International Journal of Advanced Science and Technology
Vol.113 (2018), pp.133-142
http://dx.doi.org/10.14257/ijast.2018.113.14

A Survey of Data Clustering Methods

Saima Bano and M. N. A. Khan

Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan
sayohunzai@gmail.com, mnak2010@gmail.com

Abstract

Data clustering is one of the most essential, common and interesting task to classification of patterns in different areas such as data mining, pattern recognition, artificial intelligence and etc. The objective of data clustering is to classification of similar entities. There are so many different techniques of data clustering available for different nature of applications. Data clustering techniques are categorizing into two types – Partitioning Procedures and Hierarchical Procedures. Hierarchical clustering creates hierarchy of clusters, look like tree. Results of hierarchical Clusters are shown in dendrogram shape. Partitioning method-clustering makes various partitions of objects and evaluates them by some standard. In this paper, we introduce a critical review on few papers and found some strengths and weaknesses of different clustering techniques. The purpose of this overview is to compare and evaluate each clustering techniques and find their pros and cons. This comparison concludes the better approach for future research in data clustering.

Keywords: *Data Clustering, Fuzzy Clustering, Fuzzy C-Means Clustering, K-Means, K-Mediods*

1. Introduction

Data clustering is a procedure in which we make cluster of entities ton based on their similar features. A good clustering technique will create high quality clusters with high intra-class similarity low inter-class similarity Quality of clustering depends on the similarity measure used and its implementation. Quality of a clustering process is measured by its ability to find out some or all of the unknown patterns. To creating a similarity clusters distance measured is used, which are as follow: Euclidean distance, Manhattan or taxicab distance, Mahalanobis distance, Inner product space, Hamming distance, Cosine Similarity Index, Minkowski distance. All distances are used to finds similarity in between different points but mostly Euclidian distance is used to measure objects for similarity attributes. Data Clustering is used in many different application areas; *i.e.*, Data retrieval, Image analysis, Machine learning, web search engines, Pattern recognition, computational, economic, Libraries, insurances, city planning, and earthquakes studies.

There are so many data clustering algorithms to classified data into similar groups. Normally clustering algorithms are categorized into two groups as unsupervised linear clustering algorithm and unsupervised non-linear clustering algorithm. In unsupervised linear clustering algorithms includes: Fuzzy c mean clustering algorithm, Quality threshold clustering algorithm, K- mean clustering algorithm, hierarchical clustering algorithm, and Gaussian (EM) clustering algorithm. Whereas MST created clustering algorithm are density based and kernel k- mean are clustering algorithm include in unsupervised nonlinear.

Received (December 27, 2017), Review Result (March 11, 2018), Accepted (March 13, 2018)

ISSN: 2005-4238 IJAST

Some well-know clustering algorithms are K-means clustering which create a clusters n number of objects into k clusters. Moreover, observed the nearest means to making clusters. Assigning objects to cluster by using distance. k-means strength is fast, cover local optimum and very flexible clusters during process its changes. K-means limitation is when data size greater the clusters results poor because k-means always cover local optimum clusters. While k-mediods clustering algorithms are like k-means algorithm which n objects point into k clusters, it minimizes the dissimilarities so it is more robust than k-means. Mediod is center point in k-mediod. It takes reference points instead of mean values of objects.

Common clustering problems are Interpreting results, Outlier handling, Number of clusters, Dynamic data and evaluating results, to reduce such type of issues different clustering Procedures are used. Clustering Procedures are been classified into the following categories:

- Hierarchical Procedures
    - Agglomerative hierarchical clustering
    - Divisive hierarchical clustering
- Partitioning Procedures
- Density-based Procedures
- Model-based Clustering Procedures
    - Decision Trees
    - Neural Networks
- Grid-based Procedures
- Fuzzy Clustering

1.1. Hierarchical Clustering Method

This method create a group of nested clusters structured as a visualized and hierarchical tree as a dendrogram – a diagram like tree that records the structures of joins or separations. Hierarchical procedures can be any agglomerative or divisive. Agglomerative algorithm starts with each element as a single cluster and joins them in sequence larger clusters; divisive algorithms start with the entire group and proceed to split it into successively minor clusters.

