Iris Data Classification using Unsupervised and Supervised Learning Techniques

Dr. Pankaj Kawadkar

Head & Associate Professor, Department of CSE

School of Engineering, SSSUTMS, Sehore (M.P.), India

kawadkarpankaj@gmail.com

Abstract- Data mining gives various types of clustering classification algorithm for the various number of applications such as banking, education, medical science, fraud detection, pattern representation, feature extraction for the respective filed etc., there are various algorithm such as supervised learning methods, unsupervised learning methods and semi supervised learning methods. There are various algorithm we can used with the data mining techniques to improve the performance of the system or any algorithm for the various number of field such as information retrieval and mining of the data, fraud detection, education sector, medical science, business transaction etc.

Keywords:- Data Mining, Supervised Learning, Unsupervised Learning, Health Care.

Introduction

The terms data mining, patent mining, text mining and visualization are employed for the processing of the documents. Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Clustering is a division of data into groups of similar objects. Data Mining is defined as mining of knowledge from huge amount of data. Using Data mining we can predict the nature and behaviour of any kind of data. The past two decades has seen a dramatic increase in the amount of information being stored in the electronic format. This accumulation of data has taken place at an explosive rate. It was recognized that information is at the heart of the business operations and that decision makers could make the use of data stored to gain the valuable insight into the business. DBMS gave access to the data stored but this was only small part of what could be gained from the data. Analysing data can further provide the knowledge about the business by going beyond the data explicitly stored to derive knowledge about the business.

II. Literature Review

[1] The goal of this session was to create a single venue for cross-disciplinary researchers to present research on social media mining for public health monitoring and surveillance. The session provided a forum to share new research in a variety of important public health areas, including the detection of disease outbreaks and awareness; pharma-covigilance, including interactions with natural products and dietary supplements; and various issues related to behavioral medicine, including weight loss, e-cigarette use, and well-being. Through these projects, researchers also advanced the technology needed to understand social media text, for example by developing new

NLP classifiers, new topic model variations, and new visualization systems. Given the ever-increasing amount of social media data around the world, interest in such systems will only increase over time.

[2] In this research the work has been provided on a particular dataset using classification and feature selection approach. In the mentioned method, first, instruction set is divided into two groups of healthy and sick people, then in the second stage 8192 subsets were extracted from the total features with clear cost for any subset (cost gain from feature ranking) which in the third stage by PSO algorithm for all the subsets a learning classifier (FFBP) is given for all subsets to find best subset with the highest accuracy and accessibility and the lowest cost and time.

[3] The system described by this paper uses the current knowledge about artificial neural networks, being able to suggest a diagnosis regarding skin diseases from erythemato-scuamous class. The feed-forward neural networks with back-propagation learning algorithm are frequently used by medical applications. The most important benefit is recorded for diseases with a large number of symptoms, diseases that are difficult to be identified even after a detailed analysis of a human physician.

[4] leverage the power of smart phone to enable proactive in-house heart condition monitoring. They introduce Heart-Trend, a nonparametric model to analyze and detect heart abnormality conditions like arrhythmia From photoplethysmogram (PPG) signal. It does on-demand heart status monitoring using smart phones (can also be implemented in PC/ICU monitors) and facilitates timely detection of heart condition deterioration to permit early diagnosis and prevention of fatal heart diseases. Proposed robust anomaly analytics engine accurately detects the morphological trend to find abnormal heart condition in real time through machine learning based trend prediction.

[5] Proposed the main motivation of this paper is to provide an insight about detecting heart disease risk rate using data mining techniques. Various Data mining techniques and classifiers are discussed in many studies which are used for efficient and efficacious heart disease diagnosis. As per the analysis mode, it is seen that many authors use various technologies and different number of attributes for their study. Hence, different technologies give different precision depending on a number of attributes considered.

[6] In this paper author presents the comparative evaluation study for the data mining and their functions, Each article was categorized according to the main data mining functions: clustering, association, classification, and regression; and their application in the four main library aspects: services, quality, collection, and usage behavior. Findings indicate that both collection and usage behavior analyses have received most of the research attention, especially related to collection development and usability of websites and online services respectively. Furthermore, classification and regression models are the two most commonly used data mining functions applied in library settings. Additionally, results indicate that the top 6 journals of articles published on the application of data mining techniques in academic libraries are: College and Research Libraries, Journal of Academic Librarianship, Information Processing and Management, Library Hi Tech, International Journal of Knowledge, Culture and Change Management, and The Electronic Library. Scopus is the multi-disciplinary database that provides the best coverage of journal articles identified. To our knowledge, this study represents the first systematic, identifiable and comprehensive academic literature review of data mining techniques applied to academic libraries.

