ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

Analysis and Prediction of Dravet Syndrome Dataset based on Linear and Logistic Regression

B.Dwarakanath[1], Dr.K.Ramesh Kumar[2]

Department of Information Technology, Hindustan University, Chennai, Tamil Nadu, India

ABSTRACT: Health care industry has very large amount of data and is extracted to find out unknown information in data to generate knowledge. Data mining techniques are used to discover hidden patterns that can be effectively applied to disease diagnosis which will help the general practitioner to take effective decision. Due to the significance of data mining techniques, association rule mining which is one of the data mining techniques to find the hidden rules from the data is applied to disease diagnosis by various researchers. In this paper, we propose association rules based linear and logistic regression for the diagnosis of dravet syndrome. The proposed technique for the diagnosis of Dravet syndrome contain three essential phases, namely selection of significant attributes, generation of association rules and, linear and logistic regression with the performance comparison based on time between them.

KEYWORDS: Association Rules, Association Rule Mining (ARM), Linear Regression, Logistic Regression, Dravet Syndrome.

I. INTRODUCTION

Information retrieval from huge amount of data available at different health centers is quiet challenging. Therefore it is quite impossible to pull out expert knowledge from the world of medical dataset; also there was lot of problems because of imperfection of knowledge. Recently it brings attention for computer scientists, especially in the area of knowledge mining and artificial intelligence [1]. It is because of the computers and informatics, there is improvement in overall health delivery system [2]. Data gathering as well as consequent storage, retrieval and handling of the data are enhanced by efficient computerization through database in static fashion in medical practice. Such decision-making through computers is fruitful and improves accuracy [3]. Using knowledge gained from previous known data, data classification process is mostly applied in problems of medicine, social science management and engineering. Different types of problems such as disease diagnosis, medical image recognition, and credit evaluation using classification techniques [4]. One of the disease diagnoses achievable in the medical field is Epilepsy which is clinically and genetically heterogeneous group of disorders in human. 'Severe myoclonic epilepsy of infancy' (SMEI) or 'Dravet syndrome' is a fatal infantile-onset epilepsy affecting about 1 in 20 000 to 40 000 children with a two-fold predominance in males [5].

Mostly, mutations in genes encoding ion channels in brain neurons have been analyzed in various epilepsy syndromes [6]. It delivers extraordinary advancement in terms of diagnostic molecular pathology and reduces over-reliance on traditional clinical classification [5]. An initial diagnosis of Dravet syndrome is significant for treatment and genetic counseling [7]. But the detection of Dravet syndrome among children was a challenging task. However, the Information on the pervasiveness and attributes of early, vaccination-related seizures in Dravet syndrome make better recognition of Dravet syndrome among children presenting with seizures following vaccinations. The knowledge that vaccination, although possibly the trigger for the first seizure, is not the reason for genetic epilepsy syndrome of their child, could support immunization of other children in the family [8]. Statistical programming styles are efficient and effective approach, in medical data classification. In this paper, we propose a class association rules based feature selection and regression model for the diagnosis of dravet syndrome. Here, for the given dataset the class association

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

rules are generated using the constraints given. The significant attributes are taken to generate the association rules [9]. The insignificant attributes are removed from the dataset before generation of association rules. Then the linear and logistic regression techniques are applied on the dravet syndrome data set for the diagnosis of the disease.

II. REVIEW OF LITERATURE

In this section, some classification method related with disease diagnosis based on association rules and regression analysis is presented.

R. Chaves *et al.,* [10] have designed AR-based Fs technique to overcome the negativity of size problem of a functional database. Previously, it was designed for the early detection of Alzheimer's disease (AD) in order to design a CAD system. The ARs were acquired from activated blocks of controls at different minsuff and mincounf values. Also, the maxima values are selected from the most important voxels, rejecting the rest which in turn cent 100%. Again, the rules, from a set of control subjects were got hold of, by the mean of the same set rules. In terms of activation of antecedents, the resultant AR was joined without repetition and the voxels were selected and prepared. In order to verify the dependability of the designed AR FS-based method in terms of Acc, Spe, Sen converging to ideal operating point of 91.75%, 95.12%, 89.29% respectively for 12 PLS features (SPECT) and 90%, 90.67%, 89.33%(PET), which performed better than recently reported techniques by many try out conducted.

