Literature Reviewof Design and Development of Novel Techniques Harnessing In Data Mining For Clustering and Classification of Data

Rajesh S. Walse
School of Computational Sciences (Research Center), Swami Ramanand Teerth Marathwada University, Nanded, India, (CDT MAFSU, Nagpur)

Dr. G.D. Kurundkar
Department of Computer Science, Shri Gurubuddhi Swami Mahavidyalaya, Purna Dist. Parbhani, S.R.T.M.University, Nanded, India

Dr. Parag Bhalchandra
School of Computational Sciences, Swami Ramanand Teerth Marathwada University, Nanded, India

Abstract—

The title of the Research topic is "Design and Development of Novel Techniques for Clustering and Classification of Data" In this research proposedto develop a Novel techniques, techniques in terms of applying and to develop new Algorithm in Data Mining Techniques, before applying techniques it required some input database without database it is not possible, therefore preliminary it is proposed to use some dairy related as well as veterinary related datasets for making clusters and classification techniques.

Clustering the data, people can obtain the data distribution, observe the character of each cluster and make further study on particular clusters. In addition, cluster analysis usually acts as the preprocessing of other data mining operations. Therefore, Cluster analysis has become a very active research subject in data mining. Data mining is a new technology developing with database and Artificial intelligence. It is processing procedure of extracting credible, novel, effective and understandable patterns from database. Cluster analysis is an important data mining technique used to find data segmentation and pattern information. As the development of data mining a number of clustering methods have been founded, The Study of clustering techniques from the perspective of statistics based on the statistical methods with the computer algorithm techniques and introduces the existing excellent statistical methods including factor analysis, Correspondence analysis and functional data analysis into data mining. The present study is undertaken to develop a Data Mining workflow using clustering and classification of data to solving clustering problem as well as extracting potentially interesting association rules.

Keywords—Data Mining, Clustering, Classification, KNN, Weka, K-Means, Matlab, ISODATA, SRIDHCR, HDFS

I. Introduction

Data Mining:

1. Data mining finds valuable information hidden in large Volumes of data
2. Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data
3. The Computer is responsible for findings the pattern by identifying the underlying rules and features in the data.

Literature related to the topic either through advanced books / paper:

The interest in analyzing data has grown tremendously in recent years, as business in all sectors have discovered the potential of using the data scattered in diverse business systems as one coherent whole for better understanding and management of the business. To analyze data a multitude of technologies is needed namely technologies from the areas of data mining, data warehouse, online analytical (OLAP) data visualization and customer relationship management (CRM).

Data analysis systems are increasingly based on data mining in classification. People are often prone to making mistakes during analysis or, possibly, when trying to establish relationship between

multiple features. Some applications area cannot be handled satisfactorily using classification rules as the data is too complex.

Several classification models have been proposed in the literature, including data mining techniques, decision trees, statistical model, and machine learning tools. R Algorithms presents the classification rules where we discussed number of methods and literature review. Clustering and Classification is an extensively studied problem (Mainly in statistics, machine learning and neural network), HDFS, Hadoop Distributed File System using Hadoop to classify data and make a clusters.

Classification is probably one of the most widely used data mining techniques with a lot of extensions scalability is still an important issue for database applications, thus combining classification with database techniques should be promising topic.

II. Applications Of Classification

1.To predict tumor cell as begin or malignant
2.In fraud detection related to credit card transaction analysis as legitimate fraudulent type.
3.To classify the secondary structure of protein in to alpha.

