PROTOTYPE ANALYSIS OF DIFFERENT DATA MINING CLASSIFICATION AND CLUSTERING APPROACHES

Srinivas Kolli and M. Sreedevi
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur,
Andhra Pradesh, India
E-Mail: kollisreenivas@gmail.com

ABSTRACT

Present days, large amount information stored in data sources, which is formally increased based on Knowledge Discovery from different data ware houses. To acquire required and useful data from data sources, some of the techniques, methods and some of developed tools to combine huge amount of data sets. This procedure gives demand to implement novel research field in data mining. The main aim of data mining is to extract required information from huge amount data and change them into meaningful for further use in data retrieval. Classification and Clustering is the main data mining approaches to classify and combine categorical data in a large set of data into required group set of class labels. So in this paper we provide comprehensive analysis of different classification and clustering methods in data mining to efficient data retrieval, which includes neural networks, Bayesian networks and decision trees. We also provide survey on some of semi supervised and supervised outlier detection techniques for categorical data on unlabeled data sets under large instances in data sets with required instances in real time synthetic data. We bring out the keys aspects of different outlier and data mining approaches to data exploration.

Keywords: data mining, data clustering, unsupervised, supervised, outlier techniques, knowledge discovery categorical data.

1. INTRODUCTION

Information retrieval is mining and analysis of data with different formations in meaningful parameter sequences and rules. Information retrieval is to explore, transform and upload transactions with respect to attributes from data warehouses to suitable storage and manage information in multi-dimensional data systems, provide business analysis in software updating and display data in required format. Information retrieval is multi dimensional process which presents analyzed results with reasonable activities. In data retrieval, data to me mined with following two machine learning approaches, i.e. supervised and unsupervised learning approaches. These two approaches consists number of data mining techniques like clustering, classification, outlier and association approaches for categorical data assessment in real time data streams. The general procedure of information retrieval is as shown in Figure-1.

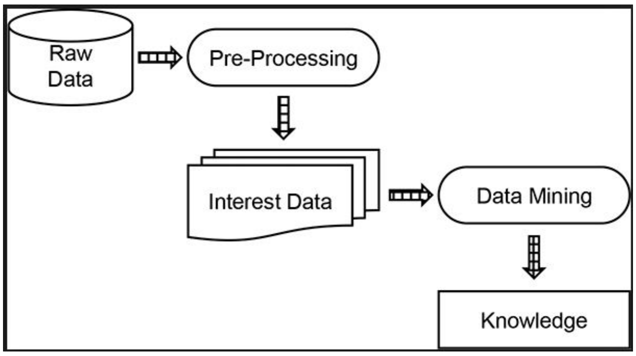

Figure-1. Data extraction procedure for different data sources.

As shown in Figure-1, normal procedure for accessing information from different data sources, there is a data selection, data pre-processing, data transformation and then data retrieval based on data interpretation based on user knowledge. In this paper we provide brief comprehensive survey of different data mining techniques, the main aim of this survey gives different classification and clustering techniques and approaches in data mining. We briefly discuss about these approaches in data exploration.

Remaining of this paper organized as follows: Section 2 describes related work regarding data exploration from different data sources, Section 3 defines techniques used in data mining, section 4 describes the final conclusion of this paper implementation with respect algorithm descriptions.

2. RELATED WORK

In this section, we discuss about different authors, researchers definitions on data exploration from different data sources. We also discuss categorical data description with different real time applications. Categorical data also known as nominal or qualitative multistate details has become increasingly typical in modern real life applications. These dataset are often rich in details and are frequently detected in domains where extensive details places are normal, e.g. ,in network attack recognition .however ,unlike express details, ongoing details attribute principles cannot be naturally mapped on to a range, making most ongoing details research methods in applicable in this setting. Data with express features have been analyzed for a long period, going back at least a century when Karl Pearson [12, 13] introduced the test for independence between express features. The traditional exploratory methods used were concurrent platforms, the chi square statistics, pie maps and unordered histogram. Friendly suggested several sophisticated mathematical methods which are Filter blueprints and Mosaic Displays to view k-way concurrent platforms and Multiple Correspondence analysis (MCA) to deal with multivariate express details places, though most methods are limited to features that take few possible principles. Fernandez talks about several exploratory methods for express details from details exploration perspective. There was number of studies directed at express details in creation community. In particular, one in creation has been to get the groups using the details found in details. one such technique is called Distance Quantification Classing(DQC) was developed by Rosario et al to get the groups found in a class varying in express details set with respect to the Forecaster variable.

