A Novel Soft Computing Based Model For Symptom Analysis & Disease Classification

Smita Prava Mishra, Debahuti Mishra & Srikanta Patnaik

Institute of Technical Education and Research, Siksha O Anusandhan University, Bhubaneswar, Odisha, India
Email: smitaprava@yahoo.com, debahuti@iter.ac.in, prof.srikantapatnaik@gmail.com

Abstract **-** In countries like India, many mortality occurs every year because of improper pronouncement of disease on time. Many people remain deprived of medication as the people per doctor ratio are nearly 1:1700. Every human body and its physiological processes show some symptoms of a diseased condition. The proposed model in this paper would analyze those symptoms for identification of the disease and its type. In this proposed model, few selected attributes would be considered which are shown as symptoms by a person suspected with a particular disease. Those attributes can be taken as input for the proposed symptom analysis and classification model, which is a soft computing model for classifying a sample first to be diseased or disease free and then, if diseased, predicting its type (if any). Number of diseased and disease free samples are to be collected. Each of these samples is a collection of attributes shown / expressed by a human body. With respect to a specific disease, those collected samples form two primary clusters, one is diseased and the other one is disease free. The disease free cluster may be discarded for further analysis. Depending on the symptoms shown by the diseased samples, every disease has some types based on the symptoms it shows. The diseased cluster of samples can reform clusters among themselves depending on the types of the disease. Those clusters then become the classes of the multiclass classifier for analysis of a new incoming sample.

Keywords **-** *Symptom analysis; Clustering; Classification; Multiclass Classifier*.

I. INTRODUCTION

Every human body and its physiological processes show some symptoms of a diseased condition. The proposed model in this paper would analyze those symptoms for identification of the disease and its type. Symptoms may be clinical parameters like blood pressure, blood glucose, scanning reports etc or linguistic expressions like nauseas, weakness, repulsion towards social gatherings etc. In this proposed model, few selected attributes would be considered which are shown as symptoms by a person suspected with a particular disease. Those attributes can be taken as input for the proposed symptom analysis and classification model, which is a soft computing model for classifying a sample first to be diseased or disease free and then, if diseased, predicting its type (if any). For achieving this goal, number of diseased and disease free samples are to be collected. Each of these samples is a collection of attributes shown / expressed by a human body. With respect to a specific disease, those collected samples form two primary clusters, one is diseased and the other one is disease free. The disease free cluster may be discarded for further analysis. Depending on the symptoms shown by the diseased samples, every disease has some types based on the symptoms it shows. The diseased cluster of samples can reform clusters among themselves depending on the types of the disease. Those clusters then become the classes for analysis of a new incoming sample.

A new sample could be a set of symptoms shown by a human being, to be tested for a particular disease. If the classifier classifies it to the disease free class then the person would be announced to be safe from that disease. If the sample would be found diseased then it has to be further classified to a particular type of the disease. Basing on the predicted type, a physician may further confirm it with clinical tests and prescribe required medication.

This paper has been organized as follows: section II discusses some related works in this context. Section III introduces some preliminaries required for discussion of the model. Section IV proposes the model. In section V explanatory analysis of the model is given. Section VI concludes the paper and discusses future directions.

II. RELATED WORK

Many works has been done on disease predictions. Few recent works in this context are as follows: Maglogiannis I. et al. [1], Yu W. et al. [2], Son Y J.

