The transformation stage removes the insignificant attributes, map the similar attributes and transform them to common data type. The transformation stage is able to resolve the problem of heterogeneity. Data level impurities or dirty data problem is also removed during this normalization stage. The normalized processing form dataset is obtained after this stage.

3.3. Verification/Validation

The normalization is a slow process in which the structure of dataset is change and some data losses occur. At one time, only single change is performed over the dataset and the verification of the change is required to maintain the dataset integrity. The verification process ensures that the transformation has not loss any sensitive and valuable information. Various error measures and accuracy analysis methods are available to validate the significance of the dataset.

Data cleaning[16][17][24] is an integral and necessary stage of machine learning algorithm to improve the accuracy, consistency and effectiveness of model. Various researchers has defined different algorithmic methods to improve the data quality. Various data cleaning methods adopted by different researchers and their significance are listed in table 1. The algorithmic methods adaptation and the scope of these methods are listed in this table.

Table 1 : Data Cleaning Methods and Significance

Author	Data Cleaning Method	Significance	Description
Al-janabi et al.[18]	Confidence Score based Weight Model	• Data Density Analysis • Repair Inconsistent Data • Cost Adaptive Data selection	• Analyze Functional Dependency between attributes • Uniform cost model is implied
Liu et al.[19]	Data Association and Repairing Model	• Reduced Cleaning Cost • Interactive Model	• Structured and Descriptive Metadata Formation • Track Data Usage • Analyze Contextual and Usage Pattern • Clean based on Association Analysis
Koshley et al.[20]	Abstract Interpretation Framework	• Performance Boosting • Rectify Integration Errors	• Similarity based Clustering • Domain Level Abstraction • Instance Cleaning
Interlandi et al. [21]	Sherlock rules and reference Table Method	• Performed Positive and Negative Cleaning • Improved Deterministic and Consistent Behavior	• Annotate Dirty instances • Apply Sherlock Rule • Combined Multiple Optimization methods
Zhang et al.[22]	Crowd sourcing using Human Intelligence Task	• Cleaning of Uncertain data • Better Truth Discovery	• Generate Crowd sourced Answers • Apply X-Partition Algorithm • Approximation algorithm for level and instance pruning
Salem et al.[23]	Functional Dependency Rule Formulation	• Performance Enhancement • Rule Updation	• Maximize Close Frequent Pattern • Apply Dependable conditional functional dependency

4. FEATURE GENERATION

The real world dataset is formed using large number of features. These features can be significant or non-contributing for a particular application. The feature selection is the prior stage of machine learning is to verify the significance of each attribute and to select the most contributing feature set. Feature selection [25] is the computational intelligent theory that benefits the data mining and machine learning methods in different aspects. The processing on selected feature not only simplifies the interpretation but also reduce the processing time and avoid the chances of over-fitting. The feature selection must be defined by considering the data understanding, storage availability, response time, and error acceptance rate. Various filters and methods are adopted by the researchers to perform feature ranking, feature selection and feature composition. Some of these feature generation methods and their scope is listed in table 2.

Table 2: Feature Generation and Selection Methods

Author	Feature Generation Method	Scope	Description
Chen et al.[26]	Multivariate Feature Scoring and Analysis	• Discover Effective Features	• Scoring based Factor Analysis • Dimension reduction
Yan et al.[27]	Group Technology	• Generate minimum feature	• Cosine Distance based Group

