International Journal of Advanced and Innovative Research (2278-7844) /
Volume 7 Issue 1

A Heuristic SOM Based Hybrid Approach for Liver Patient Disease Data Classification

Priyanka Mathur[#1], Shweta Sharma[*2], Pankaj Richhariya[#3]

[#]*CSE Department, BITS College, RGPV University*
Bhopal, India

[1]mathurp715@gmail.com

***Abstract*— Data mining is a stream where a large data analysis using processing algorithm can be performed. Data mining help in understanding the processing outcome and thus a utilization of data can be done. Health care is the segment which is sensitive area and a large number of information produce by the health care data reports. Liver Health care analysis helps us to understand the behaviour and diagnose the disease according to requirement. Different data classification approach for classifying the liver health care dataset is proposed by existing authors. Algorithm such as Naïve Bayes, Genetic rule based algorithm, Support vector machine approach and other classification approach is performed by existing researchers. The existing solution still lacking in finding a proper computation efficiency over the health care dataset. In this paper a enhance Heuristic SOM based optimization approach is proposed for health care data analysis. The proposed technique makes use of heuristic optimization and Map use from the SOM ANN algorithm. The proposed technique is experiments using the Weka tool and ILPD (Indian liver patient dataset) taken from UCI is executed. Observed result shows the efficiency of proposed technique over existing solutions..**

***Keywords*— IDLP, SOM, UCI, Data Mining, Healthcare Analysis, Data classification**

I. INTRODUCTION

Liver health care [1] is the medical healthcare stream which leads to damage in human body. Liver health care analysis approach help in understanding of data and further performing the research on it. Data mining classification help in analysing the data and utilizing them for medical cure.

Classification, regression and classification are three approaches of data mining in which instances are grouped into identified classes. Classification is a popular task in data mining especially in knowledge discovery and future plan. It provides the intelligent decision making. Classification not only studies and examines the existing sample data but also predicts the future behaviour of that sample data. It maps the data into the predefined class and groups. It is used to predict group membership for data instances [2]. In Classification, the problem includes two phases first is the learning process phase in which for analysis of training data, the rule and pattern are created. The second phase tests the data and archives the accuracy of classification patterns.

While doing classify analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main advantage of classification over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Benefits of classification:

Classification [3] is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called classifies. Help users understand the natural grouping or structure in a data set. Used

either as a stand-alone tool to get insight into data distribution or as a pre-processing step for other algorithms.

RELATED WORK

There are previous research is performed in this segment to find a better efficiency of the technique. The health care liver disease data mining classification is performed in previous solutions.

In a paper author Kush R. Varshney [4] proposed Machine learning algorithms increasingly influence our decisions and interact with us in all parts of our daily lives. Therefore, just as we consider the safety of power plants, highways, and a variety of other engineered socio-technical systems, we must also take into account the safety of systems involving machine learning. Heretofore, the definition of safety has not been formalized in a machine learning context. In this paper, we do so by defining machine learning safety in terms of risk, epistemic uncertainty, and the harm incurred by unwanted outcomes.

Mariam Adedoyin-Olowe [5] proposed more people are becoming interested in and relying on the SM for information, breaking news and other diverse subject matters. Users find out what other people's views are about certain product/service, film, school, or even more major issues like seeking other people's opinion on political candidates in national election poll. Millions of people access SM sites such as Twitter, Facebook, LinkedIn, YouTube and MySpace to search out for information, breaking news and news updates.
UrszulaStanczyk [6] proposed Weighting and Pruning of Decision Rules by Attributes and Attribute Rankings Rule classifiers express patterns discovered in data in learning processes through conditions on attributes included in the premises and pointing to specific classes. A variety of available approaches to induction enable construction of classifiers with minimal numbers of constituent rules, with all rules that can be inferred from the training samples, or with subsets of interesting elements. To limit the number of considered rules either pre-processing can be employed, with reducing rather data than rules, by selection of features or instances, or in-processing relaying on induction of only those rules that satisfy given requirements, or post-processing, which implements pruning mechanisms and rejection of some unsatisfactory rules. The paper focuses on this latter approach.

Thus the given solution executes the approach using the given solution of data classification approach find its usage and limitation in the terms of processing and accuracy.