et al. [3], Babaoglu I. et al. [4], [5], Candelieri A. et al. [10] and Xia J. et al. [12] have used Support Vector Machine(SVM) for various diagnosis purposes like heart valve diseases, classifying diabetes and pre-diabetes, medical adherence in heart failure patients, Coronary Artery Disease, predictions for chronic heart failure patients and diagnosis of erythemato-squamous diseases respectively. Also, performance of SVM has been compared with other techniques like, Back-propagation Neural Networks, *k*- nearest neighbors and Naïve Bayes classifiers [1] and found to give the best result for heart valve disease classification. SVM with other feature reduction techniques like Principal Component Analysis (PCA) can even enhance the performance of SVM [7]. Even many other techniques have been used for automatic disease prediction. Few of the recent works are like: Hu V W. et al. [2] have implemented multiple clustering algorithms on Autism Diagnostic Interview – Revised diagnostic instrument for identifying subgroups of autistic probands, to isolate more homogeneous groups of people suffering from autism for gene expression analysis. They were able to form four different clusters by applying Principal Component Analysis (PCA), followed by Hierarchical Clustering (HCL) and finally, *k* - means clustering. In the work of Srinivas K. et al. [3], a study has been done on use of classification based data mining techniques like Rule based, Decision tree, Naïve Bayes and Artificial Neural Network on massive health care data. Two extended versions of Naïve Bayes method, such as One Dependency Augumented Naïve Bayes (ODANB) classifier and Naïve Credal Classifier 2 (NCC2) have been used for data preprocessing and a heart attack prediction method have been suggested. Work of Thakkar B A. et al. [4] is a significant effort, for classification of Swine Flu patients based on Naïve Bayes classifier. They developed a prototype of Intelligent Swine Flu Prediction Software (ISWPS), which gave an accuracy of 63.33% with three categories of severity. Jiao Y. et al. [9] made a comparison between brain regional cortical thickness and volume based various models for predicting Autism Spectrum Disorder over its phenotypic range. This study concludes that out of SVM, Multi Layer Perceptrons (MLP), Functional Trees (FT) and Logistic Model Trees (LMT) machine learning techniques, thickness based diagnostic with LMT classifier gives best results with accuracy of 87%, area under ROC curve of .93, sensitivity of 95% and specificity of 75%. Vecrei A. et al. [11] proposed a wavelet based local binary patterns applied to classification with modified Marsh scheme having four classes. They automated classification of duodenal texture patches for pediatric celiac disease prediction with overall classification rates in the range of 60-65 percent. Chen H L. et al. [13] have introduced a hybrid method for diagnosing hepatitis disease by integrating Local Fisher Discriminant Analysis (LFDA) and SVM. The model is also compared with PCA and SVM hybrid model and Fisher Discriminant Analysis (FDA) and SVM hybrid model. The comparison concludes that LFDA with SVM is best classifier for hepatitis disease having accuracy of 96.77%. Mostafa F G. et al. [14] have designed a diabetes disease classifier using ant colony based classification using method by extracting a set of fuzzy rules. They named it FCS-ANTMINER and it gave an accuracy of 84.24% which outranked all previous classifiers used for diabetes disease diagnosis. The validation was done using 10 fold cross validation methods.

III. PRELIMINARIES

To understand the proposed model, there is a need of discussion on few preliminary data mining concepts. They would enable a deeper insight and clear visualization of the model. Follows are the sequentially followed data mining sub steps as applicable to the model.

A. Data Preprocessing: Real world data are generally, incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data, noisy: containing errors or outliers, inconsistent: containing discrepancies in codes or names. In data preprocessing, we do data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies, data integration: using multiple databases, data cubes, or files, data transformation: normalization and aggregation, data reduction: reducing the volume but producing the same or similar analytical results, data discretization: part of data reduction, replacing numerical attributes with nominal ones. Out of all these techniques, most important in the current scenario is data normalization & standardization.

B. Data Normalization & Standardization: Data transformation such as normalization may improve the accuracy and efficiency of mining algorithms. Such methods provide better results if the data to be analyzed have been normalized, that is, scaled to specific ranges such as [0.0, 1.0]. An attribute is normalized by scaling its values so that they fall within a small-specified range, such as 0.0 to 1.0. There are many methods for data normalization as:

- Min-max normalization performs a linear transformation on the original data. Suppose that min_a and max_a are the minimum and the maximum values for attribute *A*. Min-max normalization maps a value *v* of *A* to *v'* in the range [$new\text{-}min_a$, $new\text{-}max_a$] by computing:

v'= ((v-min_a) / (max_a – min_a)) * ($new\text{-}max_a$ – $new\text{-}min_a$) + $new\text{-}min_a$

- Z-score normalization, the values for and attribute A are normalized based on the mean and standard deviation of *A*. A value *v* of *A* is normalized to *v'* by computing:

$$v' = \left((v - \bar{A})\right)/\sigma_{\bar{A}}$$

Where, $\bar{A}$ and $\sigma_{\bar{A}}$ are the mean and the standard deviation respectively of attribute *A*. This method of normalization is useful when the actual minimum and maximum of attribute *A* are unknown.

- Decimal scaling normalizes by moving the decimal point of values of attribute *A*. The number of decimal points moved depends on the maximum absolute value of *A*. A value *v* of *A* is normalized to *v'* by computing:

v' = (v / 10^j)

Where j is the smallest integer such that Max(|v'|)<1.

C. Feature Reduction: It means reducing the number of attributes of the data. It can be done as:

- Data cube aggregation: applying roll-up, slice or dice operations.
- Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space.
- Principle component analysis (numeric attributes only): searching for a lower dimensional space that can best represent the data.