		groups • Superfluous information reduction	formation • Mutual Score Analysis
Eskandarian et al.[28]	Cubist based Feature Ranking	• Reasonable Error Rate • Improve prediction accuracy	• Fscaret package based feature ranking • Cyclic model assessment
Get et al.[29]	Time Series classification for feature extraction	• Multiple methods improved the accuracy • Better for practical applications	• Wavelet • Fractal • Statistical
Shi et al.[30]	PCABFS Method	• Accuracy Improvement • Removal of irrelevant features	• Wavelet Leaders Multifactral Formalism • PCA based Feature Selection
Zhao et al.[31]	Orthogonal Least Square Regression	• Improved the accuracy • Reduced dataset size	• Convergence Function with iterative method is defined • Constraint based least square method is implied
Xu et al.[32]	Weighted Multi-label linear Discriminant Analysis	• Lesser Error •	• Binary, Correlation, Entropy, Fuzzy • Dependency based Weight Assignment
Wang et al.[33]	Hybrid Association Rule Mining	• Reasonable Rules are identified • FMeasure is Increased • Implicit Features are Identified	• Five feature extraction algorithms are used in collaboration to constitute a rule • Implicit rules are extracted from basic rules
Barddal et al.[34]	Naïve Feature Drift Detection Approach	• Concept Features are generated and processed	• Data Streaming based Classification method is implied • Irrelevant and Redundant Features are Removed

5. CLUSTERING

Clustering[35][36][37] is a process to divide the available dataset in smaller segments or groups or partitions based on similarity analysis. Clustering is able to collect the similar data elements based on collaborative filtering and the similarity observation. In more effective form, the clustering can be applied on the featured data. Clustering algorithms are divided in two main categories called Partitioned clustering and Hierarchical Clustering. Paritional clustering uses the iterative method to split the data into clusters. Each data item can belong to single partition. The numbers of clusters required are known in advance. The distance measure or rule can be implied to determine the number of clusters. Hierarchical clustering is described as the clustering tree in which similarity matrix is composed between each data pair. The selected features are processed in a tree form to identify the number of clusters. Later on the algorithmic improvement on basic clustering methods were provided by the researchers. These improvements are in terms of cluster formation method, feature generation criteria or the processing on the specific application domain. In table 3, various clustering methods suggested by the researchers is provided along with algorithmic approach, feature generation method and the datasets on which method is implied.

Table 3 : Clustering Methods

Author	Dataset	Feature/Criteria/ Parameter	Algorithm/Approach
Siddiqi et al.[35]	Iris, Glass, Ecoli, Banknode Authentication, Image Segment, Cardiotocography, Student Evaluation, Landsat Satllite, Pen Based Digits, Diabetes, Heart Statlog, Ionosphere, Sonar, Vehicle, Waveform-5000	Distance Based and Frequency based	Greedy+Genetic
Kumar et al.[36]	2D CS Bigdata, 2D Non-CS Bigdata, High dimensional big data, forest cover, KDD cup 99, REDUCE Energy	Different data forms including synthetic datasets, 2D datasets and	CustiVAT algorithm

		multidimensional datasets	
Qian et al.[37]	Fitting contact lenses, Balloon, Space shuttle autolanding, Soybean small, Hayes-Roth, Lympography, Vote, Breast Cancer, Promoters	Categorical data	Space Structure based Categorical Clustering Algorithms
Shao et al.[38]	Parking slot dataset	Spatiotemporal data	Energy function with Euclidean distance analysis
Chan et al.[39]	Biological Datasets	Genome Featured Dataset	Multivariate Mutual Information and Shannon's mutual information
Singh et al.[40]	Text Documents, Spambase, Human Face	Distance Matrix Transformed Dataset	Aitchison distance based KMeans
Gullo et al.[41]	Iris, Wine, Glass, Ecoli, Yeast, Image, Abalone, Letter, KDDCup, Neuroblastoma, Leukaemia	Weighted Discriminate featured dataset	Prototype based Agglomerative Hierarchical Clustering
Azimi et al.[42]	Ames, Chariton, Calmar, Iris, Glass, Missa, Bridge, Thyroid, Magic, Wine, Shutltle, Pendigit, Wdbc, Yeast, Heart, Ionosphere, Libras, Spambase, Waveform	Karhunen-Loeve Transformed Dataset	Gradual Data Transformation based KMeans
Liu et al.[43]	Synthetic Datasets, Iris, Seeds, Glass, Ecoli, User Movement, ADL, NBA, Weather	Geometric Distance and Probability distance distribution transformed dataset	Kernel Distance and Jensen–Shannon divergence weighted parametric method
Santi et al.[44]	Synthetic Datasets	Heterogeneity and Dissimilarity Featured Dataset	Variable Neighborhood Search Heuristic Algorithm
Ferraro et al. [45]	*Gamonedo cheese, Temperature*	Fuzzy Structured Dataset	Fuzzy Probabilistic KMeans