3. TECHNIQUES USED FOR DATA MINING

In this section, we discuss data mining techniques like clustering, classification, association and outlier techniques for categorical data based on different attributes. Based on activities and tasks prescribed in different data streams the following classification procedures were used in real time applications.

3.1. Classification: In data visualization and analysis there are two basic formations i.e. Classification and prediction. In that classification is machine learning approach, it performs grouping based on predefined attributes in each data set, classification performs accurate exact required data for each data item from different data sets. Classification done in 2 modules for data processing.

a) Construct data model

b) Use classifier for categorization.

The main objective in accuracy of classification is to identify and applied data relations for different data sets. The mainly used classification in real time applications is binary classification in real data stream evaluation with low and high possible data values. To define relationship between attributes with categorization in different values, there are different classification groups were used for efficient categorization in different attributes.

3.1.1 Decision tree induction classification: Main commonly used technique used data mining for classification is decision tree. Based on instances in decision tree generated efficient connection with different attributes. Decision tree is a direct tree which consists main node as root; root node does not contain any incoming connection from other nodes present in tree. The outgoing node in tree is called as internal node, each internal node denotes identification of attribute and each branch denotes test instances with related attributes. Decision tree algorithm performs recursive function to maintain attribute relations with different instances. Procedure of the decision tree is as follows: first attribute must be act as root node, for effective data maintenance in decision tree root node generate and transform data to different data instances like leaf nodes (i.e. internal nodes), then internal nodes categorize data into separate spaces, until attributes consists same relational classification. Decision tree classification follows top-down, divide and conquer rule in relational attribute partition. Basic structure of decision is as shown in following Figure-2.

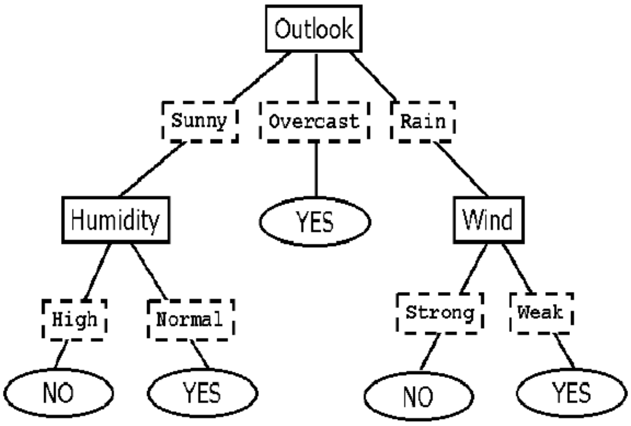

Figure-2. Decision tree procedure in required values check in relational data sets.

As shown in Figure-2, outlook is the root node to define different formations as internal nodes like sunny, overcast, rain which follows humanity decision with different attribute relations in high and normal presentation. In that if humanity is high then sunny is not available to trap the relation, if it is normal then sunny is available to trap the relations, and also wind is strong then rain is not coming based on relation between sub instances and main attribute relations with feasible data presentation. Decision tree classification is decision supporting tool to define relation in different attributes. Finally decision tree performs better and convenient support with categorical data features.

3.1.2 K-Nearest Neighbor (k-NN) classification: It is regression approach to define and maintain instance based learning to define relation between attributes from different data sources. In k-NN classification each feature assigned with same weight then a little bit confusion to separate relevant and irrelevant features from different data sources. k-NN is also used to define and detect unknown column or attribute from different data sources, and it is a simple and convenient method to describe different features partition in data retrieval, by using Euclidian distance to perform variable metric analysis in data representation, it is also local defendable data structure to overlap different attribute variables. Data representation of k-NN is as shown in Figure-3.