1.2. Partitioning Clustering Method

The partitioning Procedures commonly result in a group of M clusters, each item belonging to unique cluster. Each cluster may be denoted by a centroid or a cluster representative; this is some sort of summary description of all the entities enclosed in a cluster. The exact form of this report will depend on the nature of the entity that is being clustered.

1.3. Density-based Clustering Method

The density –based clustering algorithm discovers the clusters in arbitrary shape. By region of low density objects are separate into dense region. points are connecting on the base of certain distance threshold. Points are connected till satisfy density criteria. It handles noise and one scan. Several algorithm related to density –based is: DBSCAN, OPTICS, DENCLUE, CLIQUE, BIRCH, CURE.

1.4. Model-based Clustering Methods

Model-based methods are used to optimize the data and some mathematical model that is based on the assumption. Model-based clustering, model generated a data that assumes

and recover the original model from the data. Then clusters are defined from the recovered data.

1.5. Grid-based Clustering Method

This method concerned with value space instead of data points to making grid clustering. In such way grid clustering method first creates a grid structure then calculates cell density and after that identifies cluster centers. The main advantage of grid-based clustering is reducing a computational complexity, particularly for clustering very large data sets. Several interesting methods of grid-based clustering method are: Wave Cluster, STING, CLIQUE.

1.6. Fuzzy Clustering

Fuzzy clustering is referred to soft computing and the data points belong to more than one cluster. A famous fuzzy clustering algorithm is fuzzy c- means clusters it are mostly used in image processing tools.

2. Related Work

Karaboga *et al.*, [1] discuses Artificial Bee Colony (ABC) algorithm an optimizing procedure that simulates the quick seeking behaviors of a honeybee swarm for data classification and clustering. The artificial colony consists of three types of bees: employed bees, onlookers and scouts. Employed bee: associate with a specific food source and offers the neighborhoods of the basis in its memory. Onlookers: it's get information of food source from the employed bees in the hive and select one of the food source together the nectars. Scouts: it is responsible for finding new food, the nectar source. The total numbers of employed bees are equal to the number of food sources around the hive. In this paper, clustering problem is stated *i.e.*, N number of objects are provided and the aim of the proposed technique is to allocate each object into k-clusters followed by minimizing the sum of distances between the objects. ABC algorithm is compared with other famous heuristic algorithm such as GA and PSO based on their performance. Thirteen types of classification issues (such as glass, thyroid, and wine) from Glass Identification Data Set, Thyroid Disease Data Set, and Wine Data Set available in the UCI Machine Learning repository are used to evaluate the performance of the ABC algorithm. First 75% of data used for training data, and the remaining 25% is used for testing data. The performance is evaluated and tested using XOR, Decoder-Encoder and 3-bit parity Procedures. The authors tested the proposed algorithm to form clusters for patterns by assigning it to the class whose center is closest to the cluster center. For this purpose, they have used the Euclidean distances measure. The ABC algorithm shaped average clusters for all the problems with error percentages of 13.13% as compared to 15.99% error percentage for PSO. ABC algorithm offers much better quality clusters and shows good performance. The overall ranking of the ABC algorithm was first among the other optimization techniques.

Chen *et al.*, [2] disuses Spectral parallel Procedures which is widely used for computer vision and information retrieval. Author compare two types of approaches: sparsifying the similarity matrix and the Nystro¨m approximation to discovering a parallel spectral cluster in distributed environment. Spectral refer to the use of eigenvalues, eigenvectors, singular values and singular vectors. It is more effective in finding clusters than some traditional algorithms, such as k-means. The algorithm constructs a parallel matrix and reflects the relationship between the data points, then uses similarity matrix information to groups into k-clusters. To constructing the sparse similarity matrix using nearest neighbor the authors use compute distances of

all data points, symmetrically modify the sparse matrix and finally computes the similarities. These three steps are implemented by using MapReduce, a Google parallel computing framework. To reduce the memory use the sparsification approach keeps the most useful sparse matrix. Whereas, Nystrom approximation approach used to stores only several columns of the similarity matrix. For experimentations, three data sets were used: Corel (images), RCV1 (documents) and Picasa-Web (a Google online platform photo sharing product. By k-means clusters were generated an observed that spectral clustering finds better similarities in images. By evaluation parallel spectral clustering approach speed up to 256 machines, and efficiently handle larger problem.