6. CLASSIFICATION

Classification[46][47][48][49][50] is an important data mining issue to recognize the data class for some unknown data sample. In patter recognition, image processing and fault diagnosis, the classification is having the higher significance. Classification is about to categorize the objects and ideas based on understanding, recognition method and differentiation measure. Various statistical, mathematical and attribute driven frameworks are available to classify the data. The classification process is applied by the concrete implementation process called classifier. This process defines the mathematical function applied on different features to recognize the data class. The structural features, statistical features and domain specific features can be processed under classifier to recognize the data class. Decision tree, Bayesian network, Neural network are the common classifiers that can be used to identify the data class. In more advanced form, researchers have combined the two or more classifiers to enhance the accuracy and the strength of classification methods. Table 4 is showing the contribution in terms of different feature forms, classification methods and datasets adopted by different researchers. Researchers have modified the traditional classifiers to gain the accuracy and efficiency in classification frameworks.

Table 4 : Classification Methods

Author	Dataset	Processing Data Form	Classifier
Demidova et al.[46]	Medical Diagnostic Datasets : Heart, Breast Cancer; Credit Scoring; Signal Processing: Ionosphere	Lagrange function with Quadratic Programming	SVM+KNN
Guerfala et al.[48]	Iris and Wine Datasets	Logarithmic Spiral framed and Golden Ratio feature	RBF
Chen et al.[49]	KSC and Pavia Datasets for Land cover	Spatial Dominant Features and Stack Autoencoded Dataset	PCA+Deep Learning+Logistic Regression
Zhang et al.[50]	Hyperspectral Image Dataset	Entropy, Dissimilarity Agreement, Neighborhood	Ensemble Multiple Kernel Active Learning

		Estimation	
Antonelli et al.[51]	Financial Datasets : BLA, CARD, AF, ARB, COMM, SL, LEN, DPKG, BAN, GIV, COI	Condition selection based Generated Rules were used. Weight on rules applied for dominance observation	Fuzzy Rule Based Classifier and Multiobjective Evolutionary Algorithm
Xu et al.[52]	Wine, Haberman, Iris, Seed, Heart	Evidence Weight Distributed Featureset	Rule based Evidence Reasoning Classifier
Junior et al.[53]	Vector Datasets : Iris, Wine, Balance, Sonar, Credit, Image, Glass, Pima, WDBC, WPBC, Flags, Waveform, Heartspectf,Soybean, Segment,Blood, Heart, Haberman	Range interval division of attributes and transformed it as the vertex data. Information gain for attribute weight or feature	Attribute Based Decision Graph
Li et al.[55]	Abadone, Glass, Yeast, Winequality, Pima, Haberman, Vehicle, Poker	Instance Majority Analysis with Swarm Search.	Adaboost and Neural Network Classifiers
Lin et al.[56]	Breast Cancer, Person Activity, Protein Homology	Instance Selection and KNN filtered Dataset	Dual Classifier, DuC-GA and DuC-IB3
Shao et al.[57]	Yeast, Vehicle, Transfusion, Wine, PimaIndian, Ionosphere, Haberman, German, CMC, Vowel, Shuttle, Segment	Lagrangian weighted Datast	Twin SVM

7. OUTLIER DETECTION

The data collected from real environment can also have some abnormal behaviours or patterns. These inconsistent patterns or exceptional behaviour can be recognized as outlier, anomalies and abnormalities. Outlier identification is common machine learning practice to improve the performance of data discovery from dataset. Outlier detection is a decision analysis based process modeling associated to the specific domain. It can be applied to identify the fraud transaction, intrusion detection, health abnormality etc. Numerous algorithmic methods are developed by the researchers to determine and rank the outliers. Various metrics are proposed to generate the score for data validation and to set the line between the valid and invalid data. This scoring is useful to identify the outlier rank and to identify the type of outlier. Dimension specific methods are proposed by the author applied on different feature sets and with different distance measure to identify the outlier accurately. Outlier detection is also effective to improve the accuracy of classification and clustering methods. In many machine learning frameworks, the outlier detection is also included as the intermediate stage to enhance the accuracy and reliability of the system. Various improved algorithms and feature generation methods suggested by the researchers are listed in table 5 to enhance the outlier detection.

