Materials Today: Proceedings 51 (2022) 2198–2201

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Implementation of machine learning techniques for disease diagnosis

Shachi Mall [a,*], Ashutosh Srivastava [b], Bireshwar Dass Mazumdar [c], Manmohan Mishra [d], Sunil L. Bangare [e], A. Deepak [f]

[a] Department of Computer Science & Engineering, Institute of Technology & Management, GIDA, Gorakhpur, Uttar Pradesh, India
[b] Systems Engineering, Department Of electrical engineering, IIT- BHU, Varanasi, India
[c] Department of Computer Science &Engineering, IERT, Prayagraj 211002, India
[d] Department of Computer Application, United Institute of Management, Prayagraj 211010, India
[e] Department of Information Technology, Sinhgad Academy of Engineering, SavitribaiPhule Pune University, Pune, India
[f] Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India

ARTICLE INFO

Article history:
Received 24 September 2021
Received in revised form 7 November 2021
Accepted 15 November 2021
Available online 10 December 2021

Keywords:
Data Mining
Machine Learning
Classification
Decision Tree
Prediction
Heart Disease

ABSTRACT

Recently, data mining and machine learning techniques have found widespread use in the field of healthcare. The objective of this study is to develop an automated method for diagnosing illnesses. A Fuzzy logic-based random forest approach and a thorough examination of the patient's medical records are used to diagnose the disease. Clinical diagnosis is performed with the aid of a doctor's expertise and understanding in traditional healthcare. It is more challenging to provide good healthcare in rural and remote areas because patients are more likely to travel a long distance to visit a specialist. Because the number of medical practitioners and facilities in these areas is limited, providing an expert diagnosis in a fair period of time is challenging. The problem can be solved by using expert systems for disease diagnosis that employ data mining techniques and fuzzy logic. Decision trees are often used in machine learning to predict outcomes. Fuzzy datasets are an excellent choice for describing medical facts and expert opinions. Fuzzy decision trees build simple decision trees using fuzzy input. In this proposed system, an expert system that diagnoses disease using a random forest algorithm and fuzzy decision trees is provided. The fuzzy decision trees increase the accuracy of the diagnostic system. On the UCI repository, the proposed method is assessed and found to be more efficient in sickness prediction than current strategies. Classification accuracy has risen as temporal complexity has decreased.

Selection and peer-review under responsibility of the scientific committee of the International Conference on Advances in Materials Science

1. Introduction

Today's world faces a greater threat from various illnesses, which are spreading at an alarming rate. Every year, the number of illnesses diagnosed by health-care agencies grows. For any sickness, predicting or identifying the condition at an early stage would allow individuals to receive the best possible treatment. As a result, disease prediction has become a more vital duty to assist medical practitioners in delivering effective therapy to patients. Previously, the medical practitioner identified the condition based on the symptoms, and this was also done at random. Because of the ambiguity associated in identifying a specific ailment, the medical practitioner offers therapy depending on the symptoms and will alter until it is cured. Because the symptoms of many diseases are similar, the time necessary for therapy is similarly lengthy. This difficulty may be met with data mining tools and techniques that are effective in generating algorithms that can be used to a wide range of situations [1,2]. Table 1.

Data mining in health care is designed to decrease errors in sickness prediction, medicine prescription, individual perception, use of traditional drugs, early disease detection, and avoidance of fraudulent medical insurance claims. Data mining techniques might be used to identify ailments including heart disease, cancer, and bone fractures, among others. It incorporates methods from a variety of areas, including as machine learning, database structure, and information retrieval. The work of data mining is doing an

* Corresponding author.
E-mail addresses: shachimall@gmail.com (S. Mall), ashutosh.rs.eee@iitbhu.ac.in (A. Srivastava), bireshwardm@gmail.com (B.D. Mazumdar), mishramanmohan@gmail.com (M. Mishra), sunil.bangare@gmail.com (S.L. Bangare), deepakarun@saveetha.com (A. Deepak).

https://doi.org/10.1016/j.matpr.2021.11.274
2214-7853/
Selection and peer-review under responsibility of the scientific committee of the International Conference on Advances in Materials Science

Table 1
Accuracy of Machine Learning Classifiers.

Algorithm Name	Accuracy In %
Fuzzy Random Forest	91.7
C4.5	73.27
ID3	66.29

automatic or semi-automatic review of massive volumes of data in order to discover previously unknown intriguing patterns [3,4].