Onur Inan *et al.,* [11] have designed a feature selection called AP, to detect breast cancer, which combines the method of Apriori and PCA. At the beginning, all the inputs for feature selection was analysed by Apriori algorithm and according to one of its inputs termed as uniformity of cell size, the mitoses was destroyed. Therefore, the entered numbers was decreased to eight. After that, the most significant six data column were selected after the obtained reduced number of entered data set converted to PCA data space. The designed hybrid algorithms output was applied to multi-layered feed-forward back-propagation neutral network, which was an orthodox classier. For training & testing, a 10-fold cross-validation method has been used. The designed algorithm's functionality was calculated by Winconsin's breast cancer database, which illustrates that it offer better outcome as compared with other systems, which used to perform on cancer data using cross-validation.

Logistic regression and decision tree algorithms are been examined by Biswanath Samanta *et al.,* [12], for the forecasting of Periventricular Leukomalacia(PVL). It was examined based on the post-operative hemodynamic and artificial blood gas data, which has statistical feature extraction, feature selection using logistic regression and generation of classification rules using decision tree. By using LR & DT based selection process the rules were made quite traceable by the use of a reduced numbers of feature. If the dataset were less limited, then the test success of the decision tree algorithm the hard boundaries used to check the success of the algorithm.

Normally, classification trees were frequently used to arrange patients according to the presence or absence of a disease. To overcome the negative impact of classification trees, Peter C. Austin *et al.,* [15], has compared the functionality of classification methods with the orthodox classification trees. The functionality of the classification tree was assessed to organize patients with heart failure (HF). It was based on the heart failure subtype like HF with preserved ejection fraction (HFPEF) and HF with decreased ejection fraction. Comparing the ability of these techniques to forecast the possibility of the presence of HFPEF, with help of convectional logistic regression method is better for the prediction of the possibility for the presence of HFPEF.

III. METHODOLOGY

The data mining techniques applied for the diagnosis of the disease are the association rule mining and regression from the dravet syndrome data set.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

3.1 Association Rule Mining and Regression

Association rules are generated from huge amount of data in the databases. The frequent patterns are identified and based on which the rules are generated to extract the relationship between various attributes based on various interestingness measures. The optimum constraints used are minimum thresholds on support and confidence. In this paper, we propose linear regression and logistic regression techniques for the prediction of diagnosis of Dravet Syndrome [2]. The proposed technique for the diagnosis has three essential phases, selection of significant attributes, association rule generation and regression [14]. The figure 1 shows the proposed architectural diagram of ARM based Linear and Logistic regression. The association rules are generated by association rule mining (ARM). The linear regression is applied [13]. The logistic regression is applied and the performance based on execution time is compared between the two regression techniques.

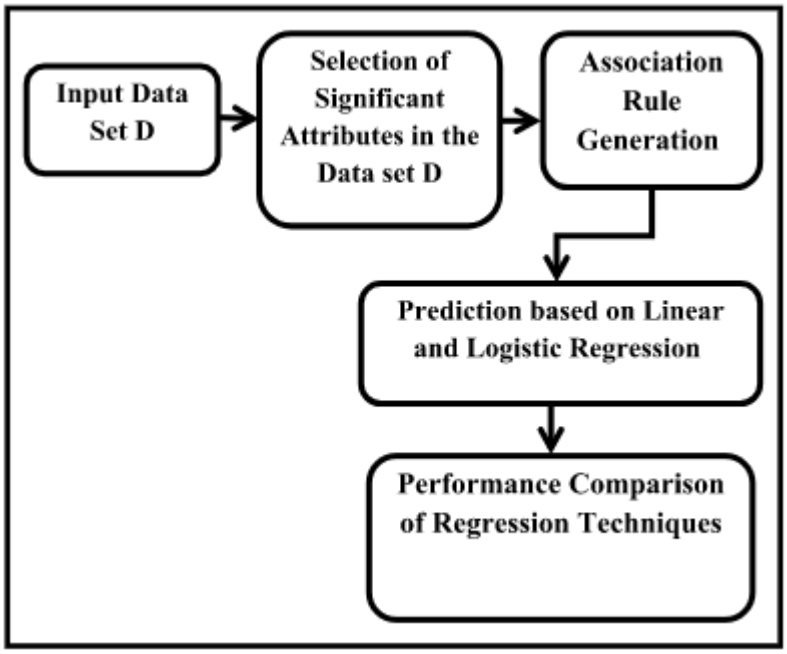

Figure 1: Proposed ARM-LILOR system

3.2 Apriori Algorithm

Aim: To find hidden rules from data.
Input: Dravet Syndrome Data set D.
Method: Apriori algorithm is used to generate association rules. It uses two parameters