Table 5 : Outlier Detection Methods

Author	Dataset	Approach	Feature
Liu et al.[58]	Multiple High Dimensional Real World Dataset	Local Projection Score based deviation degree analysis is applied on neighbors to assign ranking	High Dimensional Data, Generic support to multiple Datasets
Rokhman et al.[59]	Categorical Data	Entropy Value Frequency based Weight Matrix was Used. Range, Variation, Deviation and Square Functions used for Weight Decision	Support Homogenous Categorical datasets
Tang et al.[60]	Synthetic and Real life datasets	Local Density based analysis on nearest neighbors is defined. The reverse nearest and shared nearest method is also defined	Theoretical Feature Analysis and Kernel Density based analysis was defined.
Huang et al.[61]	Real World Datasets	Mutual Neighbor Graph based outlier cluster detection was defined.	Distance Density, Decision Graph based Neighbor Analysis
Bai et al.[62]	Real Machine Learning Datasets of UCI	Grid Partition Algorithm on density data feature. Distributed LOF computing method was defined for parallel computation	Local Reachability distance Analysis, Density k-distance measure, Load Balancing
Bouguessa et	Mix-Attribute	Bivariate beta mixture model was defined	No Feature Transformation,

al.[63]	Dataset	on mixed attribute dataset to recognize outlier. The maximum likelihood measure used for outlier classification	Discriminate Outliers from Inliers
Liu et al.[64]	Synthetic and Real Time Datasets	Least Square Algorithm on multi-time curve fitting was applied. The probabilistic density function with distance approximation was integrated	Uncertain Data Processing, Local Density and Density Degree Analysis
Saha et al.[65]	Heterogeneous Traffic Data	PCA was used in unsupervised manner for structure adaptation. K Dimensional Vector Space processed on swarm points	LOS (Level of Service) Assessment, Correlation analysis on attributes

8. CONCLUSION

Data mining framework is the compositional process that accepts the raw dataset as input and applies a series of data processing to extract the effective knowledge from it. Each process stage is able to answer the common problems associated to different application domains. In this paper, a detailed description of various data mining methods and the advanced work carried out in each direction is described. The paper has explored the research carried out in direction of data cleaning, clustering, feature selection, classification and outlier detection methods. Various feature generation methods and the data processing methods suggested by the researchers are also described in this paper extensively.

9. REFERENCES

[1] M. Lobur, Y. Stekh and V. Artsibasov, "Challenges in knowledge discovery and data mining in datasets," Perspective Technologies and Methods in MEMS Design, Polyana, 2011, pp. 232-233.

[2] S. Singh, A. K. Solanki, N. Trivedi and M. Kumar, "Data mining challenges and knowledge discovery in real life applications," 2011 3rd International Conference on Electronics Computer Technology, Kanyakumari, 2011, pp. 279-283.

[3] L. Cao, "Domain-Driven Data Mining: Challenges and Prospects," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 6, pp. 755-769, June 2010.

[4] E. Simoudis, "Industry applications of data mining: challenges and opportunities," Proceedings 14th International Conference on Data Engineering, Orlando, FL, USA, 1998, pp. 105-.

[5] M. Kalra and N. Lal, "Data mining of heterogeneous data with research challenges," 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp. 1-6.

[6] L. R. Sebastian, S. Babu and J. J. Kizhakkethottam, "Challenges with big data mining: A review," 2015 International Conference on Soft-Computing and Networks Security (ICSNS), Coimbatore, 2015, pp. 1-4.