Clustering is a data mining technique used to categorize objects into similar groups. Using the crisp clustering technique, an item can become a member of just one cluster, and the membership can be entirely inclusive or exclusive. To give a clear grasp of clustering ideas, this article provides a review of various clustering techniques as well as demonstrations of the k-means algorithm and replicating neural networks. However, object segmentation varies under the humanistic method since the item may be a member of more than one cluster at the same time, with varying degrees of connectivity. Using fuzzy clustering algorithms, data mining segmentation may be performed exactly the way individuals organize things. The fundamental idea behind fuzzy clustering is the non-unique grouping of data into a collection of clusters with membership values ranging from 0 to 1. The non-zero membership numbers, up to one, show how much the data point belongs to a cluster.

Other crisp data mining approaches, such as association rule mining, classification, and prediction, employ intervals with crisp bounds to categorize quantitative features. The intervals are used in conjunction with traditional methods to discover patterns involving these characteristics. Overestimation of boundary situations is an issue that plagues such approaches. The shift from one segment to another in natural segmentation of quantitative values is gradual rather than sudden. Only fuzzy sets and linguistic variables may be used to model these natural segments with overlapping borders [5,6,7].

The existence or absence of items in a transaction is evaluated in fuzzy influence rule analysis to produce association rules. Association rule mining approaches do not take into account objects' significance, profit, or the quantity of things in a transaction. The weight of an item, a fuzzy notion, might be used to indicate the importance of an item in a transaction. The unsupervised cluster estimation technique is used to divide the quantitative characteristics into natural language chunks, which incorporates this notion. The values of the resultant fuzzy linguistic segments are utilized to create fuzzy association rules.

2. Related work

Data mining tools in healthcare are proving to be a significant resource for both researchers and patients. Using the medical database, this aids in illness prevention, diagnosis, and treatment.

Outlier identification, as proposed by [8], is a prominent study field in data mining from big datasets and is a critical task in a variety of application domains. Outliers, formerly classified as noisy data, have emerged as a key emphasis in data mining applications. Outlier detection is useful in detecting unrecognized and unexpected data. Data preprocessing covers a wide range of data quality issues, with an emphasis on outliers and noise. The primary goal of this step is to remove items that impede data analysis.

Authors in [9] presented two methods, Distance-Based outlier detection and Cluster-Based outlier algorithm, for finding and eliminating outliers in a health care dataset using outlier score. Experiments were carried out using three built-in health care datasets, and the findings revealed that the cluster-based outlier identification algorithm outperformed the distance-based outlier detection algorithm in terms of accuracy. Author [10] presented a survey that provided a complete and articulated summary of outlier classifications in different temporal data systems, techniques implemented to identify and remove them from the data base, and set-ups with suitable detection techniques implemented in specific applications.

Author [11] created Feature-Rich Interactive Outlier Detection (FRIOD). The suggested outlier identification technique allows for user input at all critical phases of the process. It included dense cell selection, location-aware distance thresholding, and top outlier validation. Another data mining approach with several applications, such as outlier identification, is data clustering.

Author [12] concentrated on data clustering and outlier identification. The authors presented a modified K-means type method that included an additional "cluster" without taking outliers into account while determining the cluster center. The algorithm performed better when tested with actual and fake data.

Author in [13] created a parallel processing system for fuzzy associative categorization using Open Computing Language. The suggested technique used a CPU-GPU implementation to identify infectious illness occurrences, notably influenza, with the use of previously collected disease and environmental data. The study also compared the performance of the Hybrid technique to that of the other methods.

Author in [14] used association rule mining on Electronic Medical Records (EMR) to identify sets of risk variables and their corresponding subpopulations that correlate to individuals at high risk of acquiring diabetes. Due to the high dimensionality of EMRs, association rule mining produces a massive collection of rules that are required for straightforward clinical usage. Researchers [15] used linguistic fuzzy rule-based classification to predict the likelihood of a patient developing cardiac disease. The framework adopted diagnoses and also offers a brief analysis of the decision, allowing clinicians to get information from the existing history of patient data.