III. Applications Of Clustering

1.Data Mining (DNA – Analysis, marketing studies
2.Text Mining (Text type clustering)
3.Information retrieval
4.Statistical computational linguistics
5.Corpus based computational

1.	Antonio Roberto formaggio, Matheus Alves Vieira, Camilo DalelesRanno	2012	Object Based Image Analysis (OBIA) Anand Data Mining (DM) in Landsat Time Series for Mapping Soybean in Intensive Agricultural Regions [1]	Projection system first degree polynomial and Nearest neighbor re-sampling, knowledge extraction, structural classifier in DM (J 48 algorithm) Classification	New classification approach, integrated object based evaluated J48 algorithm was highly logic.	[1]
2.	Manish Saggar, Ashish Kumar Agrawal, Abhimanyu Lad	2004	Optimization of Association Rule Mining using Improved Genetic Algorithms [2]	Algorithms are rate mining	Enormous robust of GA's in mining the Association Rules. Toolkit can also handle other database	[2]
3.	CarolinFenlon, Luke O Grady, Laurence Shalloo	2016	Regression techniques for modeling conception in seasonally calving dairy cows [3]	Irish Dairy Management systems, Descriptive statistics of moorepark, Regression analysis. R-statistical programming language, mixed effects, logistic regression.	An interaction between log days in milk and the calving interval genetic trait also found to be statistically significant.	[3]
4.	Yaoguang Hu, ZhengjieGuo, Jingqian Wen, Jialin Han	2015	Research on knowledge mining for agricultural machinery maintenance based on association rules [4]	Knowledge preparation, mining implementation, K-nearest neighbor algorithms. KNN algorithm, Vector space model, Data integration, Association Rules, Hierarchical hidden mark model, Kernel algorithm. WEKA O.S. packages.	Redundancy information has been cleaned. The data of Agricultural machinery maintenance is essential information for Causation analysis. To find out high related factors around all the information influences. The normal working of agricultural machines.	[4]
5.	Niketa Gandhi, Leisa J.	2016	Rice Crop Yield Forecasting of Tropical	Koopen classification applying various Data	Calculating effect of precipitation and average	[5]

	Armstrong		Wet and Dry Climatic Zone of India Using Data Mining Techniques [5]	Mining Techniques, Weka, Data Visualization	temperature on Rice crop yield. Effect of maximum and minimum Temperature rice crop yield. This research discusses the result achieved on rice crop yield data set of Tropical wet and dry climate zone after executing the classification algorithm. Executed in Weka.	

There are various techniques like classification, Clustering, Association and Outlier analysis used for the knowledge extraction. A comprehensive literature review of various significant researches in the area of Dairy Technology/ Dairy Industries / Agricultural / Veterinary related published research papers as well as some of the research has been carried out in the title of Data mining, clustering, classification using machine learning tools and Design Novel Techniques using machine learning applications different new algorithms.

Some of the references ranging from year 2013 to 2017 (i.e. from last 5 years is presented below in categorical tabular form.)

IV.Literature Review

TABLE I. (A) COMPARISON OF DIFFERENT DATA MINING TECHNIQUES

TABLE II. (B) COMPARISON OF DIFFERENT DATA MINING TECHNIQUES

Sr. No.	Author	Year	Techniques Used	Methodology	Key findings	Reference
6.	SabriSerkan Gulluoglu	2015	Segmenting Customers With Data Mining Techniques [6]	SPSS, Excel Association rule mining algorithm apriori is used	The item which is sold the most was feta cheese, which was in 53% of all transaction. In their study, not only one antecedent association rules are generated, but also 2 and 3 antecedents are generated as well as most sold items can be in promotion together to increase sales by attracting customers.	[6]
7.	Zahid Halim, Dr. Rauf Baig, Shariq Bashir	2006	Sonification: A Novel Approach towards Data Mining [7]	RPAI Algorithm	Auditory icons have a great potential as a strategy for creating informative strategy for data mining and classification. RPAI algorithm presented here to predicts the probability of rains based on weather conditions of previous 48 hours.	[7]
8.	Kriti Bhargava, Stepan Ivanov, William Donnelly, ChamilKulatunga	2016	Using Edge Analytics to Improve Data Collection in Precision Dairy Farming[8]	Data analysis in WSN, Data compression algorithm, Sensor Analytics, Edge mining techniques, Data reduction on our collar device to	Implementation of WSN Technology in the precision dairy farming. Implementation of light weight edge mining algorithms.	[8]