According to Senthilnath *et al.*, [3] Firefly Algorithm (FA) is best for optimization problem in clustering that is a latest nature inspired optimization algorithm which simulates the flash pattern and characteristics of fireflies. FA is used for difficult optimization problem in clustering. The algorithm works in following three steps. The agents are randomly distributed in search space. The objects were separate into classes which goal is to find clusters center. FA is unsupervised technique so the dataset is distributed into training dataset and test dataset. Thirteen types of data sets (The Balance data set, Cancer-Int data set and *etc*.,) from UCI machine learning repository were used to shown the results of the optimization techniques. The performance is measured by using Classification Error Percentage (CEP) with all the 13-benchmark data sets. CEP helps to evaluate which method has generated the optimal cluster centers. The performance of FA is compared with other two well-known optimization algorithms i.e. Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO). The accuracy and robustness of FA could be efficiently used for clustering and FA has good global performance than other optimization techniques. The performance of clustering generally depends on the size and value of training data set. FA efficiently generate clusters center.

Kim *et al.*, [4] presented Density-based clustering algorithms such as DBSCAN and OPTICS are widely used in clustering. Density-based clustering Procedures are used to discover clusters of arbitrary shape and dense regions of data points. On the base of given density parameters, discover clusters which are dense in region. For a large dataset author proposed a density-based clustering algorithm, which discovers densities of cluster and well suited for framework .by using MapReduce framework it is difficult to parallelize clustering. When dealing with large amount of data, it is hard to parallelize clustering algorithm by using MapReduce framework. A Density-based clustering algorithm, DBCURE find clusters with varying densities and is suitable for parallelizing the algorithm with MapReduce. To parallelized DBCURE using MapReduce, the authors have developed DBCURE-MR which finds several clusters together by expanding every core point in parallel. Whereas, traditional density-based algorithms find each cluster individually. Three datasets CLOVER, WINDOW and BUTTERFLY have been used by the authors to evaluate clustering. The overall findings of the study are that DBCURE and DBCURE-MR finds clusters centers efficiently and scales up well with the MapReduce framework.

Kaymak *et al.*, [5] used Fuzzy clustering algorithm to divides the dataset into groups so that the clusters describe a structure within the data. Similar to fuzzy logic, in fuzzy clustering every point has certain degree of belonging to different clusters, rather than belonging to just one cluster. Fuzzy clustering is widely used in various fields like finance and marketing. However, there are certain issues in fuzzy algorithm that need to be taken care of such as volume and shape of the clusters, distribution of the data patterns, initialization of the clustering algorithm and selecting the number of clusters in the data.

Marghescu *et al.*, [6] explore advanced supporting tools for predicting currency crises, which is based on an experimental study of the currency crisis in 23 rising markets around the world spanned over half century. The authors built fuzzy C-mean (FCM) model to partition data points into specific overlapping groups and then classify data clusters into early-warning clusters (EWCs) and tranquil clusters (TCs). FCM model is used for predicting the overall economic crisis by testing and evaluating a large number of samples.

Niknam *et al.*, [7] discuses K-Mean clustering which is simple and efficient technique to create k-clusters, and covers local optimal solutions. K-mean cluster technique is highly depend on the initial position and finds local optimal solution. Niknam and Amiri (2010) present a new hybrid evolutionary approach based on FAPSO (fuzzy adaptive particle swarm optimization), ACO (ant colony optimization) and k-means algorithms called FAPSO-ACO–K which finds global optimal and correctly centered clusters.