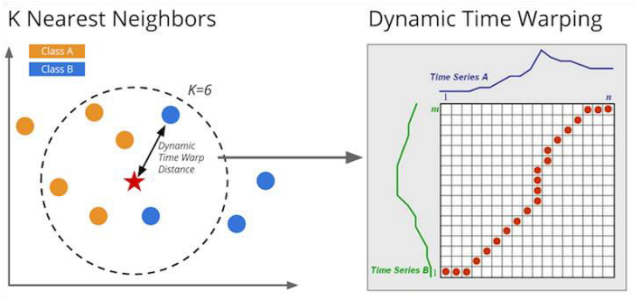

Figure-3. k-NN classification procedure to define different features.

Figure-3 defines basic k-NN structure to evaluate dynamic time wrapping used to classify human actions like sitting walking and lying with different time series in reliable data streams. The above figure contains k=6 with object classification then each object is assigned with separate class then classification occurs with nearest neighbor based on simple and small samples attributes in data separation.

3.1.3 Support Vector Machine (SVM) classification: If we want classify attributes into different classes based on their class label formation with lose of their individuality and generality then SVM classification.

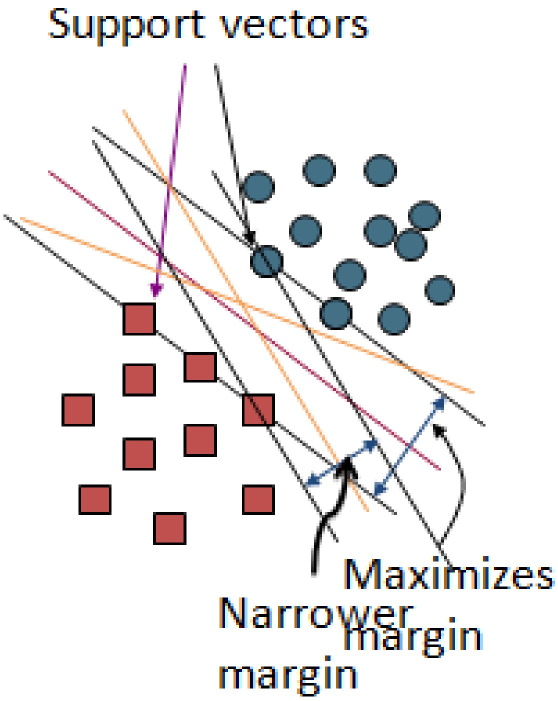

Figure-4. Data representation based on class labels with different features.

As shown in Figure-4, SVM supports for large classifiers to separate functional feature presentation. In this we specify different features as training samples i.e. support vectors based on margin specifications in relevant data presentation from different data sources. This classifier separates hyper planes with different data points, by using quadratic optimization calculation for training data points (x0) with large margin multiplication of parameters.

3.1.4 Component Naïve Bayesian classifier: It is probabilistic data representation model that acquires attribute representation based on their weights in graph presentation. This classifier follows graphs data structure for data representation, nodes are the random variables to represent unknown values with different parameters and their conditional weights in data set evaluation. Data is taken by each individual node based on randomly generated set of values as taken by input and then get conditional values as output. Naïve classifier consists 3 main functions to represent data: unspecified values, attributes, structure representation. It is simple model to solve complex data values. It is flexible and convenient model for probabilistic data representation.