				perform localized data compression. Edge mining algorithm provides a foundations for future real time responsiveness of the system The performance of LSIP data compression using R-analysis.		
9.	Niketa Gandhi, Leisa J Armstrong	2016	A review of the application of data mining techniques for decision making in agriculture[9]	Machine learning, ANN, SVM, Association Rule Mining. The relations of factors of factors of on crop	The review has outlined a number of promising techniques that have been used to understand various climate and other production	[9]
10.	SanyamBharara, A. Sai Sabitha, Abhay Bansal	2017	A review on Knowledge extraction for Business operations using Data Mining[10]	Knowledge extraction classification, clustering KM applications, DM techniques, association text mining.	In this research work data mining and knowledge management concepts are applied for business management. The other DM techniques like time series, outlier analysis can also be applied in business operation field.	[10]

Table III (C) Comparison of Different Data Mining Techniques

Sr. No.	Author	Year	Techniques Used	Methodology	Key findings	Reference
11.	R. Senthil Kumar, Dr. C. Ramesh	--	A Study on Prediction of Rainfall Using Data Mining Technique[11]	K-means clustering algorithm, Naïve Bayes algorithm	(ANOVA) analysis of various popular data mining algorithms is presented for rainfall prediction	[11]
12.	Wei Zhang, Shuping Li, Xue Wang, Chunyan Xia	2011	Application of Data Mining in Agricultural Topic Tracking[12]	Heuristic attribute reduction algorithm	Agricultural topic tracking by the data mining technique carryout.	[12]
13.	AyodeleLasisi and Rozaidaghazali, FolaLasisi, TututHerawan, Mustafa Mat Deris	2015	Knowledge Extraction of Agricultural Data Using Artificial Immune System[13]	Data Mining, Clonal selection algorithm, artificial immune recognition system, K-Nearest Neighbor, SVM, NN	DM techniques provides such extraction capabilities and several have been used over the years such as K-Means, KNN, SVM, and drafted for gaining insights in to the agricultural data provided by researchers in New Zealand.	[13]

14.	Hisayoshi Kato, Hironori, Hiraishi and Fumio Mizoguchi	2001	Log summarizing agent for web access data using Data Mining techniques [14]	Log summarizing agent request from user, web log database, log classification agent	Interesting user access pattern is in users each domain	[14]
15.	R. Sujatha, Dr. P. Isakki Devi	2016	A Study on Crop Yield Forecasting Using Classification Techniques [15]	Data mining, classification, clustering, classification algorithm, Rule based classifier, Bayesian networks, SVM, NN, ANN, Genetic Algorithms	How improving agriculture efficient by prophesying and improves yields by previous agricultural information, select best crop by farmer depending on whether situation and provides required information to prefer the suitable season to do excellent farming.	[15]
16.	Liquiong Tang, Phillip Abplanalp	2014	GPS Guided Farm Mapping and Waypoint Tracking Mobile Robotic System[16]	MATLAB simulation RTD-GPS	The future application of robust, reliable and smart Robotic systems in agricultural and dairy industry is huge with some advanced processing the combination of both GPS and IMU could significantly improve the accuracy of the navigation system.	[16]
17.	V. Vijay Hari Ram, H. Vishal	2015	Regulation of Water in Agriculture Field Using Internet of Things [17]	RFID, GPS, IoT Sensors	Working the concepts of the Internet of things to its extent and improve the funding of the device by using peripheral device.	[17]
18.	Srisruthi.S, N. Swarna, G.M. SusmithaRos, Edna Elizabeth	2016	Sustainable Agriculture using Eco-friendly and Energy Efficient Sensor Technology [18]	Pressure sensor, Temperature sensor, eight sensor colours	In Comparison to industrialized agriculture, sustainable agriculture, promises economic stability for farmers to lead a better quality of life.	[18]

