A novel approach for Liver disorder Classification using Data Mining Techniques

A.S.Aneeshkumar[#1], Dr. C.Jothi Venkateswaran[*2]

[1]*Research Scholar, P.G. & Research Dept. of Computer Science, Presidency College, Chennai, India*
aneesh_kumar777@yahoo.com
[2]*Associate Professor & Head, P.G. & Research Dept. of Computer Science, Presidency College, Chennai, India*

Abstract— Data mining is an integrated platform for all other soft computing techniques and which is used to identify the expected or probable values from a large storage by using computational algorithms. This paper describes the categorization of liver disorder through feature selection and fuzzy K-means classification. In any of the medical diagnosis activity, some features may directly or indirectly influence and some others may not influence. So the process of influenced attribute subset selection is an essential factor for more accurate prediction. Suitability of the dataset and the selection of algorithm are the two key factors for any predictive analysis.

Keywords— Snake peel enabled Hybrid Ant colony optimization and Genetic Algorithm, Fuzzy K-means classification

1. Introduction

Discussion on Data mining and its applications are widely occurred in all other areas of engineering, science, medicine and management. The word data mining, refers the retrieval of expected result or information from the large storage.

Liver disorder is considered as one of the major problem which unknowingly growing in modern society. The depth and severity of the disease will be rectified in matured stages only, because the early symptoms are not much more effective. So the identification of liver disorder in initial phase is a vital activity to make awareness about the disease. The liver problems may vary from simple allergy of toxic to liver cirrhosis. In this work we attempt to classify five types of liver disorders in its initial stage with physical, psychodynamic and clinical attributes. Normally this is a complex task to medical practitioners to determine these five types at the beginning stage.

The collected data from a reputed hospital in southern region of Tamil Nadu is used here for this study. The dataset consists of 48 attributes and 6078 instances.

Data mining algorithm is very suitable for disease recognition and treatment prescription. But the difficulty is in the selection of appropriate algorithm to be trained for that particular dataset. Because each of the data mining approach will deal with different methodology and so the significance of training data is the key issue in algorithm selection.

2. Feature Selection

Feature selection is a process of identifying a minimum subset from the multi-dimensional origin of features [1]. The best possible subset is selected with relevancy and redundancy phase. A factor is considered to be relevant if that has an influence over the decision. Otherwise, it named as irrelevant. However, redundancy is measured from the correlation of the attributes to achieve overall decision [2].

From the results of previous works, we identified that the accuracy is satisfactory with some selected attributes, but the result is not that much effective with these 48 attributes. In this situation we understood the necessity of feature subset selection and so we choose thirty attributes which are directly or indirectly influencing the disease.

As part of feature selection, we used Snake peel enabled hybrid Ant colony optimization and genetic algorithm. The performance of this algorithm reflects in the prediction accuracy of the disease. The selected attributes are considered as the input for develop and train a suitable classification model.

3. Fuzzy K-means classification

K-means clustering is a known unsupervised learning technique in data mining. In clustering, n samples are divided into k categories [3], where each input attribute belongs to one cluster and it may not be a part of other clusters [4]. If same attribute value is seen in multiple clusters, it will be difficult task to classify the liver disorder type with that attribute. In our study, the feature selection used to identify influencing attributes but even the problem with scattering values of symptoms are appeared in multiple clusters and so the fuzzy classification on clusters is attempted. In fuzzy, the symptoms belonging to the clusters doesn’t carry more than one crisp values and so it may fall between 0 and 1 for the classification [5].

3.1 K-means Clustering

In K-mans clustering, n observed attributes of liver disorders are partitioned into k clusters and determined its most appropriate centroids. We are having five clusters in our research work, known as AFLD (Alcoholic Fatty Liver Disorder), Hep-F (Hepatitis- Female), Hep-M (Hepatitis - Male), NAFLD-F (Non-Alcoholic Fatty Liver Disorder - Female), and NAFLD-M (Non-Alcoholic Fatty Liver Disorder-Male) .

The following steps are carried out for K-means clustering.