3.2 Clustering: Clustering is combined data into different labels of similar objects with suitable data presentation. Information modeling represents statistical, mathematical and numerical analysis for data evaluation. For efficient data evaluation. In experimentally clustering plays efficient performance in different data retrieval approaches like data exploration, data retrieval in location services with customer relationship management web data analysis pattern recognition in various real time applications. Based on data representation and evaluation survey in real time data set processing more number of clustering techniques was used. In this paper, we give brief describe clustering approaches as follows:

3.2.1 K-Means clustering: It is the main clustering method for different scientific and industrial real time applications. As the name suggest that it represents k-number of clusters with $C_{i,j}$ with c_{ij} data points called as center of data representation. K-Means clustering defines statistical data representation for numerical data only, it does not work well for categorical data attributes in real time In k-means clustering there are 2 versions presents to optimize different data evaluations: 1. First version is worked based on Euclidian distance presentation to assign all data points with their centroid, and also calculate centroid formation for newly arrived objects for data points. Second version of k-means clustering is to give in-depth data analysis to move different data points currently attached with already presented data point from overall data set evaluation.

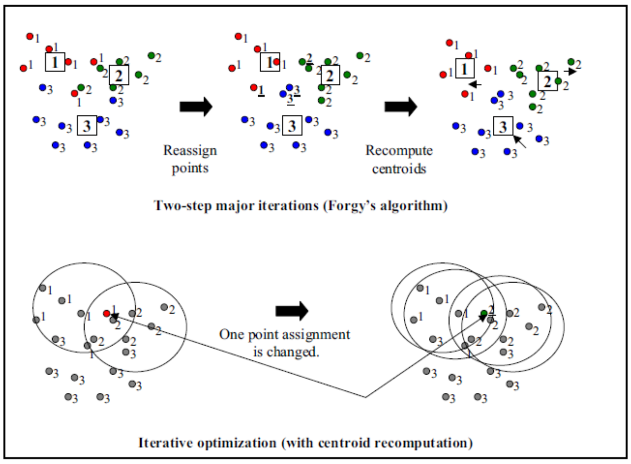
Reassign points
Recompute centroids
Two-step major iterations (Forgy's algorithm)
One point assignment is changed.
Iterative optimization (with centroid recomputation)

Figure-5. Iterative data presentation in k-means clustering feature presentation.

Figure-5 shows both versions of k-means with assignment of newly added data points with different movements in real time data streams. K-means clustering method consists following usually aspects:

a) Initial guess of data assignments

b) Local data optimization for both local and global data assessments.

c) Sensitive data representation with respect to outliers.

d) Result sequences for unbalanced data representation.

Finally k-means defines efficient cluster initialization data representation effectively for full data.

3.2.2 Hierarchal clustering: Hierarchal clustering defines cluster hierarchy with different groups known as dendrogram. Each group consists sub modes based on attribute partitioning with their parent connection parent node. Hierarchal grouping defines data categorized as agglomerative and divisive for data representations with same type of features. Agglomerative grouping approach defines one-to-one point with appropriate groups in same data representations where as divisive clustering defines single data with all data points with recursive execution for appropriate group analysis in same attributes.

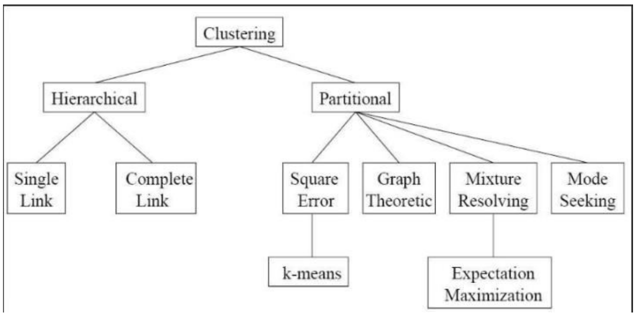

Figure-6. Hierarchical data presentation with different labels.

Hierarchical representation of different links like singleton, square, mixture and data presentation with complete features in real time data presentation and partitioning. Finally in hierarchical clustering point by point with attribute data representation to define linkage based on similar attributes.