Table IV (D) Comparison of Different Data Mining Techniques

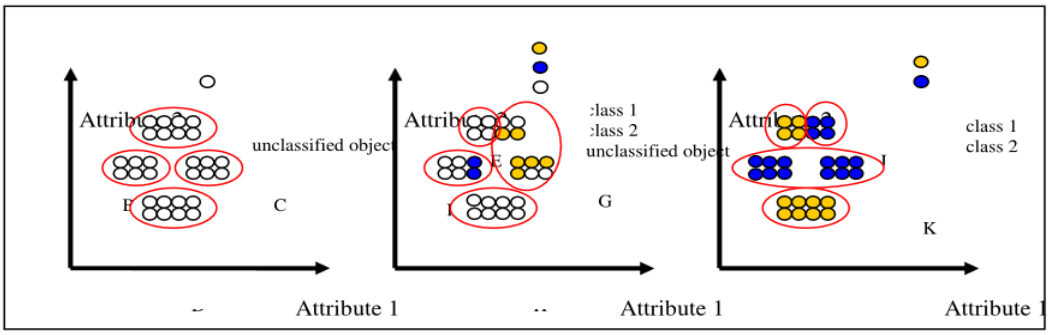
Sr. No.	Author	Year	Techniques Used	Methodology	Key findings	Reference
19.	Esther Hochsztain	2015	A Mining approach to evaluate geoportals usability [19]	Experimental data Analysis, Data Mining, Web Mining Association rule	Iterative proposal to evaluate geoportals usability based on the data and web mining propose a framework and association rule usability.	[19]
20.	Lorenzo Di Silvestro, Michael Burch, Margherita Caccamo, Daniel Weiskopf, Fabian Beck, Giovanni Gallo	2014	Visual Analysis of Time-dependent Multivariate Data from Dairy Farming Industry [20]	2D time series, visual Analytics, Visualization tools.	Analyzing data collected by the dairy industry with the aim of optimizing the cattle-breeding management and maximizing profit in the production of milk. Used a visual analytics approach to support animal researchers in analyzing multivariate time varying data.	[20]
21.	M.H. Ariff, I. Ismarani, N.	2014	RFID Based Systematic Livestock	RFID sensors, tags.	RFID based Systematic livestock's Health management software system	[21]

	Shamsuddin		Health Management System [21]		has been presented to determine the health status of the livestock. The system is user friendly, reliable, simple and can be used in a livestock farm for a long time. System generate power from laptop battery.	
22.	Chen Jinyin, Lin Xiang, Zheng Haibing, BaoXintong	2017	A novel cluster center fast determination clustering algorithm [22]	K-means algorithm normal distribution, Density distribution CH-CCFDAC algorithm	10 clusters centres represent 10 images with different numbers. This shows that the 10 clusters CH-CCFDAC can accurately find clustering centres. Confidence interval points are selected, then the cluster centers are determined automatically.	[22]
23.	S. DilliArasu, Dr. R. Thirumalaiselvi	2017	A Novel imputation method for effective prediction of coronary kidney disease [23]	WAELI (Weighted average Ensemble learning imputation technique.	Innovative solution to handle missing values for CKD data set. Proposed WAE LI algorithm predicts the missing value using single value imputation.	[23]

V. Study of existing problems in the Research domain: -

The content domain is the body of knowledge skills or abilities being measured or examined by a test, experiment or research study. A researcher would want a content domain to cover all aspects to the subject area as well as be well defined and objective.