Kocheturov *et al.*, [8] analyze stock markets of the USA and Sweden by studying the dynamics of a cluster structure in financial markets followed by finding its correlation to crisis and non-crisis periods. The network examination has become a great tool for learning financial markets in the last 15 years. The authors build a network structures from a correlation matrix of the stocks markets, which has a predefined number of connected components. Structure is a forest of stars with weighted edges where every node represents a stock and weights are equal to the measured similarity between the stocks.

Nanda *et al.*, [9] presented a data mining methodology for classifying Indian stocks market into unlike clusters. The clustering methodology classifies stocks on certain investment criteria. Bombay Stock Exchange (BSE) was together from Capitaline Databases Plus and data for 106 stocks were collected for the fiscal year 2007–2008. K-means, self-organizing maps (SOM) and Fuzzy C-means clustering approaches were used to cluster stock market data. The authors collected a mixed data from different sectors like BSE BANKEX, BSE Auto, BSE Pharma, BSE IT, BSE Midcaps and BSE. By comparing the portfolio performances with the BSE Sensex benchmark index, the results shows that K-means method turns out to be better.

Sastry *et al.*, [10] utilized clustering techniques for detecting difference in product sales and also to identify and compare sales over a particular time. Clustering is well suited to group items that seem to fall naturally together, when there is no specified class for any new item. Authors used annual sales data of steel products to analyze Sales Volume & Value with respect to dependent attributes like products, customers and quantities sold. The demand for steel products is cyclical and depends on many factors like customer profile, price, discounts and tax issues. Authors have analyzed sales data with clustering algorithms like K-Means & EM (expectation–maximization) that revealed many interesting patterns useful for improving sales revenue and achieving higher sales volumes. K-Means & EM (partition Procedures) algorithms are better suited to evaluate sales data in comparison with density based Procedures.

Fallahpour *et al.*, [11] in their study talk about applied clustering approach to classify 79 selected stocks of Iran's stock market into a number of clusters. The data collected from currency crisis in Iran’s economy that negatively influenced Iran Stock Exchange dramatically during the period 22/09/2012 to 22/03/2013. Applied three well know clustering Procedures namely K-medoids, K-means and X-means were used. The techniques were evaluated by the application of Intraclass inertia, which show the density of each clustering method. By comparison, of Intra-class inertia was take that K-Means algorithm has a enhanced quality than K-medoids and X-means techniques. By the use of some defined indexes namely Silhouette and

Davis-Bouldin, efficient number of clusters were extracted. Most desirable clusters from five stock market result shows that k-means can create an efficient portfolios.

Stetco *et al.*, [12] performed classification of registered Companies in London Stock Exchange to indentify the group of similarly performing companies based on their historical stock price record. Fuzzy clustering analysis were carried out using a correlation-based metric to obtain a more insightful classification of the companies into groups with fuzzy boundaries, giving realistic and detailed view of their relationships. Analyze cluster and discovered groups in terms of the volatility of their returns using both standard deviation and exponentially weighted moving average. This approach has the potential to be of practical importance to classification as it can detect fuzzy clusters of correlated stocks that have lower inter-cluster correlation, analyze their volatility and sample potentially less risky combination of assets.

Miguéis *et al.*, [13] analyze lifestyle segmentation of customer using data mining technique. A decent relationship between customers and companies is a vital element of attractiveness and organization of loyalty relationships with customers is a focal tactical purpose. Therefore, companies are improving service levels and wishing to be at the top edge in order to certify a good business relationship with customers. The authors [13] extracted information from a large transaction database and propose a market segmentation technique for retailing based on customer's lifestyle. The authors used a variable clustering method to infer customers' lifestyle. The authors [13] identify typical shopping baskets based on products which are more repeatedly bought together. Lifestyle segment is assigned to customers based on their purchasing history. The propose model is implemented in European retailing company.