3.2.3 Probabilistic clustering: It is the main approach to extract data from different data sources based on randomly changed attributes. The main important function of proposed probabilistic techniques is, it provides mixture model to extract generalized features from heterogeneous data sources. Practically each attribute consists multiple and multivariate constant data with dynamic information length. To solve efficient problem specification in categorical data attributes then probabilistic approach follows markov chain model to represent data in transition matrix representation with dependable data variables. Based on multivariate feature of data exploration in probabilistic approach has following features:

a) This model handle dynamic complex structure based problems to explore data from different sources.

b) It is worked with consecutive chances of data evolution, to represent data into different set of data points based on their feature presentation.

c) To assign random data presentation for continuous iterative process using mixture model

d) The results prescribed with cluster interpretable system applications

From the data mining aspect to extract and combine similar data set presentation follows probabilistic perspective based on parameter sequences in Bayesian network classification. To distribute Bernoulli, Poison, and Gaussian distributed functions with different k-values on Bayesian methodology to explore and evaluate multivariate heterogamous data sources.

3.2.4 Co-Occurrence cluster for categorical data: In co-occurrence grouping walk about categorical data, which is the repeated relation based on dynamic variable size change i.e. transaction with infinite set of attributes (items) from unique set of universal data evaluation. For transactional data maintenance in real time applications using co-occurrence with point by point feature relations. To evaluate categorical data to analyze web extraction, data analysis and other applications. The clustering algorithm like ROCK (Robust Clustering calculation) to evaluate categorical data presentation with many features and also formation of mean with specified cluster is constrained using hierarchal clustering. To find neighbor data points using similarity measures between different parameters with Shared Nearest neighbor approach in real time data streams.

3.2.5 Constrained based clustering: In real time applications, peoples were rarely interested in unconditional solutions, then grouping is the main solution for business activities. To define this type of data exploration from data sources, Tung et al. 2001 introduced constraint based clustering for individual objects and attributes that are recently purchased from transactional data evaluation with different parameter constraints. Based on aggregative functions like min, max, avg for each cluster, which includes individual object constraints to individual partitioning to remove nearest neighbors from cluster representation.

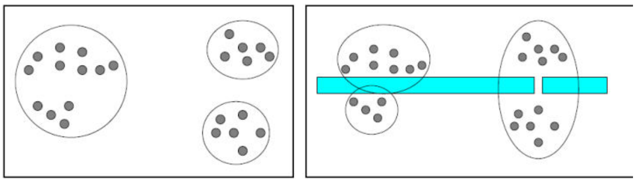

Figure-7. Obstacle data representation with distance similarity measures.

Main application of constraint based clustering is to define spatial data in the presence of obstacles, instead of regular Euclidean distance, short length path between different data points i.e. obstacle distance shown in Figure-7. The difference between three clusters in absence of obstacle in the presence of different data points. Liu *et al.* [2000] recommended another stylish relationship to monitored learning. They considered binary focus on feature described as Yes on factors susceptible to clustering, and defined as No on non-existent artificial factors consistently allocated in a whole attribute space. A choice shrub classifier is used to the full artificial information. Yes-labeled leaves correspond to groups of feedback information. The new technique CLTree (Clustering based on decision Trees) eliminates areas of inhabiting the feedback information with artificial No-points such as: (1) including factors progressively following the shrub construction; (2) making this process exclusive (without physical inclusions in feedback data); (3) problems with uniform distribution in higher measurements. There are more number of clustering approaches were introduced to solve categorical data presentation.

3.3 Outlier detection approaches: Based on different domains with different entity varies based on dependency, this outlier detection procedure aggressively follow following factors such as availability of information, input data type and resource availability by application domains. The following techniques were used to define outliers from real time data streams.

3.3.1 Statistical outlier detection: This uses certain type of mathematical submission and computes the factors by supposing all information factors have been produced by a mathematical submission. Here outliers are factors that have a low possibility to be produced by overall withdrawals Statistical outlier recognition technique is also known as the parametric strategy. This technique is developed using the submission of information factor available for handling .Detection design is meant to fit the data with referrals to submission of information .Gaussian mixture model was suggested by yaminishiet [1].where each point is given a developed ranking information factors which have a high ranking announced as outlier. Discovering outlier centered on general design within information factors was developed by [2] where it combine both Gaussian combination design and supervised method.