In the datamining world, clustering and classification are two types of methods. Both these methods characterize objects in to groups by one or more features. The key difference between clustering and classification is that clustering is an unsupervised learning techniques used to group similar instances based on features, where as classification is a supervised learning technique used to assign predefined tags to instances based on features. Data is business. Abundant data carries abundant problems. Data mining involves various techniques and they have associated problems carried with them along effective data mining becomes very vital in today and tomorrow scenario. This becomes possible with effective usage of clustering and classification of data to make the data move robust, meaningful and usable with least wastage.



a. Traditional clustering b. Semi-supervised clustering c. Supervised clustering

Fig. I: Traditional, Semi-Supervised, and supervised clustering

VI. Conclusion And Future Scope:

Research work can improve the performance of traditional algorithms like k-means and presents a hybrid approach. The frequently used algorithm is k-means, which can deal with small convex datasets preferably. It reduces the error rate and achieves accuracy. This research compares the efficiency of these clustering and classification data mining algorithms by applying them to datasets.

The Hadoop HDFS using map reduce techniques of this experiment results will generate the best classification accuracy and show the high robustness and generalization capacity to the other algorithms like ISODATA and SRIDHCR.

a. The data algorithms need to be more efficient and scalable to effectively extract the information from huge amount of data in databases.
b. Dealing with huge datasets that require distributed approaches
c. Mining information from heterogeneous databases and global information systems
d. Processing of large, complex and unstructured data into a structured format

References:

1. Antonio Roberto formaggio, Matheus Alves Vieira, Camilo DalelesRanno,, Object Based Image Analysis (OBIA) Anand Data Mining (DM) in Landsat Time Series for Mapping Soybean in Intensive Agricultural Regions , 978-1-4673-1159-5/12/$31.00 2012 IEEE pg(2257-2260)2012
2. Manish Saggar, Ashish Kumar Agrawal, Abhimanyu LadOptimization of Association Rule Mining using Improved Genetic Algorithms, 0-7803-8566-7/04/$20.00 IEEE pg (3725-3729),2004
3. CarolinFenlon, Luke O Grady, Laurence ShallooRegression techniques for modeling conception in seasonally calving dairy cows, 2016 IEEE 16th International Conference on Data Mining Workshops, 2375-9259/16 2016.174 pg (1191-1196),2016
4. Yaoguang Hu, ZhengjieGuo, Jingqian Wen, Jialin Han Research on knowledge mining for agricultural machinery maintenance based on association rules, 978-4799-8389-6/15/$31.00 2015,IEEE . pg (885-890) 10th Conference on Industrial Electronics and Applications (ICIEA) , 2015
5. Niketa Gandhi, Leisa J. Armstrong Rice Crop Yield Forecasting of Tropical Wet and Dry Climatic Zone of India Using Data Mining Techniques, 2016 IEEE 16th International Conference on Advances in Computer Applications (ICACA) IEEE, 978-I-5090-3770-4/16 pg (357-363), 2016
6. SabriSerkanGulluogluSegmenting Customers With Data Mining Techniques, ISBN: 978-4799-6376-8/15/$31.00 IEEE pg (154-159), 2015
7. ahid Halim, Dr. Rauf Baig, ShariqBashirSonification: A Novel Approach towards Data Mining, IEEE-ICET—2006 2nd International Conference on Emerging Technologies 1-4244-0502-5/06/$20.0 13-14, November 2006 pg (548-553), 2006
8. Kriti Bhargava, Stepan Ivanov, William Donnelly, ChamilKulatungaUsing Edge Analytics to Improve Data Collection in Precision Dairy Farming, 2016 IEEE 41st Conference on Local Computer Networks Workshops DOI 10,1109/LCNW, 2016.9 pg(137-144), 2016
9. Niketa Gandhi, Leisa J ArmstrongA review of the application of data mining techniques for decision making in agriculture, 978-1-5090-5256-1/16/$31.00 IEEE, 2nd International Conference on Contemporary Computing and Informatics (ic3i) 2016 pg (1-6), 2016
10. SanyamBharara, A. Sai Sabitha, Abhay Bansal A review on Knowledge extraction for Business operations using Data Mining, 978-1-5090-9/17/$31.00 IEEE, 2017 7th International Conference on Cloud Computing, Data Science & Engineering –Confluence pg (512-518), 2017
11. R. Senthil Kumar, Dr. C. RameshA Study on Prediction of Rainfall Using Data Mining Technique, Department of Computer Science and Engineering Satyabama University Chennai pg(09), Research Scholar
12. Wei Zhang, Shuping Li, Xue Wang, ChunyanXiaApplication of Data Mining in Agricultural Topic Tracking, 978-1-4577-1587-7/11/$26.00 IEEE International Conference on Computer and Network Technology pg (38-41), 2011
13. AyodeleLasisi and Rozaidaghazali, FolaLasisi, TututHerawan, Mustafa Mat DerisKnowledge Extraction of Agricultural Data Using Artificial Immune System, 978-1-4673-7682-2/15/$31.00 IEEE 2015 12th International conference on Fuzzy Systems and Knowledge Discovery pg (1653-1658), 2015
14. Hisayoshi Kato, Hironori, Hiraishi and Fumio MizoguchiLog summarizing agent for web access data using Data Mining techniques, 0-7803-3/01/$10.0 IEEE 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) pg (2642-2647), 2001