1. Initialize starting condition by defining the number of clusters and randomly select the initial cluster centers. Euclidean distance is used to observe the distance between the attributes,

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2}$$

$$\text{i.e} \quad = \sqrt{\textstyle\sum_{i=1}^{n} (a_i - b_i)^2}$$

where a, b are the two points in Euclidian space and d represents distance

2. Generate a new partition by assigning each data point to the nearest cluster center.
3. Recalculate the centers for clusters receiving new data points and for clusters losing data points.
4. Repeat the steps 2 and 3 until a distance convergence criterion is met.

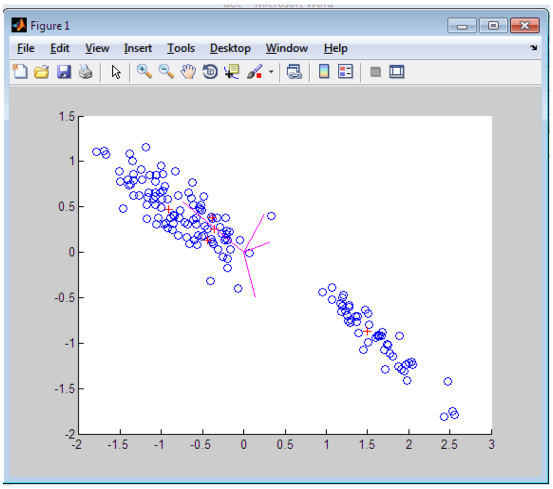


Fig.1: Clustering Diagram

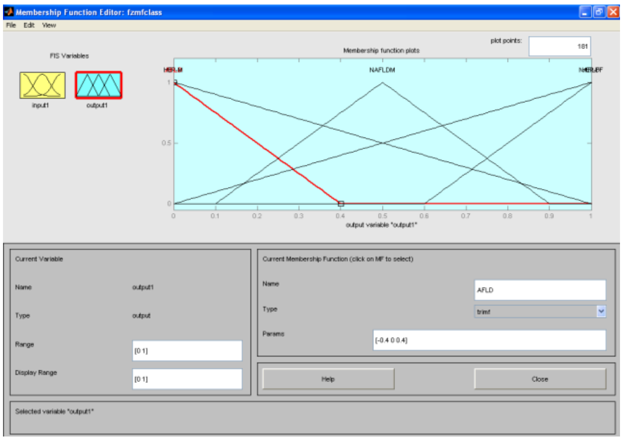
3.2 Fuzzy Classification

Fuzzy Logic is always used as the fundamental model of reasoning, which gives an approximation rather than exact value and so it is very much closer to human reasoning for in real world problems [6]. The achieved performance is computed with the significance of member function and membership rules. The two well used fuzzy Interface systems are Mamdani and Tagaki-Sugeno [7]. The Mamdani fuzzy inference system is largely recognized and well-matched to individual cognition because of its distinctiveness [8]. The Tagaki-Sugeno fuzzy inference system performs well with linear techniques and assures the output continuity of surface [9]. But in case of Tagaki-Sugeno model, there are some difficulties such as dealing with multi-parameter evaluation, weight assignment for inputs and fuzzy rule generation. Mamdani fuzzy interface system expresses its understandability and ultimateness in result and so here we adopt it for this classification study. The expression for fuzzy logic is defined as,

$$D = \{(x, \mu_D(x)) I_x \in X, \mu_D(x) \in [0,1]\}$$

Where X represents the universal set, x is an element which belongs to X, D is a fuzzy subset in X, $\mu_D(x)$ considered as the membership function of fuzzy set D.

The membership function of Mamdani is using Triangular, Trapezoidal, Gaussian distribution and etc., for mapping the input space into a output space. According to our dataset we choose Triangular membership for the input and output variables.



plot points: 181

Current Variable
Name output1
Type output
Range [0 1]
Display Range [0 1]

Current Membership Function (click on MF to select)
Name AFLD
Type trimf
Params [-0.4 0 0.4]

Selected variable "output1"

Fig.2: Triangular membership function

Triangular membership functions for each type of liver disorder is identified with,

$$triangle(x; a, b, c) = \begin{cases} 0, & x \le a. \\ \dfrac{x-a}{b-a}, & a \le x \le b. \\ \dfrac{c-x}{c-b}, & b \le x \le c. \\ 0, & c \le x. \end{cases}$$

Where a, b and c are the three scalar parameters of vector x.