[7] V. Kolici, F. Xhafa, L. Barolli and A. Lala, "Scalability, Memory Issues and Challenges in Mining Large Data Sets," 2014 International Conference on Intelligent Networking and Collaborative Systems, Salerno, 2014, pp. 268-273

[8] Lina Zhou, Shimei Pan, Jianwu Wang, Athanasios V. Vasilakos, Machine learning on big data: Opportunities and challenges, Neurocomputing, Volume 237, 2017, Pages 350-361

[9] In Lee, Big data: Dimensions, evolution, impacts, and challenges, Business Horizons, Volume 60, Issue 3, 2017, Pages 293-303

[10] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody, Critical analysis of Big Data challenges and analytical methods, Journal of Business Research, Volume 70, 2017, Pages 263-286

[11] Sabeur Aridhi, Engelbert Mephu Nguifo, Big Graph Mining: Frameworks and Techniques, Big Data Research, Volume 6, 2016, Pages 1-10

[12] D. Bachlechner and T. Leimbach, "Big data challenges: Impact, potential responses and research needs," 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), Balaclava, 2016, pp. 257-264

[13] A. L'Heureux, K. Grolinger, H. F. Elyamany and M. A. M. Capretz, "Machine Learning With Big Data: Challenges and Approaches," in IEEE Access, vol. 5, no. , pp. 7776-7797, 2017.

[14] S. J. Horng, "Big data: Challenges and practical application," 2015 International Conference on Science in Information Technology (ICSITech), Yogyakarta, 2015, pp. 11-13

[15] Y. Ma, Y. Tan, C. Zhang and Y. Mao, "A data mining model of knowledge discovery based on the deep learning," 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), Auckland, 2015, pp. 1212-1216

[16] Jason Van Hulse, Taghi Khoshgoftaar, Knowledge discovery from imbalanced and noisy data, Data & Knowledge Engineering, Volume 68, Issue 12, 2009, Pages 1513-1542

[17] S. Swapna, P. Niranjan, B. Srinivas and R. Swapna, "Data cleaning for data quality," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 344-348.

[18] S. Al-janabi and R. Janicki, "A density-based data cleaning approach for deduplication with data consistency and accuracy," 2016 SAI Computing Conference (SAI), London, 2016, pp. 492-501.

[19] H. Liu, A. K. Tk, J. P. Thomas and X. Hou, "Cleaning Framework for BigData: An Interactive Approach for Data Cleaning," 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, 2016, pp. 174-181.

[20] D. K. Koshley and R. Halder, "Data cleaning: An abstraction-based approach," 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, 2015, pp. 713-719.

[21] M. Interlandi and N. Tang, "Proof positive and negative in data cleaning," 2015 IEEE 31st International Conference on Data Engineering, Seoul, 2015, pp. 18-29.

[22] C. J. Zhang, L. Chen, Y. Tong and Z. Liu, "Cleaning uncertain data with a noisy crowd," 2015 IEEE 31st International Conference on Data Engineering, Seoul, 2015, pp. 6-17.

[23] Rashed Salem, Asmaa Abdo, Fixing rules for data cleaning based on conditional functional dependency, Future Computing and Informatics Journal, Volume 1, Issue 1, 2016, Pages 10-26

[24] Er-Wei Bai, Hans Johnson, Weiyu Xu, Mathews Jacob, A Preliminary Study on Cleaning up Erroneous Data and Filling in Missing Values in A Medical Record, IFAC-PapersOnLine, Volume 48, Issue 20, 2015, Pages 493-498

[25] Aparna U. R. and S. Paul, "Feature selection and extraction in data mining," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, 2016, pp. 1-3.

[26] Y. Chen and Hong Cui, "Intelligent feature extraction and knowledge mining by multivariate analyses," 2009 IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, 2009, pp. 33-39.

[27] J. Yan and W. Li, "Group Technology Based Feature Extraction Methodology for Data Mining," 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Shandong, 2008, pp. 235-239.