Shim *et al.*, [14] develop CRM approach using association rules and sequential patterns for a small size online shopping mall. In 2002, dot-com bubble burst was established and contained several small-sized online shopping centers. Many of them become known in the market because they have good relations with customers and decent characteristics of online marketplace including significantly reduced menu cost for products/services and search costs and easily access services/products in the world. But some of the online shopping malls have not constantly succeeded and need to close them. Several of them have poor customer relationship management strategies and need to close them. The authors [14] analyze customer transactions data of the online shopping mall and propose sequential patterns and association rules. Firstly, authors [14] define the VIP customers based on recency, frequency and monetary (RFM) values. Formerly, they design a model which categories customers into non-VIP and VIP. The authors [14] used various data mining techniques such as logistic regression, bagging decision tree, artificial neural network and decision tree with each of these as a base classifier. Lastly, they identify patterns and rules for VIPs customer from the transactional data using association rules and sequential patterns and then design CRM strategies for the online shopping mall. Reference [16-25] outlined various software engineering and machine learning techniques in different domains.

3. Critical Evaluation

Comparative evaluation of different clustering techniques as discussed above is shown in Table 1.

Table 1. Critical Evaluation of Clustering Methods

Ref #	Proposed technique	Strengths	Limitations	Possible Improvement
[1]	Artificial Bee Colony (ABC) algorithm	ABC algorithm, fast, robust, enhances accuracy and used for optimizing multivariable functions.	Local search performance depends on neighborhood search and greedy selection and the global search performance of algorithm depends on random search process.	To establish true strengths of ABC algorithm, it would be better to compare it with some other optimization techniques such as Genetic Algorithms and Ant Colony Optimization.
[2]	Firefly Algorithm (FA)	The accuracy and robustness of FA is efficient for clustering center and performance is better than other optimization techniques.	FA depends on population size and This will affect the computation time of this algorithm.	To get better results FA would be compared with some other optimization techniques like ant colony optimization, genetic algorithms etc.
[3]	Parallel clustering algorithm called DBCURE-MR.	The performance of DBCURE-MR speedup with an order of magnitude. And it is more effective for experimenting with real-life datasets.	Time complexity is maximum.	Because lack of availability of large memory, the buffer size is only kept to 500MB in every machine for simulation. To get a better result, the memory size on each machine be enlarged.
[4]	Fuzzy c-means (E-FCM) extension and – Kessel (E-GK) algorithms	Fuzzy C-Mean algorithm is to reduce the sensitivity of the resulting clustering	Algorithm miss-classification of patterns due to unsupervised learning.	Real world applications of extended clustering algorithms need to be explores to truly analyze properties of the algorithms.
[5]	Fuzzy C- mean (FCM) model	FCM model is used for predicting the overall economic crisis by testing and evaluating a large number of samples.	The high rate of false alarms in the test sets is a major weakness of this model.	Currently, the model is tested for only three Asian countries — Indonesia, Korea and Taiwan. To get a better estimate of the results, data about currency crisis in different countries from other continents

				could be included.
[6]	FAPSO (fuzzy adaptive particle swarm optimization), ACO (ant colony optimization) and k-means algorithms called FAPSO-ACO–K	To find a better cluster partition and solve nonlinear partitioned clustering problem.	The algorithm still finds hard optimization problem.	To get better understanding about its potential use, the proposed algorithms could be compared with other evolutionary algorithm such as artificial bee colony and firefly algorithm etc.
[7]	Dynamic cluster structures	Dynamic cluster structures proved to be more stable during the crisis periods like the world financial crisis, the Subprime mortgage crisis, the Dot-com crisis and the banking crisis in Sweden The modularity-based approach reveals that the highest values of modularity and relatively small number of communities help detect the abrupt changes in the markets such as the Dot-com crisis is related to both types of the markets. The world financial crisis is not observable from these trends so well — as there are only local quality threshold (Qt) maxima and minima of the community number. Thus the main difference between the modularity and our measures is that the former gives more contrasting picture of the crises.	Traditionally, dynamic clustering algorithms are calculated for large datasets. It is difficult to cluster the tiny dataset due of the loss of the statistical characteristics and probability features.	Author studied the Swedish financial market with 266 companies. To overcome the world financial crisis PMP should be applied more companies and should test the market performance with other countries.
[8]	Clustering-based stock selection method	The proposed cluster-based approach considerably reduces the time required for creating a portfolio. Validity indexes were used in each case to find the optimal number of clusters.	Clustering algorithms were performing on limited dataset. The validity indexes were no reliable in some cases.	To get more optimal number of clusters in stock market data, it would be better to analysis it with some other clustering techniques.
[9]	K-Means and EM (expectation–maximization)	Reduces the risk of redundancy errors and allows global integration of product sales.	Systems are restrictive and not flexible in implementation and usage. ERPs are difficult to adjust to the specific workflow and the main causes of their failure. Systems can be difficult to use.	To get better analysis of sale data and resultant revenue collection needs to be tested on other software such as SAP.
[10]	K-medoids, K-means and X-means	Markowitz model creates optimize portfolios on applied best techniques and k-means meets the portfolios	By applied method, k-means is not guaranteed to be global optimum. Markowitz model is based	Other well-known clustering Procedures like