[7] Here author discuss the Polynomial Neural Network is a self-organizing network whose performance depends strongly on the number of input variables and the order of polynomial which are determined by trial and error. In this paper, classification of real world problem has been

solved using GA based Polynomial Neural Network (PNN).The performance of the proposed algorithm has been studied using gradient descent (back-propagation) and genetic algorithm. The experimental results show that GA based PNN achieves better classification accuracy on real world problem as compared to BP based PNN for all the considered four cases of UCI datasets. An attempt should be made to reduce the classification time of the PNN-GA method, which is a part of future works.

[8] In this research, evolution of only the number of hidden layers and hidden nodes has been considered with regard to the adaptive optimisation of an ANN. But it is well known that these are not the only parameters that can be optimized in a given ANN. Therefore in the future, this research work can include the adaptive optimisation of other ANN parameters like the learning rate, learning momentum and activation functions, in order to realize the goal of achieving a completely optimized ANN. The results obtained from the 'Global Best' and the 'Local Best' optimisation approaches suggest that the 'Global Best' approach for adaptive optimisation of ANNs is more successful in obtaining higher accuracy levels. When considering the application case studies, the 'Global Best' approach has achieved a maximum classification accuracy of 97.33% for the Iris Data classification, and 97.72% accuracy on the full data set of the Ionosphere Data classification while achieving a classification accuracy of 94.70% on the test data set of the same case study. When compared with previous research work which has been carried out on the same case studies, the above mentioned accuracy values prove to be better than nearly all of the past results. Therefore it can be concluded that the 'Global Best' approach has the potential to obtain a structurally optimized neural network.

[9] Author here discuss on swarm intelligence, a new training strategy for neural networks is presented in this paper. Accounting for uncertainty in measurements, particle swarm optimization (PSO) approaches using interval and fuzzy numbers are developed. Applications are focused on the description of time-dependent material behavior with recurrent neural networks for uncertain data within interval and fuzzy finite element analyses. Network training with PSO allows to create special network structures with dependent parameters in order to consider physical boundary conditions of investigated materials. In this paper, a new training strategy for artificial neural networks is presented. It is based on swarm intelligence. PSO approaches for interval and fuzzy numbers are developed accounting for uncertainty in measurements. These approaches have the flexibility of modifying all parameters during training of recurrent neural networks. Additionally, special network structures can be created, which is important for using neural networks as constitutive models. An application for time-dependent material behavior is presented. Results of verifications with a fuzzy fractional Newton element and a 3D linear elastic material model show high Approximation quality of the developed neural network approaches. The new approaches can be applied to measured interval and fuzzy data. Recurrent neural networks for uncertain data can be utilized as constitutive models within interval, fuzzy, and fuzzy stochastic finite element analyses.

[10] This paper has investigated the use of the ABC algorithm for training neural networks, and shown how it can be optimized to give better results than those found in previous studies. However, in most cases, the best ABC generalization performance levels obtained are not significantly different to standard BP that has been properly optimized for the given problems. The BP parameters were first optimized "by hand" in the same way as the ABC, and then by simulated evolution by natural selection. First, with one standard BP learning rate for the whole network, the by-hand and evolutionary optimization results were not significantly different, with the ABC performing significantly better than BP for one dataset, and not significantly different for the other five. Then with two evolved BP learning rates, one for each network layer, it was found that the

results for some datasets were significantly better than with only one learning rate, and that the BP performances were significantly better than the ABC for two datasets, and not significantly different for the other four.

III. Research Motivation

From the past years there is medical science interesting domain for researcher, due to large number of population are affected from its. In our country most of the peoples are not able to get medical treatment on time at everywhere, or most of the peoples suffer from the various types of diseases but not getting any prevention to recover from diseases. To cover large number of peoples we need rich infrastructure such as large number of hospitals, medical lab, equipment, diagnostic tools, machinery, huge memory to stored large number of information, available data everywhere at each time. There are also the tools and techniques for the medical diseases diagnosis are play very vital role in this sector. There is various techniques such as data mining tools which further classified the data in the supervised and unsupervised manner, some evolutionary techniques such as genetic algorithm mostly used for the big data analytics, neural network classifier, support vector machine, decision tree classifier, rule based classifier and some swarm intelligence family methods for the optimization such as ant colony optimization, particle swarm optimization and honey bee classification. Data mining provides various models for the diagnosis of the disease such as Supervised learning, Unsupervised learning, Ensemble learning and Hybrid classification. Support vector machine commonly used to solve prediction and classification problems in efficient way due to its automatic learning system.