[28] Sajad Eskandarian, Peyman Bahrami, Pezhman Kazemi, A comprehensive data mining approach to estimate the rate of penetration: Application of neural network, rule based models and feature ranking, Journal of Petroleum Science and Engineering, Volume 156, 2017, Pages 605-615

[29] Li Ge, Li-Juan Ge, Feature extraction of time series classification based on multi-method integration, Optik - International Journal for Light and Electron Optics, Volume 127, Issue 23, 2016, Pages 11070-11074

[30] Hongtao Shi, Hongping Li, Dan Zhang, Chaqiu Cheng, Wei Wu, Efficient and robust feature extraction and selection for traffic classification, Computer Networks, Volume 119, 2017, Pages 1-16

[31] Haifeng Zhao, Zheng Wang, Feiping Nie, Orthogonal least squares regression for feature extraction, Neurocomputing, Volume 216, 2016, Pages 200-207

[32] Jianhua Xu, A weighted linear discriminant analysis framework for multi-label feature extraction, Neurocomputing, 2017

[33] Wei Wang, Hua Xu, Wei Wan, Implicit feature identification via hybrid association rule mining, Expert Systems with Applications, Volume 40, Issue 9, 2013, Pages 3518-3531

[34] Jean Paul Barddal, Heitor Murilo Gomes, Fabrício Enembreck, Bernhard Pfahringer, A survey on feature drift adaptation: Definition, benchmark, challenges and future directions, Journal of Systems and Software, Volume 127, 2017, Pages 278-294

[35] U. F. Siddiqi and S. M. Sait, "A New Heuristic for the Data Clustering Problem," in IEEE Access, vol. 5, no. , pp. 6801-6812, 2017.

[36] D. Kumar, J. C. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie and T. C. Havens, "A Hybrid Approach to Clustering in Big Data," in IEEE Transactions on Cybernetics, vol. 46, no. 10, pp. 2372-2385, Oct. 2016.

[37] Y. Qian, F. Li, J. Liang, B. Liu and C. Dang, "Space Structure and Clustering of Categorical Data," in IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 10, pp. 2047-2059, Oct. 2016.

[38] W. Shao, F. D. Salim, A. Song and A. Bouguettaya, "Clustering Big Spatiotemporal-Interval Data," in IEEE Transactions on Big Data, vol. 2, no. 3, pp. 190-203, Sept. 1 2016.

[39] C. Chan, A. Al-Bashabsheh, Q. Zhou, T. Kaced and T. Liu, "Info-Clustering: A Mathematical Theory for Data Clustering," in IEEE Transactions on Molecular, Biological and Multi-Scale Communications, vol. 2, no. 1, pp. 64-91, June 2016.

[40] J. P. Singh and N. Bouguila, "Proportional data clustering using K-means algorithm: A comparison of different distances," 2017 IEEE International Conference on Industrial Technology (ICIT), Toronto, ON, 2017, pp. 1048-1052.

[41] Francesco Gullo, Giovanni Ponti, Andrea Tagarelli, Sergio Greco, An information-theoretic approach to hierarchical clustering of uncertain data, Information Sciences, Volume 402, 2017, Pages 199-215

[42] Rasool Azimi, Mohadeseh Ghayekhloo, Mahmoud Ghofrani, Hedieh Sajedi, A novel clustering algorithm based on data transformation approaches, Expert Systems with Applications, Volume 76, 2017, Pages 59-70

[43] Han Liu, Xianchao Zhang, Xiaotong Zhang, Yi Cui, Self-adapted mixture distance measure for clustering uncertain data, Knowledge-Based Systems, Volume 126, 2017, Pages 33-47

[44] Éverton Santi, Daniel Aloise, Simon J. Blanchard, A model for clustering data from heterogeneous dissimilarities, European Journal of Operational Research, Volume 253, Issue 3, 2016, Pages 659-672

[45] Maria Brigida Ferraro, Paolo Giordani, Possibilistic and fuzzy clustering methods for robust analysis of non-precise data, International Journal of Approximate Reasoning, Volume 88, 2017, Pages 23-38

[46] L. Demidova and Y. Sokolova, "A novel SVM-kNN technique for data classification," 2017 6th Mediterranean Conference on Embedded Computing (MECO), Bar, Montenegro, 2017, pp. 1-4

[47] G. Tuysuzoglu and Y. Yaslan, "Biomedical data classification using supervised classifiers and ensemble based dictionaries," 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, 2017, pp. 1-4.