3.3.2 Distance based outlier: This outlier recognition strategy assesses a point depending on the distances to its others who live nearby. Primary design of range based outlier is given shown in Figure-8.

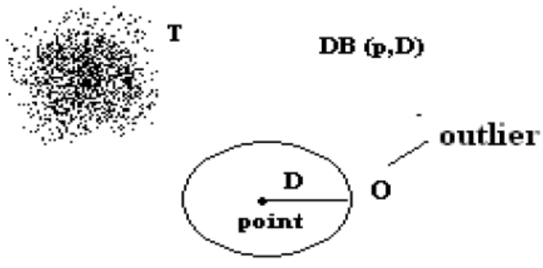

Figure-8. Outlier detection with distance based methodologies.

Precise range centered techniques are rely on the well known closest next door neighbor concept .Ng and knor propose[5] a well known range measurement to identify outliers. They determine outlier as the item which was higher in range to its others who live nearby. The stacked loop (NL) criteria, determines the range between each and every couple of things and then set as outliers those that are far away from most things. The NL criteria has problems with regard to the variety of things, which makes it unsuitable for exploration very huge data source such are seen in govt review information, system information, and medical test information. This outlier technique was provided in knorr and Ng as a product O in dataset T is a DB (p, D) outlier if at least portion p of the item in T lie at a range higher than D from O. The Parameter p used here is the little portion of things that must existing outside an outliers D-neighborhood.

Finally we develop better clustering approaches to define links and properties for categorical data in real time data streams. We develop link based clustering, ensemble clustering and other advanced clustering approaches to define efficient attribute evaluation from data sources.

4. CONCLUSIONS

In this paper, we discuss about different clustering, classification approaches to support categorical data in real time applications. In classification, we discuss different types of models like decision tree, k-Nearest Neighbor, SVM, and CNB to support data presentation based on different class labels with different features. In clustering, we discuss about k-means, Hierarchal, Probabilistic, and co-occurrence clustering approaches to combine different relevant items from different data sets. Based on these cluster approaches, we discuss about outlier detection procedures in various scenarios like statistical outlier, distance based outlier in real time applications. As further improvement of our research, we implement advanced clustering and classification approaches to do data processing effectively.

REFERENCES

[1] Pavel Berkhin. Survey of Clustering Data Mining Techniques. Author's address: Pavel Berkhin, Accrue Software, 1045 Forest Knoll Dr., San Jose, CA.

[2] Sakshi, Prof. Sunil Khare. 2015. A Comparative Analysis of Classification Techniques on Categorical Data in Data Mining. International Journal on Recent and Innovation Trends in Computing and Communication.

[3] Madhavi Alamuri, Bapi Raju Surampudi and Atul Negi. 2014. A Survey of Distance / Similarity Measures for Categorical Data. 2014 International Joint Conference on Neural Networks (IJCNN) July 6-11, Beijing, China.

[4] K.T.Divya1, N.Senthil Kumaran2. 2016. Survey On Outlier Detection Techniques Using Categorical Data. International Research Journal of Engineering and Technology (IRJET). 03(12).

[5] S. Sarumathi, N.Shanthi, M.Sharmila. 2013. A Comparative Analysis of Different Categorical Data Clustering Ensemble Methods in Data Mining. International Journal of Computer Applications (0975-8887). 81(4).

[6] Sandro Vega-pons & Jose reuiz Shulcloper. 2011. A Survey of Clustering Ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence. 25(3): 337-372.

[7] Harun Pirim, Dilip Gautam, Tanmay, Bhowmik, Andy D. Perkins, Burak Ekşioglu, & Ahmet Alkan. 2011. Performance of an ensemble clustering algorithm on biological datasets. Mathematical and Computational Applications. 16(1): 87-96.

