International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

A SURVEY ON MACHINE LEARNING INTELLIGENCE TECHNIQUES FOR MEDICAL DATASET CLASSSIFICATION

P.M. Benson Mansingh[1], Dr. M. Yuvaraju[2]

[1]Assistant Professor, Dept of ECE, Sri Ramakrishna Institute of Technology, Coimbatore
[2]Assistant Professor, Dept of EEE, Anna University Regional Campus, Coimbatore

---

Abstract - *In this paper, a survey has been done on the application of Artificial intelligence computing techniques for diagnostic of disease by classifying the bio medical datasets. Many Artificial Intelligence techniques were reviewed for medical dataset classification. This Exploration assembles typical work that shows how the Artificial Neural Network is applied to the solution of different diagnostic disease with classification. It also detects the methods and the techniques of ANN that are used frequently to solve the special problem related to the medical dataset classification. Extreme Learning Machine (ELM) is used in almost for learning the medical datasets to the network. Similarly PSO is used to optimize the attributes or the parameter of the datasets for classification. Several diseases like Breast cancer, Heart disease, Diabetes….etc using ANN approach are result in use of SVM (Support vector Machine) and BP network.*

Key Words: Medical datasets, Machine learning, Artificial Intelligence (AI), Extreme learning Machine (ELM), Artificial Neural Network (ANN)

1. INTRODUCTION

Artificial Intelligence (AI) is the science and engineering making intelligent machines, especially intelligent programs. And diagnosis is followed with the development of algorithm and techniques that are able to determine whether the behaviour of a system is correct. The application of computational or machine intelligence in medical diagnosis is a new trend for medical dataset classification. Classification system can help to minimizing possible errors that can be done because of inexperienced experts. And also provide medical dataset to be examined in shorter time and in detail. One of the application areas of analyzing database and medical dataset classification is automated diagnostic systems.

The aims of these studies are assisting to doctors in making diagnostic decision with a subject to assure the diagnosis aid accurately. In medical field, many researchers applying different technique to improve the accuracy for the given data, by accurate values give after classification to know the affected patient and improve the diagnosis. That the classification efficiency is allowed through Sensitivity, Specificity and Accuracy for classification functions. A good classifier should give hundred percent results for all the three.

2. MEDICAL DATASET

Medical classification, or medical coding, is the process of transforming descriptions of medical diagnoses and procedures into universal medical code numbers. The diagnoses and procedures within the health care record, such as the transcription of the physician's notes, laboratory results, radiologic results, and other sources are usually taken from a variety of source.

Medical classification systems are used for a variety of applications in medicine, public health and medical informatics, including: statistical analysis of diseases and therapeutic actions reimbursement; e.g., based on diagnosis-related groups knowledge- based and decision support systems direct surveillance of epidemic or pandemic outbreaks. The medical datasets are taken from the UCI machine repository.

2.1 MEDICAL DATASET ANLAYSIS METHODS

Medical dataset analysis method is included with three important steps. First step includes dataset pre processing, second follows feature selection and final step include classifying the dataset. This all described in following sections.

2.2 DATASET PRE PROCESSING

Pre processing the thousands of medical datasets are combined into one relational table, by pre processing it delete the mismatched data and also it removes the multivalued attribute. And it replaces the missing value by its mean, median and its standard deviation (SD).

Table -1: Different medical datasets

LIST OF DISEASES	Number of Instances	Number of Attributes	Number of Classes
Breast Cancer	699	10	2
Diabetes	1151	20	3
Lung Cancer	32	56	4
Heart	303	75	2
Hepatitis	155	19	2
Thyroid	7200	21	4
Breast Cancer	699	10	2

2.3 FEATURE SELECTION

After normalizing the datasets, Feature selection method is used to get the most important features of the dataset. That the method mentions the significant feature before going to classification. Several methods are used for feature selection are F-Score, Threshold fuzzy entropy, PCA, GDA etc.

F-SCORE

Feature selection by F-Score is used to find the optimal subset of input variable with the best feature by removing no predictive information. That it gives the good accuracy value for classifying the medical datasets. It follows some steps for feature selection..

- All features are taken and calculated using the given formula.
- Mean is calculated using the formula.
- Compare the mean value of the feature with the original value.
- It measures the discrimination or relevant feature values related to 2 different features.