So here we got the following membership rules from the above diagram.

Table 1: generated Fuzzy rules

Rule 1: If input1 = mf1 and input1!= mf2 and input 1 != mf3 then output is AFLD
Rule 2: If input2 = mf2 and input2 != mf1 and input2 != mf3 then output is Hep-F
Rule 3: If input3 = mf3 and input3 != mf1 and input3 != mf3 then output is Hep-M
Rule 4: If input4 = mf1 and input4 = mf2 and input4 != mf3 then output is NAFLD-M
Rule 5: If input5 = mf2 and input5 = mf3 and input5 != mf1 then output is NAFLD-M

Finally the application of fuzzy over the cluster is used to identify the accuracy of the classified data, where we received above 94 percentage of accuracy in all classes.

4. Results and Discussions

Various results achieved from this work is identified and computed for liver disorder prediction. As part of this, we identified and selected directly or indirectly influencing attributes of liver disorder and which are used to train the constructed novel model of Fuzzy K-means classification. The used novel hybrid approach of feature selection chooses thirty physical and clinical attributes out of 48. In figure 1, we can see the cluster formation with randomly selected centroid. This process and centroid adjustment will repeated until the last instant also be clustered. In figure2, Triangular Mamdani membership function is applied for the five types of liver disorder and identified rules for them are presented in table 1. Then we performed a classification accuracy analysis in these selected classes and obtained better result.

5. Conclusion

According to World Health Organization's report liver disorder is listed in top fifteen life threatening disease as per and its symptoms are hidden in earlier stages. So it is a very difficult task to determine liver disorder initially. In later stages also the determination is happened with multiple diagnosis process only. Developing countries like India; the availability of more diagnosis facilities in remote area are rare. In order to that the basic attributes always shows very much similarity in its range values for multiple diseases. Similarly various liver disorders also share same attribute values and it needs more effort to classify liver disorder type correctly with basic attributes. So Fuzzy based classification gives better performance in these confusing classes and achieved above 94 percentage accuracy for each type of liver disorder.

Acknowledgement

We express our sincere thanks to the Director Dr.(Capt.) K. J. Jayakumar M.S., M.N.A.M.S., F.A.I.S. and Chief Manager Dr. R. Rajamahendran, B.Sc., M.B.B.S., D.M.C.H., D.H.H.M., P.G.D.H.S.C.(Diab.), F.C.D., Sir Ivan Stedeford Hospital, Chennai for providing permission to collect data. We are grateful to the chief Manager for his guidance and also would like to thank other hospital staffs for their valuable suggestions and help throughout this study.

References

[1] H. Liu and H. Motoda, *Feature Selection for Knowledge* Discovery and Data Mining. Boston: Kluwer Academic Publishers 1998.

[2] Majdi Mafarja, Derar Eleyan, "Ant Colony Optimization based FeatureSelection in Rough Set Theory", International Journal of Computer Science and Electronics Engineering (IJCSEE) Volume 1, Issue 2, 2013

[3] C.Kruegel, F.Valeur, G.Vigna,‖Intrusion Detection and Correlation challenges and Solution‖ University of California, Santa Barbara, Springer Science USA, 2005.

[4] Farhad Soleimanian Gharehchopogh, Neda Jabbari, Zeinab Ghaffari Azar, " Evaluation of Fuzzy K-Means And K-Means Clustering Algorithms In Intrusion Detection Systems", International Journal of Scientific & Technology Research Volume 1, Issue 11, December 2012

[5] Vance Faber,‖Clustring and the Continuous K-means Algorithm‖, Los Almas since Number22, pp: 138-144, 1994.