[8] Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, & Chris Price. 2012. A Link based cluster ensemble approach for categorical data clustering. IEEE Transactions on knowledge and data engineering. 24(3).

[9] Sandro Vega-pons & Jose reuiz Shulcloper. 2011. A Survey of Clustering Ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence. 25(3): 337-372.

[10] J. Xie, B. Szymanski and M. J. Zaki. 2010. Learning dissimilarities for categorical symbols. Journal of Machine Learning Research-Proceedings Track. 10: 97-106.

[11] D. Ienco, R. G. Pensa and R. Meo. 2012. From context to distance: Learning dissimilarity for categorical data clustering. ACM Transactions on Knowledge Discovery from Data (TKDD). 6(1): 1.

[12] F. Cao, J. Liang, D. Li, L. Bai and C. Dang. 2012. A dissimilarity measure for the k-modes clustering algorithm. Knowledge- Based Systems. 26: 120-127.

[13] T. Herawan, M. M. Deris and J. H. Abawajy. 2010. A rough set approach for selecting clustering attribute. Knowledge-Based Systems. 23(3): 220-231.

[14] F. Cao, J. Liang, L. Bai, X. Zhao and C. Dang. 2010. A framework for clustering categorical time-evolving data. IEEE Transactions on Fuzzy Systems. 18(5): 872-882.

[15] Z. Khorshidpour, S. Hashemi and A. Hamzeh. 2011. An approach to learn categorical distance based on attributes correlation. in Electrical Engineering (ICEE), 2011 19th Iranian Conference on, May. pp. 1-6.

[16] Satish Kumar David, Amr T.M. Saeb, Khalid Al Rubeaan. 2013. Comparative Analysis of Data Mining

Tools and Classification Techniques. Computer Engineering and Intelligent Systems. 4(13).

[17] Lashari S. A. and Ibrahim R. 2003. Comparative Analysis of Data Mining Techniques for Data Classification. International Conference on Computing and Informatics, ICOCI.

[18] Vrushali Bhuyar. 2014. Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate. International Journal of Emerging Trends & technology in Computer Science (IJETTCS). 3(2).

[19] Sagar S. Nikam. 2015. A Comparative Study of Classification Techniques in Data Mining Algorithms. ISSN: 0974-6471, 8(1).

[20] Anand V. Saurkar, Vaibhav Bhujade, Priti Bhagat, Amit Khaparde. 2014. A Review Paper on Various Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering. 4(4).

[21] 2015. Effective Data Mining for Proper Mining Classification Using Networks. International Journal of Data Mining & Knowledge Management Process (IJDKP). 5(2).

[22] Mahnoosh Kholghi, Hamed Hassanzadeh, Mohammad Reza Keyvanpour. 2010. Classification and Evaluation of Data Mining Techniques for Data Stream Requirements. International Symposium on Computer, Communication, Control and Automation.

[23] Syeda Farha Shazmeen, Mirza Mustafa Ali Baig, M. Reena Pawar. 2013. Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis. IOSR Journal of Computer Engineering (IOSR-JCE) e- ISSN: 2278-0661, 10(6): 01-06.

[24] S. Archana, Dr. K. Elangovan. 2014. Survey of Classification Techniques in Data Mining. International Journal of Computer Science and Mobile Applications. 2(2).

[25] K. Wisaeng. 2013. A Comparison of Different Classification Techniques. International Journal of Soft Computing and Engineering (IJSCE). 3(4).

[26] Mining regular closed patterns in transactional databases. Intelligent systems and controls 7th international conference on intelligent systems proceedings in IEEE, DOI: 10.1109/ISCO.2013.6481184 publishes in IEEE Xplore.

[27] 2013. Closed regular pattern mining using vertical format. International Journal of Computer Science & Engineering Technology (IJCSET). 4(7), ISSN 2229-3345, pp. 1051-1056.

[28] Parallel and distributed closed regular pattern mining in large databases. IJCSI International Journal of Computer Science Issues, 10(2, No 2), March 2013, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784, pp. 264-269.