THRESHOLD FUZZY ENTROPY BASED FEATURE SELECTION

Feature is selected based on the Fuzzy C means Clustering with three framework are followed

- Mean selection
- Half selection
- Neural network

2.4 PCA

PCA used to convert the set of observation of possibly correlated variable into the set of linearly uncorrelated values. And it reduces the dimension of the datasets with minimal loss of information and selects the most relevant feature. The reduction of feature dimension by extract the sub set of feature that describe as the best feature and evaluated with high accuracy.

2.5 GDA

General Discriminant Analysis is used as a linear analysis model to the Discriminant analysis problem and also for classification problem. By using GDA we can set a complex model for the set of predictor variable.

FEATURE CLASSIFICATION

After the feature is extracted using the feature selection techniques. The best feature is selected using the classification algorithms like Neural Network, Support Vector Machine (SVM), K-NN (K-Nearest Neighborhood) , Decision Tree. By using the following techniques the best feature are selected and grouped into one and classifying the datasets with high accurate of disease rate. That classification having both the supervised and unsupervised learning algorithms. Classification techniques for supervised learning are Cluster, Fuzzy C Means etc... And unsupervised learning algorithms are ANN, K-NN, Decision tree etc…

2.6 NEURAL NETWORK

Artificial Neural Network is a powerful tool for solving the complex problem with linear input-output efficient relationship. That neural network is based on connections with human brain which having N number of neurons. Neural network are followed with three layers they are input layer, output layer and hidden layer. Hidden layer is used to map the input function to the output. Neural network having following features it support several network architecture for supervised and unsupervised learning, it uses parallel computing for feature training process, dynamic network to store the data. And it uses unsupervised learning to train the new input that adjusts itself continuously. Mathematical rules like learning and training functions are used to adjust the weight and bias automatically.

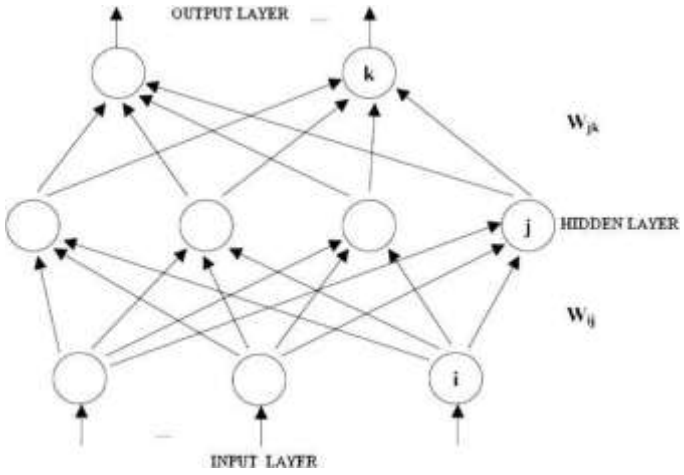

Fig -1: Neural Network

2.7 SVM

Classification with SVM by finding the hyper plane that increases the dimensions of two plane classes. The hyper plane vector are defines support vectors. Maximize the width of the margin to get the optimal hyper plane. SVM are allowed with linear hyper plane, and then it have a unique global value. It built a model of new sample into one class, making a non probabilistic linear classifier.

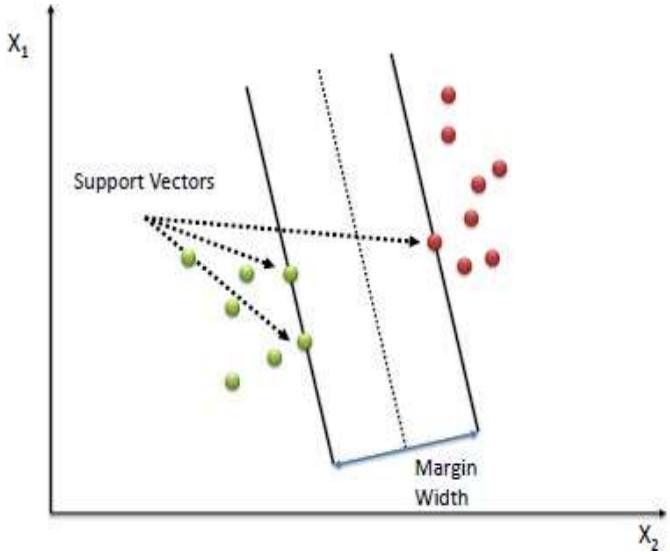

Fig -2: Support Vectors

2.8 K-NN

K-NN is the non linear parameter used for classification and regression. In both, for classification and regression the input are combined with K closest neighbor values of the feature space. Classification is followed according to the classification condition. The output is assigned by membership class. The value of K is either Positive or small. They classify the object or feature according to the value.