15. R. Sujatha, Dr. P. IsakkiDeviA Study on Crop Yield Forecasting Using Classification Techniques, 978-1-4673-8437-7/16/$31.00 IEEE pg (4), 2016
16. Liquiong Tang, Phillip AbplanalpGPS Guided Farm Mapping and Waypoint Tracking Mobile Robotic System, 978-1-4799-4315 2014 IEEE 9th Conference on Industrial Electronics and Applications (ICIEA), pg (1676-1681) , 2014
17. V. Vijay Hari Ram, H. Vishal Regulation of Water in Agriculture Field Using Internet of Things, 978-1-4799-7758-1/15/$31.00 IEEE , International Conference on Technological Innovations in ICT for Agriculture and Rural Development (TIAR 2015), pg (112-115) , 2015
18. Srisruthi.S, N. Swarna, G.M. SusmithaRos, Edna Elizabeth Sustainable Agriculture using Eco-friendly and Energy Efficient Sensor Technology, 978-1-5090-0774-5/16/3$31.00 IEEE International Conference on Recent Trends in Electronics Information Communication , Technology, 2013 , India pg (1442-1446), 2016
19. Esther HochsztainA Mining approach to evaluate geoportals usability , 978-1-4673-8111-6/15 $31.00 IEEE International Workshop on Data Mining with Industrial Applications pg (1-7), 2015
20. Lorenzo Di Silvestro, Michael Burch, Margherita Caccamo, Daniel Weiskopf, Fabian Beck, Giovanni GalloVisual Analysis of Time-dependent Multivariate Data from Dairy Farming Industry, 5th International Conference on Information Visualization theory and applications (IVAPP – Pages 99-106 2014 – International Conference on Information Visualization Theory and Applications – IVAPP-*Italy,Germanypg (99-106)*, 2014, ISBN: 978-989-758-055-5
21. M.H. Ariff, I. Ismarani, N. ShamsuddinRFID Based Systematic Livestock Health Management System, 978-1-4700-6100 -1/14/$31.00 IEEE Conference on Systems, Process and Control (ICSPC-2014) pg (6), 2014
22. Chen Jinyin, Lin Xiang, Zheng Haibing, BaoXintongA novel cluster center fast determination clustering algorithm, ELSEVIER, Applied Soft Computing 57 (2017) 539-555- Journal www.elsevier.com/locate/950c- 2017. Pg (539-555), 2017
23. S. DilliArasu, Dr. R. ThirumalaiselviA Novel imputation method for effective prediction of coronary kidney disease, 978-1-5090-6221-8/17/$ 31.00 C -2017 IEEE 2nd International Conference on Computing and Communications Technologies (ICCT/17) pg (127-136), 2017