[48] M. W. Guerfala, A. Sifaoui and A. Abdelkrim, "Data classification using logarithmic spiral method based on RBF classifiers," 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), Hammamet, 2016, pp. 416-421.

[49] Y. Chen, Z. Lin, X. Zhao, G. Wang and Y. Gu, "Deep Learning-Based Classification of Hyperspectral Data," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 6, pp. 2094-2107, June 2014

[50] Y. Zhang, H. L. Yang, S. Prasad, E. Pasolli, J. Jung and M. Crawford, "Ensemble Multiple Kernel Active Learning For Classification of Multisource Remote Sensing Data," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 2, pp. 845-858, Feb. 2015

[51] M. Antonelli, D. Bernardo, H. Hagras and F. Marcelloni, "Multiobjective Evolutionary Optimization of Type-2 Fuzzy Rule-Based Systems for Financial Data Classification," in IEEE Transactions on Fuzzy Systems, vol. 25, no. 2, pp. 249-264, April 2017

[52] Xiaobin Xu, Jin Zheng, Jian-bo Yang, Dong-ling Xu, Yu-wang Chen, Data classification using evidence reasoning rule, Knowledge-Based Systems, Volume 116, 2017, Pages 144-151

[53] João Roberto Bertini, Maria do Carmo Nicoletti, Liang Zhao, Attribute-based Decision Graphs: A framework for multiclass data classification, Neural Networks, Volume 85, 2017, Pages 69-84

[54] Raluca-Mariana Stefan, A Comparison of Data Classification Methods, Procedia Economics and Finance, Volume 3, 2012, Pages 420-425

[55] Jinyan Li, Simon Fong, Raymond K. Wong, Victor W. Chu, Adaptive multi-objective swarm fusion for imbalanced data classification, Information Fusion, Volume 39, 2018, Pages 1-24

[56] Wei-Chao Lin, Chih-Fong Tsai, Shih-Wen Ke, Mon-Loon You, On learning dual classifiers for better data classification, Applied Soft Computing, Volume 37, 2015, Pages 296-302

[57] Yuan-Hai Shao, Wei-Jie Chen, Jing-Jing Zhang, Zhen Wang, Nai-Yang Deng, An efficient weighted Lagrangian twin support vector machine for imbalanced data classification, Pattern Recognition, Volume 47, Issue 9, 2014, Pages 3158-3167

[58] H. Liu; X. Li; J. Li; S. Zhang, "Efficient Outlier Detection for High-Dimensional Data," in IEEE Transactions on Systems, Man, and Cybernetics: Systems , vol.PP, no.99, pp.1-11

[59] N. Rokhman, Subanar and E. Winarko, "WMEVF: An outlier detection methods for categorical data," 2016 International Conference on Informatics and Computing (ICIC), Mataram, 2016, pp. 37-42.

[60] Bo Tang, Haibo He, A local density-based approach for outlier detection, Neurocomputing, Volume 241, 2017, Pages 171-180

[61] Jinlong Huang, Qingsheng Zhu, Lijun Yang, DongDong Cheng, Quanwang Wu, A novel outlier cluster detection algorithm without top-n parameter, Knowledge-Based Systems, Volume 121, 2017, Pages 32-40

[62] Mei Bai, Xite Wang, Junchang Xin, Guoren Wang, An efficient algorithm for distributed density-based outlier detection on big data, Neurocomputing, Volume 181, 2016, Pages 19-28

[63] Mohamed Bouguessa, A practical outlier detection approach for mixed-attribute data, Expert Systems with Applications, Volume 42, Issue 22, 2015, Pages 8637-8649

[64] Jing Liu, HuiFang Deng, Outlier detection on uncertain data based on local information, Knowledge-Based Systems, Volume 51, 2013, Pages 60-71

[65] Pritam Saha, Nabanita Roy, Deotima Mukherjee, Ashoke Kumar Sarkar, Application of Principal Component Analysis for Outlier Detection in Heterogeneous Traffic Data, Procedia Computer Science, Volume 83, 2016, Pages 107-114