		risk minimization by portfolio diversification.	on diversification; when portfolios are not well diversified then k-means is not able to return efficient stock portfolios.	fuzzy clustering, expectation–maximization etc. can be used to study how to minimize currency crisis in the stock exchange.
[11]	Principal Component Analysis (PCA), compression and fuzzy clustering	The authors computed 43 eigenvectors with 90% of the variability in stock data. By the using PCA, it removes noise and to improves data analysis performance.	1. If the time scale is kept small (such as hourly, daily) then not only does it becomes computationally more costly to analyze but also global trends might be difficult to discover. In contrast, larger time intervals may erase local differences that may be useful for classification. Stock price is not a good indicator of performance, as it does not reflect the size and revenue of a company; neither is it a good comparator when considering multiple stocks. 2.In Fuzzy clustering, data points have more than single clusters so correlation cannot be used directly as a metric in cluster analysis as it does not satisfy the non-negativity condition of metric functions.	To identify better stock prices for long period, the model-based techniques can be used.

4. Conclusion and Future Work

A review of different clustering techniques proposed in the literature shows that each technique has its own advantages and disadvantages. I would like to propose a new hybrid technique like evolutionary techniques, optimization methods such as Genetic Algorithms and Ant Colony Optimization to compare with other techniques to get better result of data clustering. Real world applications of extended clustering algorithms need to be explored to analyze properties of the algorithms. The available existing techniques regarding data clustering will overcome the issues regarding performance, scalability and dimensionality. My focus will be to improve data clustering technique to incorporate performance, scalability issues. The key conclusion of this paper is to assessments of data clustering techniques, which are used in data mining, pattern. Reason behind the review is to implement a fresh hybrid technique in an operational method keeping in view of different data clustering strengths and limitations based on their practicality and productivity of current techniques are also analyzed critically.

References

[1] D. Karaboga and C. Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", Applied Soft Computing, vol. 11, no. 1, **(2011)**, pp. 652-657.
[2] W. Y. Chen, Y. Song, H. Bai, C. J. Lin and E. Y. Chang, "Parallel spectral clustering in distributed systems", Pattern Analysis and Machine Intelligence, IEEE Transactions, vol. 33 no. 3, **(2011)**, pp. 568-586.