Authors in [16] investigated the application of a general-purpose, supervised feed forward neural network with one hidden layer, namely the Radial Basis Function (RBF) neural network. It is adaptive and employs a reduced number of locally adjusted components. With Wisconsin breast cancer data, the RBF neural network's performance was evaluated and compared to the most widely used Multilayer Perceptron (MLP) network model and the traditional logistic regression. According to the findings, the sensitivity and specificity of both neural network models outperformed logistic regression in terms of predictive power. Even when tested on a different dataset, the neural network models outperformed the logistic regression. The comparative study given for analysis shows that the RBF neural network has strong predictive skills and that the time consumed by RBF is less. However, RBF has drawbacks such as sensitivity to dimensionality and difficulties dealing with big datasets.

Researchers in [17] sought to evaluate the efficacy of multiple data mining categorization approaches utilizing three distinct machine learning tools across four different healthcare datasets. The criteria utilized are the percentage of accuracy and error rate of each categorization technique used. The trials are carried out with the help of the 10 fold cross validation approach. The approach that is best suited to a given dataset is chosen based on its greatest classification accuracy and lowest error rate. According to the experimental results, different categorization algorithms react differently on different datasets depending on the nature of their characteristics and size. The classification approach with the highest accuracy rate and lowest error rate across a dataset was chosen as the best classification technique for that dataset.

In the example of forecasting diabetes in patients, [18] developed a Decision support system that uses the strengths of both OLAP and Data mining to anticipate the future condition and provide relevant information or effective decision making. In addition, the system contrasted the outcomes of the ID3 and C4.5 decision tree algorithms.

On several breast cancer datasets, author in [19–22] investigated the performance of decision tree classifier-CART with and without feature selection in terms of accuracy, time to create a model, and tree size. The experimental findings indicated that feature selection, a pre-processing approach, substantially improves classification accuracy. They determined that the adoption of any of the feature selection methods improved classifier accuracy over the classifier accuracy without feature selection. Various tests were carried out on three distinct breast cancer datasets in order to determine if the same feature selection approach can lead to the greatest accuracy for diverse datasets in the same domain. According to the findings, a given feature selection may not result in the greatest accuracy for all breast cancer datasets. The most appropriate feature selection technique for a given dataset is determined by the amount of attributes, attribute types, and occurrences. As a result, each additional dataset is evaluated, one must experiment with multiple feature selection approaches to discover the optimal one to improve classifier performance rather than just evaluating the previously proven one linked to the same domain. Once the optimal feature selection technique for a given dataset has been found, it may be utilized to improve classifier accuracy [23–25].

3. Proposed methodology and result analysis

Fig. 1 depicts a framework for illness prediction. This framework takes as input a set of student performance data. This student data set has been preprocessed to reduce noise and make the input data set consistent. The input data set is then subjected to different machine learning techniques such as Fuzzy Random Forest, ID3, and C4.5. Data classification is carried out. The classification results of several methods are compared.

Fuzzy Random Forest Algorithm is as follows:

Input: Dataset

Output: Diagnosed disease data

Step 1: Set Sum of records = m, Sum of attributes = p

Step 2: Let 'm' conclude the quantity of attributes at a node of a decision tree

Step 3: For every decision tree do

3.1 Select randomly the subgroup of coaching information that characterizes the M records and

3.2 uses the rest of information to measuring mistake of the tree,

3.3 Calculate threshold β for fuzzy decision tree.

Step 4: Generate fuzzy decision tree

3.1. While there exist candidate nodes DO

4.1.1 choose one amongst them employing a search scheme

4.1.2 make it's sub node. Sub-nodes get-together the leaf threshold are leveled as leaf-nodes

4.1.3 the remaining sub-nodes are viewed as new candidate nodes

Step 5: End

This analytical work makes use of the UCI machinery heart disease data collection [23]. The 303-record Cleveland database is utilized as input in the ID3, C4.5, Fuzzy Random Forest method. Fig. 2 depicts the level of accuracy attained. It is clear that the fuzzy random forest algorithm is outperforming ID3 and C4.5 algorithm as far as accuracy is concerned.