- Support
- Confidence

Step 1: Find all item sets that have minimum support
Step 2: A frequent itemset is an item set whose support is ≥ minsup.
Step 3: Find all 1-item frequent item sets; then all 2-item frequent item sets, and so on.
Step 4: In each iteration *k*, only consider itemsets that contain some *k*-1 frequent itemset.
Step 5: Use frequent item sets to generate rules
Step 6: Frequent itemsets ≠ association rulesOne more step is needed to generate association rules. Step 7: For each proper nonempty subset *A* of *X*,

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

Let B = X - A
A $\rightarrow$ B is associations rule if
Confidence (A $\rightarrow$ B) $\geq$ minconf,
support (A $\rightarrow$ B) = support(A$\cup$B) = support(X)
confidence (A $\rightarrow$ B) = support(A $\cup$ B) /
support(A)
Output: Association Rules, attributes.

3.3 Linear and Logistic Regression

In this section the significant attributes in the dataset D is given to the regression based classifier which would predict the output. The dataset after selecting the significant attributes, it is given to the regression based classification prediction. The regression based classifier would predict based on the significant attributes and obtains the correct class. The two regression models which are used are linear regression method and logistic regression method for classification. The formula to predict the class based on the linear regression technique is

$$LR = \left(\sum_{i=1}^{I} att_i^d * wt_i \right) + \varepsilon$$

The formula to predict the class based on logistic regression technique is

$$LOR = \varepsilon * \log\left(\frac{1}{1+e^{\left(\sum_{i=1}^{I} att_i^d * wt_i \right)}} \right) + (1-\varepsilon) * \log\left(1 - \frac{1}{1+e^{\left(\sum_{i=1}^{I} att_i^d * wt_i \right)}} \right)$$

In the above equation LOR denotes the predicted class; and I denotes the total number of data; and att_i^d denotes i^{th} attribute of d^{th} data; and wt_i denotes the weight value of i^{th} attribute; and ε denotes the error term that is calculated at each iteration.

IV. RESULTS AND DISCUSSION

The dataset used to experiment the proposed technique is 'Dravet Syndrome GT' which is taken in real time. This dataset contains the details about ten patients who were affected by dravet syndrome. The attributes in the dataset are as follows: patient id, patient name, gender, age in years, age in months, number of months the patient fed mother's milk, first episode seizure occurrence month, number of visits to hospitals with respect to age in months, seizure occurrence, general reasons for seizure occurrence, general symptoms, tests, seizure type, vitamin drugs, seizure drugs, seizure status, activities and therapy. The figure 2 shows thee sample association rules generated for the data set as given below.

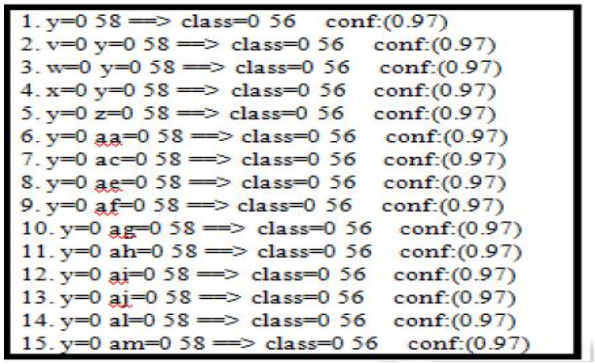

Figure 2 : Association rules generated

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

The Figure 3. shows the performance comparison based on execetion time for linear and logistic regression as below.

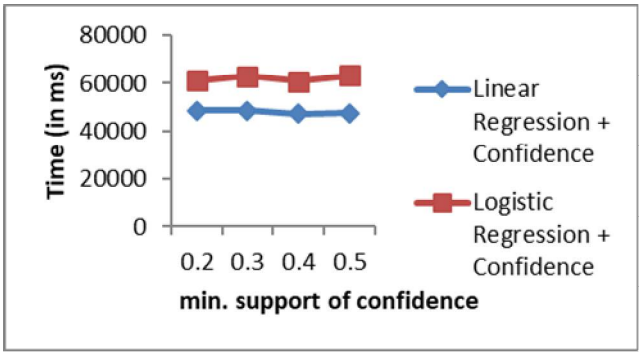

Figure 3. Execution time for Linear and Logistic regression

V. CONCLUSION

The association rules were generated and linear and logistic regression were applied with interestingness measure for the medical diagnose of dravet syndrome. The frequent patterns and association rules were generated by applying apriori algorithm and association rule mining techniques, rules that can be effectively applied to disease diagnosis which will help the physicians to take effective decision. In this, we studied and applied the linear and logistic regression for the prediction of diagnosis of Dravet Syndrome. In the proposed regression technique, linear regression has minimum execution time than logistic regression for prediction.