[3] J. Senthilnath, S. N. Omkar and V. Mani, "Clustering using firefly algorithm: performance study", Swarm and Evolutionary Computation, vol. 1, no. 3, **(2011)**, pp. 164-171.
[4] Y. Kim, K. Shim, M. S. Kim and S. Lee, "DBCURE-MR: an efficient density-based clustering algorithm for large data using MapReduce", Information Systems, vol. 42, **(2014)**, pp. 15-35.
[5] U. Kaymak and M. Setnes, "Extended fuzzy clustering algorithms", ERIM Report Series Reference No. ERS-2001-51-LIS, **(2000)**.
[6] D. Marghescu, P. Sarlin and S. Liu, "Early-warning analysis for currency crises in emerging markets: A revisit with fuzzy clustering", Intelligent Systems in Accounting, Finance and Management, vol. 17, no. 3-4, **(2010)**, pp. 143-165.
[7] T. Niknam and B. Amiri, "An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis", Applied Soft Computing, vol. 10, no. 1, **(2010)**, pp. 183-197.
[8] A. Kocheturov, M. Batsyn and P. M. Pardalos, "Dynamics of cluster structures in a financial market network", Physica A: Statistical Mechanics and its Applications, vol. 413, **(2014)**, pp. 523-533.
[9] S. R. Nanda, B. Mahanty and M. K. Tiwari, "Clustering Indian stock market data for portfolio management", Expert Systems with Applications, vol. 37, no. 12, **(2010)**, pp. 8793-8798.
[10] S. H. Sastry, P. Babu and M. S. Prasada, "Analysis & Prediction of Sales Data in SA P-ERP System using Clustering Algorithms", arXiv preprint arXiv:1312.2678, **(2013)**.
[11] S. Fallahpour, M. H. Zadeh and E. N. Lakvan, "Use of Clustering Approach for Portfolio Management", International SAMANM Journal of Finance and Accounting, vol. 2, no. 1, **(2014)**.
[12] A. Stetco, X. Zeng and J. Keane, "Fuzzy cluster analysis of financial time series and their volatility assessment", In Systems, Man, and Cybernetics (SMC), IEEE International Conference, **(2013)**, pp. 91-96.
[13] V. L. Miguéis, A. S. Camanho and J. F. e Cunha, "Customer data mining for lifestyle segmentation", Expert Systems with Applications, vol. 39, no. 10, **(2012)**, pp. 9359-9366.
[14] B. Shim, K. Choi and Y. Suh, "CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns", Expert Systems with Applications, vol. 39, no. 9, **(2012)**, pp. 7736-7742.
[15] C. Pete, C. Julian, K. Randy, K. Thomas, R. Thomas, S. Colin and W. Rüdiger, "CRISP-DM", NCR, SPSS, DaimlerChrysler. USA, **(2000)**, pp. 1-76.
[16] M. N. A. Khan and S. Ullah, "A log aggregation forensic analysis framework for cloud computing environments", Computer Fraud & Security, vol. 2017, no. 7, **(2017)** July, pp. 11-16.
[17] M. N. A. Khan and I. Wakeman, "Machine learning for post-event timeline reconstruction", In First Conference on Advances in Computer Security and Forensics Liverpool, UK, **(2006)**, pp. 112-121.
[18] S. Rahman and M. N. A. Khan, "Review of live forensic analysis techniques", International Journal of Hybrid Information Technology, vol. 8, no. 2, **(2015)**, pp. 379-88.
[19] M. N. A. Khan, C. R. Chatwin and R. C. Young, "Extracting Evidence from Filesystem Activity using Bayesian Networks", International journal of Forensic computer science, vol. 1, **(2007)**, pp. 50-63.
[20] M. N. A. Khan, "Performance analysis of Bayesian networks and neural networks in classification of file system activities", Computers & Security, vol. 31, no. 4, **(2012)**, pp. 391-401.
[21] M. N. A. Khan, C. R. Chatwin and R. C. Young, "A framework for post-event timeline reconstruction using neural networks", Digital Investigation, vol. 4, no. 3-4, **(2007)**, pp. 146-157.
[22] M. S. Bashir and M. N. A. Khan, "Triage in live digital forensic analysis", International journal of Forensic Computer Science, vol. 1, **(2013)**, pp. 35-44.
[23] M. Rafique and M. N. A. Khan, "Exploring static and live digital forensics: Methods, practices and tools", International Journal of Scientific & Engineering Research, vol. 4, no. 10, **(2013)**, pp. 1048-1056.
[24] R. Shehzad and M. N. A. Khan, "Integrating knowledge management with business intelligence processes for enhanced organizational learning", International Journal of Software Engineering and Its Applications, vol. 7, no. 2, **(2013)**, pp. 83-91.
[25] M. Khalid, S. ul Haq and M. N. A. Khan, "An assessment of extreme programming based requirement engineering process", International Journal of Modern Education and Computer Science, vol. 5, no. 2, **(2013)**, pp. 41.