IV. Proposed Work

The terms data mining, patent mining, text mining and visualization are employed for the processing of the documents. This chapter will try to give some explanations of the terms and explain why “data mining” was chosen for the title of the study. Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Clustering is a division of data into groups of similar objects.

Data mining gives various types of classification algorithm, according to the diversity of data and variety of data. The variety of data induced the problem of classification issue and degreed the performance of classification algorithm. The classification technique gives different types of algorithm such as support vector machine, decision tree, KNN and ensemble based classifier. Now a day’s ensemble based classifier used for the process of classification. The ensemble based classifier used three types of ensemble technique, bagging, boosting and random forest. The all three technique of ensemble proceed in different manners. Some authors also used clustering technique for the process of ensemble classifier. The feature attribute plays major role in classification technique, the diverse feature creates many problems related to the process of classification such as outlier, and boundary value and core point, without elimination of these problems the classification ratio can’t improve. For the minimization and removal of such types of problem used feature optimization process. The process of feature optimization gives the value of minimization of maximum value of feature attribute. For the optimization of feature attribute used improved genetic algorithm, the improved genetic algorithm defines the fitness constraints for the selection of feature attribute for the process of classification. Good ensemble methods are that in which each individual classifier are accurate and diverse.

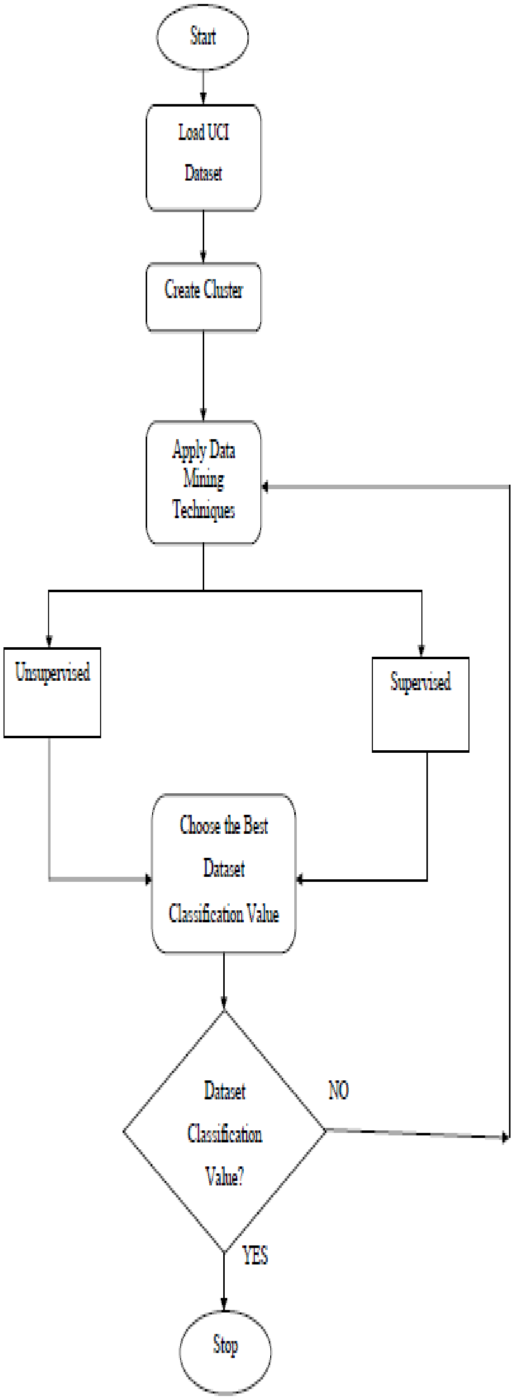
Start
Load UCI Dataset
Create Cluster
Apply Data Mining Techniques
Unsupervised
Supervised
Choose the Best Dataset Classification Value
Dataset Classification Value?
NO
YES
Stop

Fig 1: Proposed model for the Dataset classification.