PROPOSED METHODOLOGY

The approach is a tool used in exploratory phase of data mining. It projects input space on prototypes of a low-dimensional regular grid that can be effectively utilized to visualize and explore properties of the data. When the number of SOM units is large, to facilitate quantitative analysis of the map and the data, similar units need to be grouped, i.e., classified. In this research, different approaches to classification of the SOM are considered. In particular, the use of hierarchical agglomerative classification and portative classification using -means are investigated. The two-stage procedure—first using SOM to produce the prototypes that are then classified in the second stage—is found to perform well when compared with direct classification of the data and to reduce the computation time.

Self-Organizing Maps (SOM's)
• Categorization method
• A neural network technique
• Unsupervised

The self-organizing map (SOM) is especially suitable for data survey because it has prominent visualization properties. It creates a set of prototype vectors representing the data set and carries out a topology preserving projection of the prototypes from the -dimensional input space onto a low-dimensional grid. This ordered grid can be used as a convenient visualization surface for showing different features of the SOM (and thus of the data), for example, the classify structure.
The classification is carried out using a two-level approach, where the data set is first classified using the SOM, and then, the SOM is classified.

- SOM is a classification technique, which can be used to provide insight into the nature of data. We can transform this unsupervised neural network into a supervised LVQ neural network.
- The network architecture is just like a SOM, but without a topological structure.
- Each output neuron represents a known category (e.g. apple, pear, orange).
- Input vector $\underline{x} = (x_1, x_2, x_2, \cdots, x_n)$
- Weight vector for the jth output neuron $\underline{w}_j = (w_{1j}, w_{2j}, w_{3j}, \cdots, w_{nj})$
- C_j = Category represented by the j^{th} neuron. This is pre-assigned.
- T = Correct category for input $\underline{x}$
- Define Euclidean distance between the input vector and the weight vector of the j^{th} neuron.

The most important benefit of this procedure is that computational load decreases considerably, making it possible to classify large data sets and to consider several different pre-processing strategies in a limited time. Naturally, the approach is valid only if the classifies found using the SOM are similar to those of the original data. In the experiments, a comparison between the results of direct classification of data and classification of the prototype vectors of the SOM is performed, and the correspondence is found to be acceptable.

EXPERIMENTAL SETUP

An Experiment using the JDK framework and Weka tool implementation of the approach is used for analysis. The approach is used for analysis is compared with existing tree based approach available in tool configuration.

Dataset: UCI repository available dataset [7] is taken for the processing of execution.

Dataset for performing the comparison analysis '**ILPD (Indian Liver Patient Dataset)**' dataset has been used. The datasets has been collected from different resources. In the work weka tool is used for the analysis and evaluation[8].

The used dataset is extracted from the UCI machine learning repository and further processed in simulation tool which is Weka tool for data mining. The ILPD dataset contains a liver patient dataset which is total 583 patient records including 416 are the liver patient and other 167 are non-liver

patients. A north east Andhra Pradesh hospital data is collected and grouped in male & female patient. This dataset is containing 11 attributes i.e. different 11 properties are included in the dataset including patient age. A below table is the sample dataset shows the LDPI dataset which is taken for process.

The proposed technique and some random tree based approach is executed [9][10].

RESULT ANALYSIS

As per the algorithm computation using the weka tool, an experiment result analysis is presented and it shows the computation. The weka tool [11] is used for the experiment purpose and result examines are shown below.

Statistical Analysis:

In the given below section shows the experiment result observed from the execution.

Algorithm/Parameter	Bayes Algorithm	Genetic Algorithm	Heuristic SOM Algorithm
Precision	78.36	82.27	89.65
Recall	71.78	83.6	91.90
Accuracy	77.10	87.34	93.80

Table 1: Result Comparison analysis between the algorithm executions

As per given table 1 above, the discussion comparison analysis is done and presented in statically format. The presented table shows the efficiency of proposed technique while comparing with existing data classification technique over liver disease dataset which is ILPD Indian liver patient dataset.

Graphical Analysis:

A graphical analysis of the observed result shown the efficiency of proposed algorithm and its outcome.