2.9 PERFORMANCE ACCURACY

The Diagnosis of the diseases is identified by medical datasets with several feature selection methods and classification techniques. The diagnosis accuracy may be varied from one exact technique to another technique. Different feature extraction techniques that combine with the classification method provide better results. The Table 3 shows that the various accuracy levels of feature extraction and classification methods during the emotion identification process.

Table -2: Accuracy of Methods

S. NO	SELECTION METHOD	CLASSIFICATION ALGORITHM	ACCURACY	CITIATION
1	Kernel F-Score	SVM	76.03%	[1]
2	Fuzzy entropy	K-NN	75.45%	[7]
3	F –Score	NN	85.90%	[3]
4	Weighted F-Score	RBF	79.12%	[8]
5	PCA	SVM	67.8%	[3,4]
6	GDA	NN	70.8%	[5]

3. CONCLUSION

Classification of medical dataset can be identified by extracting the different kind of feature from the datasets. For extracting the features from the datasets Fuzzy C Means Clustering will give the highest accuracy, after preprocessing the signal it has to be smoothed and optimized for the particular feature, using different optimization techniques like SVM, PSO, etc.. After getting the optimized result apply the PSO and Fuzzy Cognitive map it will provide the high accuracy of classification. These classification of medical dataset will used for diagnosis of disease in early stage.

ACKNOWLEDGEMENT

We would earnestly thank our Supervisor for all his valuable comments and active support.

REFERENCES

[1] Kemal Polat *, Salih Gunes ”A new feature selection method on classification of medical datasets: Kernel F-score feature selection” in 2009 Elsevier.

[2] P. Jaganathan and R. Kuppuchamy, “A threshold fuzzy entropy based feature selection for medical database classification,” Computers in Biology and Medicine, vol. 43, no. 12, pp. 2222– 2229, 2013.

[3] Peng Tao1, Huang Yi” A Method Based on Weighted F-score and SVM for Feature Selection” in 2015 CCDC.

[4] H. Temurtas, N. Yumusak, and F. Temurtas, “A comparative study on diabetes disease diagnosis using neural networks, Expert Systems with Applications, vol. 36, no. 4, pp. 8610–8615, 2009.

[5] K. Polat, S.G¨unes¸, andA.Arslan, “Acascade learning systemfor classification of diabetes disease: generalized discriminant analysis and least square support vector machine,” Expert Systems with Applications, vol. 34, no. 1, pp. 482–487, 2008.

[6] M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” Expert Systems with Applications, vol. 36, no. 2, pp. 3240– 3247, 2009.

[7] S. Salcedo-Sanz, A. Pastor-Sánchez, L. Prieto, A. Blanco- Aguilera, and RGarc´ıa-Herrera, “Feature selection in wind speed prediction systems based on a hybrid coral reefs optimization—extreme learning machine approach,” Energy Conversion and Management, vol. 87, pp. 10–18, 2014

[8] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in Proceedings of the IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948, IEEE, Perth, Australia, December 1995.

[9] Kayaer, K., & Yıldırım, T. (2003). Medical diagnosis on pima Indian diabetes using general regression neural networks, artificial neural networks and neural information processing (ICANN/ICONIP) (pp. 181–184), Istanbul, Turkey, June 26–29.

[10] Vapnik, V. (1995). The nature of statistical learning theory. New York: Springer.

[11] Ster, B., & Dobnikar, A. (1996). Neural networks in medical diagnosis: Comparison with other methods. In Proceedings of the international conference on engineering applications of neural networks (pp. 427–430).

[12] West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnosis decision support system: A breast cancer diagnosis application. European Journal of Operational Research(162), 532–551.

BIOGRAPHIES

P.M.Benson Mansingh completed B.E (Electronics and Communication Engineering) from St.Peter's engineering college in 2010. He received his Master's Degree in Network Engineering from Anna University Regional Campus, Coimbatore in 2014. Currently he is pursuing his Ph.D degree in ANNA UNIVERSITY, Chennai and working as Assistant Professor in Sri Ramakrishna Institute of Technology.

Dr.M.Yuvaraju, Assistant Professor, Department of Electrical & Electronics Engineering, Anna University Regional Centre Coimbatore. He has guided several UG and PG students and currently guiding 5 Ph.D Research Scholars in Anna university, Chennai.