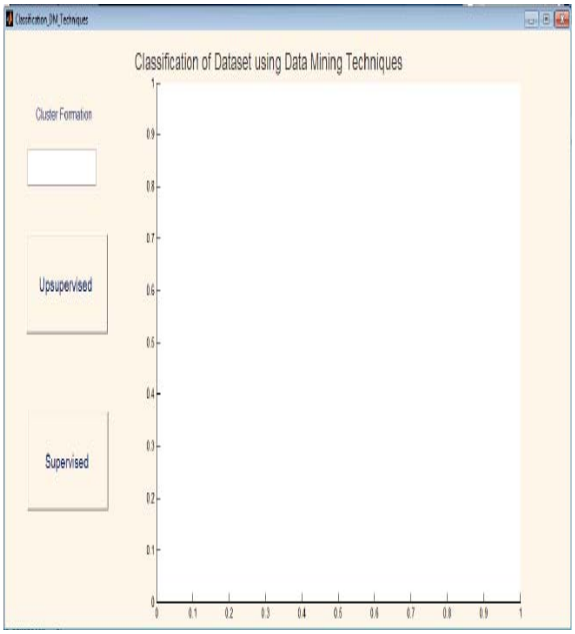

Fig 2: This windows shows that the our proposed work simulation environment.

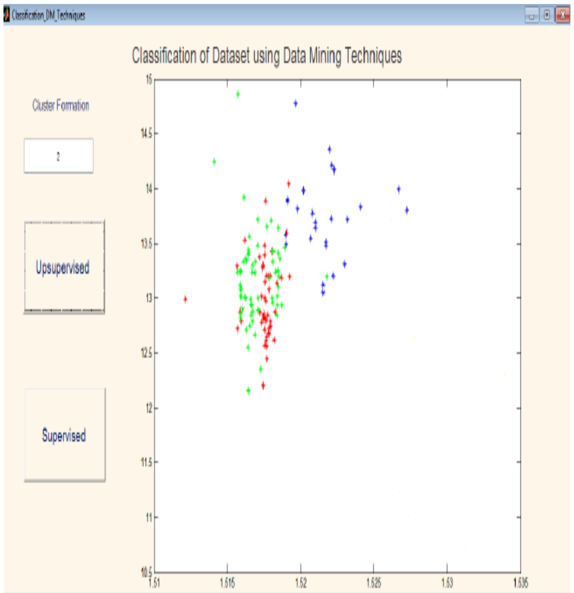

Fig 3: This windows show that the result of Un-supervised methods in a graphical user environment with accuracy in the experimental process using Iris dataset.

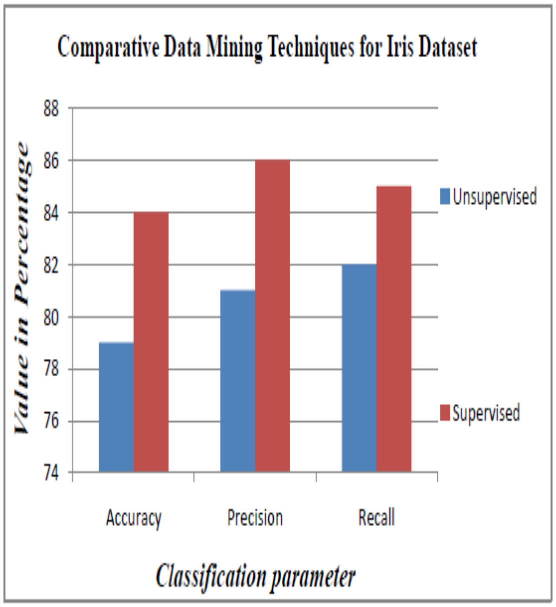

Fig 4: Shows that the comparative result for Data mining techniques using Iris dataset for the both unsupervised and supervised techniques for the performance parameter such as accuracy, precision and recall.

V. Conclusion

In this paper, we present the comparative result analysis study for the physical and life data such as Iris datasets is performed. This process of disease diagnosis of various dataset is done by using data mining techniques such as unsupervised learning and supervised learning; here our proposed method supervised learning gives better results than the existing techniques.

References:

[1] Liling Li, Sharad Shrestha, Gongzhu Hu, “Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques”, IEEE, 2017. pp 363-369.

[2] Michael J. Paul, Abeed Sarker, John S. Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L. Smith,Graciela Gonzalez, “Social Media Mining For Public Health Monitoring And Surveillance”, Pacific Symposium On Biocomputing 2016. Pp 468-477.

[3] Mohammed G. Ahamad, Mohammed F. Ahmed And Mohammed Y. Uddin, “Clustering As Data Mining Technique In Risk Factors Analysis Of Diabetes, Hypertension And Obesity”, European Journal Of Engineering Research And Science, 2016. Pp 88-94.

[4] S M.Inzalkar, Jai Sharma, “A Survey On Text Mining- Techniques And Application”, International Journal Of Research In Science & Engineering, 2016. Pp 1-9.