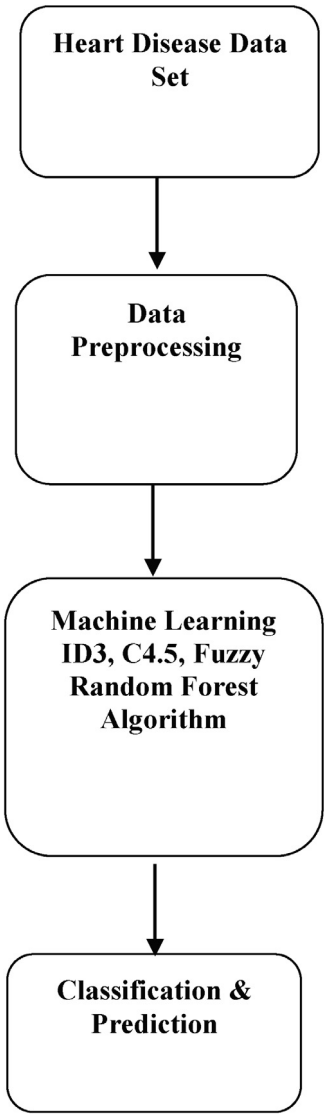

Fig. 1. Framework for disease prediction.

4. Conclusion

In machine learning, decision trees are frequently used to predict outcomes. Fuzzy datasets are a great way to describe medical facts and expert opinions. Fuzzy decision trees use fuzzy input to construct basic decision trees. An expert system that diagnoses illness utilizing a random forest algorithm and fuzzy decision trees is presented in this suggested system. The fuzzy decision trees improve the diagnostic system's accuracy. The suggested technique is evaluated on the UCI repository and shown to be more efficient in illness prediction than current solutions. From experimental results, it is clear that the fuzzy random forest algorithm is outperforming ID3 and C4.5 algorithm as far as accuracy is concerned. The proposed fuzzy random forest algorithm has achieved an accuracy rate of 91.7 percent for heart disease classification and prediction.