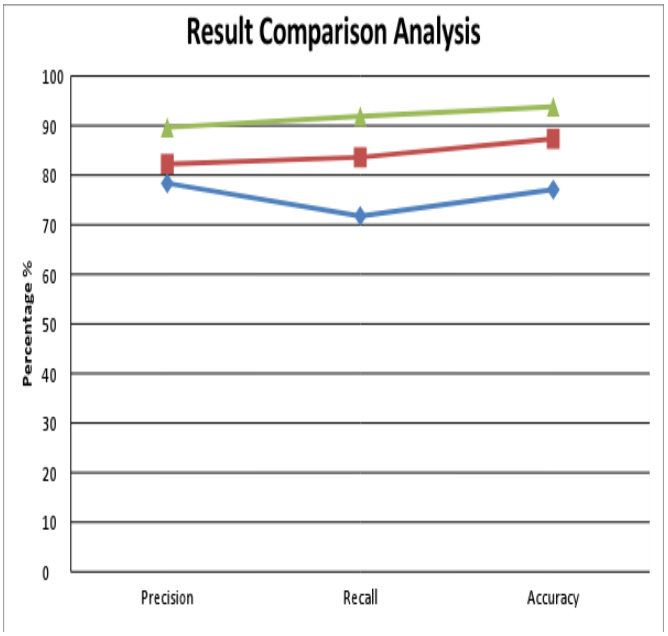

Figure 1: Comparison analysis line graph

In this figure 1 above, the comparison between the execution algorithm such as genetic algorithm, naïve Bayes approach and the proposed heuristic based SOM ANN Algorithm is given. The proposed graphical and statically presented result shows the

efficiency of given proposed approach over existing solution.

In the above section, discussed algorithm and proposed solution obtained in the given parameter execution is shown. The result execution shows the efficiency of proposed technique.

CONCLUSION & FUTURE WORK

Classification is a field in data mining, where a classification of large amount of data can be performed using the data mining classification approach. In this paper a liver healthcare detection and mining the useful information from the dataset is performed. Liver healthcare dataset is taken for experiment purpose and a computation using the existing solution Bayes and Tree based classification is performed. The proposed solution which is Heuristic based SOM ANN approach is given and compared with the given solution. The experiment setup on Weka tool and comparison analysis study shows the efficiency of proposed algorithm over the given simple classification solutions. A performance parameter such as accuracy and mean error shows the efficiency of proposed Technique. A further study direction to implement the with other real time dataset and providing the solution as a open source tool which can help in analysis of liver disease by users.

ACKNOWLEDGMENT

Here by study of this research, I am thankful to the guide cooperation and helping me to understand the data analysis and finding an efficient outcome from the research.

REFERENCES

[1]. Tapas RanjanBaitharua ,Subhendu Kumar Pani," Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset", International Conference on Computational Modeling and Security (CMS 2016), Elsevier.

[2]. A.A. Balamurugan, R. Rajaram, S. Pramala, et al., NB+: An improved Naïve Bayesian algorithm, Knowledge-Based Systems 24 (5) (2011) 563±569.

[3].. Sta´nczyk, U.. Rough set and artificial neural network approach to computational stylistics. In: Ramanna, S., Jain, L.C., Howlett, R.J.,editors. Emerging Paradigms in Machine Learning; vol. 13 of Smart Innovation, Systems and Technologies. Springer Berlin Heidelberg;2013, p. 441± 470.

[4]. Kush R. Varshney Data Science Theory and Algorithms proposed On the Safety of Machine Learning:Cyber-Physical Systems, Decision Sciences, andData Products, researchgate,2017.

[5]. Mariam Adedoyin-Olowe1, Mohamed Medhat Gaber1 and Frederic Stahl2 proposed A Survey of Data Mining Techniques for Social Media Analysis, 2014.

[6]. UrszulaSta´nczyk,"Weighting and Pruning of Decision Rules by Attributes and Attribute Rankings" ,2016, Springer.

[7]. https://archive.ics.uci.edu/ml/datasets/liver+disorders

[8]. S. Karthik, A. Priyadarishini and J. Anuradha and B. K. Tripathy," Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease

and its Types", Pelagia Research Library ,Advances in Applied Science Research, 2011.

[9]. BendiVenkataRamana and Prof. M.Surendra Prasad Babu," Liver Classification Using Modified Rotation Forest", International Journal of Engineering Research and Development, ISSN: 2278- 067X, Volume 1, Issue 6 (June 2012), PP.17-24.

[10]. HuanLiu,HiroshiMotoda and Rudy Setiono,"Feature Selection: An Ever Evolving Frontier in Data Mining",JMLR: Workshop and Conference Proceedings 10: 4-13 The Fourth Workshop on Feature Selection in Data Mining.

[11]. http://www.ijarcsse.com/docs/papers/Volume_ 3/9_Septe mber2013/V3I9-0189.pdf