[5] T. Sajana, C. M. Sheela Rani And K. V. Narayana, “ A Survey On Clustering Techniques For Big Data Mining”, Indian Journal Of Science And Technology, 2016. Pp 1-12.

[6] Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, And Guang-Zhong Yang,” Big Data For Health”, Ieee Journal Of Biomedical And Health Informatics, Vol. 19, 2015. Pp 1193-1206.

[7] Dr. S. Vijayarani,.Dhayanand, “ Data Mining Classification Algorithms For Kidney Disease Prediction”, International Journal On Cybernetics & Informatics, 2015. Pp 13-27.

[8] Katrina Sin, Loganathan Muthu, “ Application Of Big Data In Education Data Mining And Learning Analytics – A Literature Review”, Ictact Journal On Soft Computing, 2015. Pp 1035-1050.

[9] Sushilkumar Kalmegh, “Analysis Of Weka Data Mining Algorithm Reptree, Simple Cart And Randomtree For Classification Of Indian News”, International Journal Of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015. Pp 438-457.

[10] Lambodar Jena, Narendra Ku. Kamila, “ Distributed Data Mining Classification Algorithms For Prediction Of Chronic- Kidney-Disease”,

International Journal Of Emerging Research In Management &Technology, 2015. Pp 110-119.

[11] Ashish Dutt, Maizatul Akmar Ismail, And Tutut Herawan, “A Systematic Review On Educational Data Mining”, Ieee, 2017. Pp 15991-1604.

[12] Basheer Mohamad Al-Maqaleh, Ahmed Mohamad Gasem Abdullah “Intelligent Predictive System Using Classification Techniques For Heart Disease Diagnosis” International Journal Of Computer Science Engineering, 2017. Pp 145-151.

[13] Arijit Ukil, Soma Bandyopadhyay, Chetanya Puri And Arpan Pal “Heart-Trend: An Affordable Heart Condition Monitoring System Exploiting Morphological Pattern”, Ieee, 2016, Pp 6260-6264.

[14] Theresa Princy. R And J. Thomas “Human Heart Disease Prediction System Using Data Mining Techniques”, Iccpct, 2016, Pp 1-5.

[15] Aigerim Altayeva, Suleimenov Zharas And Young Im Cho “Medical Decision Making Diagnosis System Integrating K-Means And Naïve Bayes Algorithms”, Iccas, 2016, Pp 1087-1092.

[16] Michael K. K. Leung, Andrew Delong, Babak Alipanahi, Brendan J. Frey, “Machine Learning In Genomic Medicine: A Review Of Computational Problems And Data Sets”, Ieee Vol-104, 2016. Pp 176-197.

[17] Majid Ghonji Feshki And Omid Sojoodi Shijani “Improving The Heart Disease Diagnosis By Evolutionary Algorithm Of Pso And Feed Forward Neural Network”, Ieee, 2016, Pp 48-53.

[18] Arijit Ukil, Soma Bandyopadhyay, Chetanya Puri And Arpan Pal “Heart-Trend: An Affordable Heart Condition Monitoring System Exploiting Morphological Pattern”, Ieee, 2016, Pp 6260-6264.

[19] Theresa Princy. R And J. Thomas “Human Heart Disease Prediction System Using Data Mining Techniques”, Iccpct, 2016, Pp 1-5.

[20] Aigerim Altayeva, Suleimenov Zharas And Young Im Cho “Medical Decision Making Diagnosis System Integrating K-Means And Naïve Bayes Algorithms”, Iccas, 2016, Pp 1087-1092.

[21] Zhenyundeng, Xiaoshuzhu , Debocheng, Mingzong, Shichaozhang, “Efficient Knn Classification Algorithm For Big Data”, Elsevier Ltd. 2016, Pp 143-148.

[22] Anna L. Buczak, Erhan Guven, “A Survey Of Data Mining And Machine Learning Methods For Cyber Security Intrusion Detection”, Ieee Communications Surveys & Tutorials, 2016. Pp 1153-1176.

[23] Jean Damascene Mazimpaka, Sabine Timpf, “Trajectory Data Mining: A Review Of Methods And Applications”, Journal Of Spatial Information Science, 2016. Pp 61-99.

[24] Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, S. S. Iyengar, “Computational Health Informatics In The Big Data Age: A Survey”, Acm Computing Surveys, Vol. 49, 2016. Pp 1-36.

[25] Sudha Ram, Wenli Zhang, Max Williams, Yolande Pengetnze, “Predicting Asthma-Related Emergency Department Visits Using Big Data”, Ieee Journal of Biomedical And Health Informatics, Vol. 19, 2015. Pp 1216-1218.