[6] I Elamvazuthi, P.Vasant and J.Webb, "The Application of Mamdani Fuzzy Model for Auto Zoom Function of a Digital Camera, International Journal of Computer Science and Information Security, Volume 6, Issue3, 2009.

[7] Yuanyuan Chai, Limin Jia, and Zundong Zhang, Mamdani Model based Adaptive Neural Fuzzy Inference System and its Application, International Journal of Computational Intelligence 5:1 2009, pp. 22-29.

[8] E.H.Mamdani and S.Assilian, An experiment in linguistic synthesis with a fuzzy logic controller, International Journal of Man-Machine Studies, 7(1):1-13,1975.

[9] T. Takagi and M. Sugeno, Derivation of fuzzy control rules from human operators control actions, Proc. IFAC Symp. on Fuzzy lnformation,Knowledge Representation and Decision Analysis, 55–60, July 1983.

[10] S.Aravindh and G.Michael, "Hybrid of Ant Colony Optimization and Genetic Algorithm for Shortest Path in Wireless Mesh Networks", Journal of Global Research in Computer Science, Volume 3, No. 1, Pp.31-34, January 2012

[11] Shahla Nemati, Mohammad Ehsan Basiri, Nasser Ghasem-Aghaee and Mehdi Hosseinzadeh Aghdam, "A novel ACO–GA hybrid algorithm for feature selection in protein function prediction", Expert Systems with Applications 36, pp.12086–12094 Elsevier 2009.

[12] Zainudin Zukhri and Irving Vitra Paputungan, "A Hybrid Optimization Algorithm based on Genetic Algorithm and Ant Colony Optimization", International Journal of Artificial Intelligence & Applications, Vol. 4, No. 5, pp.63-75,September 2013

[13] Zaiyadi M. F and Baharudin B, "A Proposed Hybrid Approach for Feature Selection in Text Document Categorization", International Scholarly and Scientific Research & Innovation , Vol.4, Issue 12, pp.100-104, 2010.

[14] Ali A Mokdad, Alan D Lopez, Saied Shahraz, Rafael Lozano, Ali H Mokdad, Jeff Stanaway, Christopher JL Murray and Mohsen Naghavi, "Liver cirrhosis mortality in 187 countries between 1980

and 2010: a systematic analysis", European Journal of Medical Research, Volume 12, Issue 9, pp. 145-152. 2014

[15] Alp Aslandogan Y and Gauri A.Mahajani, "Evidence Combination in Medical Data Mining", Proceedings of the International Conference on Information Technology: Coding and Computing IEEE, ITCC- 2004.

[16] Andrea Dimartini, Mary Amanda Dew, Lubina Javed, Mary Grace Fitzgerald and Ashok Jain, "Pretransplant Psychiatric and Medical Comorbidity of Alcoholic Liver Disease Patients Who Received Liver transplant", Psychosomatics 45:6, Page no: 217-523, 2004.

A.S.Aneeshkumar is a Ph.D. Research Scholar in the P.G. & Research Department of Computer Science, at Presidency College (Autonomous), Chennai. He holds a Master Degree in Information Technology from the University of Madras and a Master of Computer Applications degree from Bharathiar University, Coimbatore. He finished M.Phil in Computer Science from Vinayaka Mission University. In addition to that he completed a separate Master degree in Psychology and Criminology. He presented numerous papers at various national and international conferences and have a handful of publications in National and International level Journals with good impact factor. He is having teaching experience of more than five years and four years of Research activities.

Dr. C. Jothi Venkateswaran, Dean of the Post Graduate and Research Department of Computer Science and Applications at Presidency College (Autonomous), Chennai.. He has been serving more than 27 years of teaching experience and more than 13 years of research experience in the field of Data mining and Database Management System. He has published many articles in the National and International Journals with good impact factor and some of them in Scopus index .He presented papers in many Conferences in India and abroad. In addition to that, he produced numerous M.Phil. and Ph.D. under his guidance from various Universities. He is appointed in various positions of Government organizations, Universities, Higher Educational Institutions and Non-Governmental organizations.