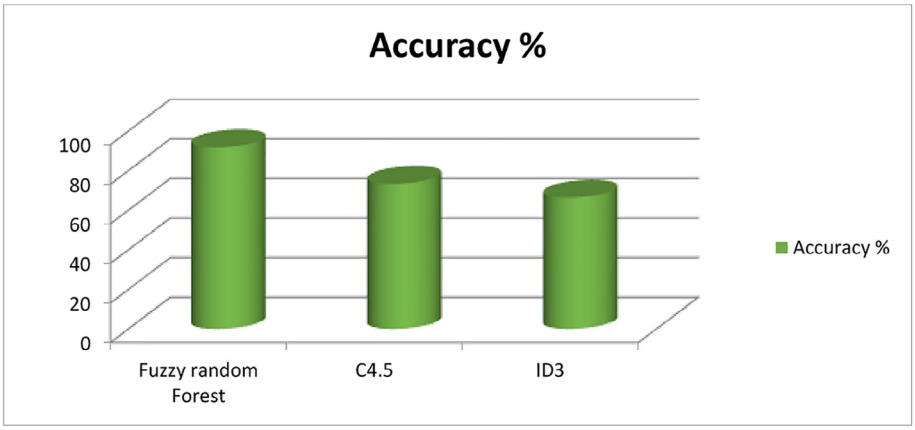

Fig. 2. Disease Data Classification Results.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Shachi Mall: Conceptualization. **Ashutosh Srivastava:** Data curation, Visualization. **Bireshwar Dass Mazumdar:** Methodology, Writing – review & editing. **Manmohan Mishra:** Investigation, Writing – original draft. **Sunil L. Bangare:** Validation, Supervision. **A. Deepak:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Singh, and, R., Kumar,, Heart Disease Prediction Using Machine Learning Algorithms, 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452–457, 10.1109/ICE348803.2020.9122958.
[2] V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.
[3] P. Motarwar, A. Duraphe, G. Suganya, M. Premalatha, Cognitive Approach for Heart Disease Prediction using Machine Learning, 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1–5, 10.1109/ic-ETITE47903.2020.242.
[4] Ravi Manne, Snigdha Kantheti and Sneha Kantheti 6 11 101 108.
[5] R.J.P. Princy, S. Parthasarathy, P.S. Hency Jose, A. Raj Lakshminarayanan, S. Jeganathan, Prediction of Cardiac Disease using Supervised Machine Learning Algorithms, 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 570–575, 10.1109/ICICCS48265.2020.9121169.
[6] Myla M Arcinas, "A Blockchain Based Framework For Securing Students Educational Data", Linguistica Antverpiensia, ISSN: 2295-5739 Volume 2021, Issue 2, pp. 4475-4484.
[7] Myla M Arcinas Guna Sekhar Sajja Shazia Asif Sanjeev Gour Ethelbert Okoronkwo The Role of Data Mining in Education for Improving Students Performance for Social Change Turkish Journal of Physiotherapy and Rehabilitation, ISSN 32 3 2651–4451, 6519 6526.
[8] Rashi Bansai, Nishant Gaur, Shailendra Narayan Singh, Outlier Detection: Applications and techniques in Data Mining, IEEE Conference on Cloud System and Big Data Engineering, 2016, pp. 373–377.
[9] A. Christy, G. Meera Gandhi, S. Vaithyasubramanian, Cluster Based Outlier Detection Algorithm for Healthcare Data, Elsevier 50 (2015) 209–215.
[10] Manish Gupta, Jing Gao, Charu C. Aggarwal, Jiawei Han, Outlier Detection for Temporal Data: A Survey, IEEE Trans. Knowl. Data Eng. 26 (9) (2014) 2250–2267.
[11] Xiaodong Zhu, Ji Zhang, Hongzhou Li, Philippe Fournier-Viger, Jerry Chun-Wei Lin, Liang Chang, FRIOD: A Deeply Integrated Feature-Rich Interactive System for Effective and Efficient Outlier Detection, Access IEEE 5 (2017) 25682–25695.
[12] Guojun Gan, Michael Kwok-Po Ng, k-means clustering with outlier removal, Pattern Recogn. Lett. 90 (2017) 8–14.
[13] Erhan Guven, Anna L. Buczak, An OpenCL Framework for Fuzzy Associative Classification and its Application to Disease Prediction, Procedia Comput. Sci. 20 (2013) 362–367.
[14] Gyorgy J. Simon, Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M. Regina Castro, Peter W. Li, Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus, IEEE Transactions on Knowledge And Data Engineering 27 (1) (2015) 130–141.
[15] José Antonio Sanz, Mikel Galar, Aranzazu Jurio, Antonio Brugos, Miguel Pagola, Humberto Bustince, Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system, Appl. Soft Computing 20 (2014) 103–111.
[16] J. Padmavathi, A Comparative study on Breast Cancer Prediction Using RBF and MLP, Int. J. Sci. Eng. Res. vol. 2, no. 1 (2011).
[17] Shelly Gupta, Dharminder Kumar, Anand Sharma, Data mining Classification techniques applied for breast cancer diagnosis and prognosis, Indian Journal of Computer Science and Engineering (IJCSE), ISSN 0976–5166 2 (2) (2011).
[18] Chetan M. Thakar Shailesh S. Parkhe Ankit Jain Khongdet Phasinam G. Murugesan Randy Joy Magno Ventayen 2021 10.1016/j.matpr.2021.06.272.
[19] S.T. Jagtap, C.M. Thakar, O., El, imrani,, K., Phasinam,, S., Garg, and, R., J., M., Ventayen,, A Framework for Secure Healthcare System Using Blockchain and Smart Contracts, 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 922–926, 10.1109/ICESC51422.2021.9532644.
[20] Rupa Bagdi & Pramod Patil, Diagnosis of Diabetes Using OLAP and Data Mining Integration, International Journal of Computer Science & Communication Networks 2 (3) (2012) 314–322.
[21] D. Lavanya, K. Usha Rani, Analysis of Feature Selection With Classfication: Breast Cancer Datasets, Indian Journal of Computer Science and Engineering (IJCSE), ISSN 0976–5166, 2, 2011, p. 5.
[22] P. P. Shinde, K. S. Oza and R. K. Kamat, "Big data predictive analysis: Using R analytical tool," 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2017, pp. 839-842, doi: 10.1109/I-SMAC.2017.8058297.
[23] P.P. Shinde, K.S. Oza, R.K. Kamat, Leveraging cell phones for surveillance, International Conference on Intelligent Sustainable Systems (ICISS) 2017 (2017) 6–9, https://doi.org/10.1109/ISS1.2017.8389337.
[24] K. Arumugam Mohd Naved Priyanka P. Shinde Orlando Leiva-Chauca Antonio Huaman-Osorio Tatiana Gonzales-Yanac 2021 10.1016/j.matpr.2021.07.361
[25] https://archive.ics.uci.edu/ml/datasets/heart+disease.