VI. FUTURE WORK

The association rule can be generated with missing values with representativity measure and the regression techniques can be used for the dravet data set.. The performance of the algorithm can be analyzed with various measures such as , lift and leverage.

REFERENCES

[1] B. K. Tripathy, D. P. Acharjya and V. Cynthya, "A Framework For Intelligent Medical Diagnosis Using Rough Set With Formal Concept Analysis", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.2, No.2, 2011.
[2] R.D. Lele, Computer in Medicine. Tata McGraw Hill Publishing Company Ltd., New Delhi, 1988.
[3] D.M. Shah & S.M. Patel, "IT in Healthcare", Presented at the Proceeding of the National Seminar on Computer Applications, 2001.
[4] Murat Karabatak, M. Cevdet Ince, "An expert system for detection of breast cancer based on association rules and neural network", Expert Systems with Applications, vol. 36, pp. 3465–3469, 2009.
[5] Chloe M Mak, KY Chan, Eric KC Yau, Sammy PL Chen, WK Siu, CY Law, CW Lam and Albert YW Chan, "Genetic diagnosis of severe myoclonic epilepsy of infancy (Dravet syndrome) with SCN1A mutations in the Hong Kong Chinese patients, Hong Kong Medical Journal, Vol. 17, No. 6, 2011.
[6] Ingrid E. Scheffer, "Diagnosis and long-term course of Dravet syndrome", European journal of pediatric neurology, pp.1-4, 2012.
[7] Ceulemans B, Boel M, Claes L, Dom L, Willekens H, et al. "Severe myoclonic epilepsy in infancy", Toward an optimal treatment. J Child Neurol, Vol. 19, pp.516–521, 2004.
[8] Nienke E. Verbeek, Nicoline A. T. van der Maas, Floor E. Jansen, Marjan J. A. van Kempen, Dick Lindhout and Eva H. Brilstra," Prevalence ofSCN1A-Related Dravet Syndrome among Children Reported with Seizures following Vaccination: A Population-Based Ten-Year Cohort Study", PLOS ONE Journal, Vol. 8, No.6, 2013.
[9] Agarwal, R., Imielinski, T., and Swami, A. "Mining association rules between sets of items in large databases." In: SIGMOD, pp. 207-216, 1993.
[10] R. Chaves, J. Ramírez, J.M. Górriz, C.G. Puntonet,
"Association rule-based feature selection method for
Alzheimer's disease diagnosis", Expert Systems with Applications, vol. 39, pp. 11766–11774, 2012.
[11] Onur Inan, Mustafa Serter Uzer and Nihat Yilmaz, "A
New Hybrid Feature Selection Method Based On

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

Association Rules and PCA for Detection of Breast Cancer", International Journal of Innovative Computing, Information and Control ICIC International, Vol. 9, No.2, pp. 727-739, 2013.

[12] Biswanath Samanta, Geoffrey L. Bird, Marijn Kuijpers, Robert A. Zimmerman, Gail P. Jarvik, Gil Wernovsky, Robert R. Clancy, Daniel J. Licht, J. William Gaynor and Chandrasekhar Nataraj, "Prediction of per ventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms", Artificial Intelligence in Medicine, Vol.46, pp. 201—215, 2009.

[13]J. M. Hahne, F. Bießmann, N. Jiang, Member, IEEE "Linear and Nonlinear regression for simultaneous and proportional myoelectric control", IEEE Transactions on Neural and Rehabilitation Engineering,Vol. 22, No. 2, March 2014.

[14] Swati Gupta, "A Regression Modeling Technique on Data Mining." International Journal of Computer Applications, Volume 116 – No. 9, pp.27-29, 2015.

[15] Peter C. Austin, Jack V. Tu, Jennifer E. Ho, Daniel Levy and Douglas S. Lee, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes", Journal of Clinical Epidemiology, Vol. 66, pp.398-407, 2013.

BIOGRAPHY

B. Dwarakanath received the B.E. and M.Tech. degrees in Computer Science and Engineering from Bangalore University and Vellore Institute of Technology in 2000 and 2004, respectively. During 2001-2002, he stayed in Muthayammal Engineering College. His research interests are Data Mining, Image Processing and Network Programming. He is now associated with Hindustan University.

Dr. K. Ramesh Kumar received the Ph.D. degree in Computer Science from Alagappa University. During 2010-2011, he stayed in Karunya University. His areas of interest are Data Mining, Ubiquitous and Cloud Computing. Currently, He is now associated with Hindustan University.