Further reduction of the number of attribute values can be done as:

- Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).
- Clustering: grouping values in clusters.
- Aggregation or generalization

D. Cluster Analysis: Cluster analysis organizes data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups. The representation can then be investigated to see if the data group according to preconceived ideas or to suggest new experiments. Basic Clustering techniques are as:

- **Partitional**: Given a database of objects, a partitional clustering algorithm constructs partitions of the data, where each cluster optimizes a clustering criterion, such as the minimization of the sum of squared distance from the mean within each cluster. One of the issues with such algorithms is their high complexity, as some of them exhaustively enumerate all possible groupings and try to find the global optimum. Even for a small number of objects, the number of partitions is huge. That's why, common solutions start with an initial, usually random, partition and proceed with its refinement. A better practice would be to run the partitional algorithm for different sets of initial _ points (considered as representatives) and investigate whether all solutions lead to the same final partition. Partitional Clustering algorithms to locally improve a certain criterion. First, they compute the values of the similarity or distance, they order the results, and pick the one that optimizes the criterion.
- **Hierarchical**: Hierarchical algorithms create a hierarchical decomposition of the objects. They are either *agglomerative* (*bottom-up*) or *divisive* (*top-down*):
  - *Agglomerative* algorithms start with each object being a separate cluster itself, and successively merge groups according to a distance measure. The clustering may stop when all objects are in a single group or at any other point the user wants. These methods generally follow a greedy-like bottom-up merging.
  - *Divisive* algorithms follow the opposite strategy. They start with one group of all objects and successively split groups into smaller ones, until each object falls in one cluster, or as desired. Divisive approaches divide the data objects in disjoint groups at every step, and follow the same pattern until all objects fall into a separate cluster.

E. Classification: Classification is a data mining function that derives a model which assigns items in a collection to target categories or classes. The derived model is based on the analysis of a set of data objects whose class labels are known. Those are called the training data.

- **Decision trees:** Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values.
- **Bayesian Networks:** A Bayesian Network (BN) is a graphical model for probability relationships among a set of variables features. The Bayesian

network structure S is a directed acyclic graph (DAG) and the nodes in S are in one-to-one correspondence with the features X. The arcs represent casual influences among the features while the *lack* of possible arcs in S encodes conditional independencies. Moreover, a feature (node) is conditionally independent from its non-descendants given its parents ($X1$ is conditionally independent from $X2$ given $X3$ if $P(X1|X2,X3)=P(X1|X3)$ for all possible values of $X1$, $X2$, $X3$).

- **Nearest Neighbour Classifier:** Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an *n*-dimensional pattern space. When given an unknown sample, a *k*-nearest neighbour classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of euclidean distance.

- **Support Vector Machine (SVM) classifier:** SVM classification is based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. SVM finds the vectors (support vectors) that define the separators giving the widest separation of classes.

- **Rule- based Classifier:** A rule-based classifier uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form:

  IF *condition* THEN *conclusion.*

The "IF"-part (or left-hand side) of a rule is known as the rule antecedent or precondition. The "THEN"-part (or right-hand side) is the rule consequent. In the rule antecedent, the condition consists of one or more *attribute tests* that are logically ANDed. The rule's consequent contains a class prediction.

- **Classification by Backpropagation :**

  Backpropagation is a neural network learning algorithm. A neural network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. Neural network learning is also referred to as connectionist learning due to the connections between units.

- **Classiication by Association Rule Analysis:**

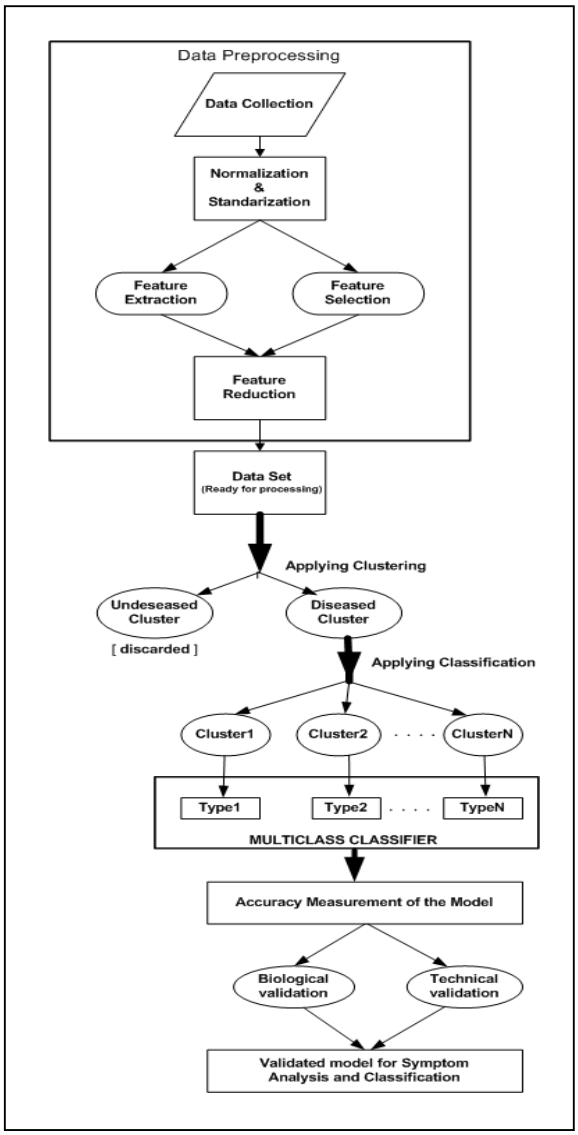
Frequent patterns and their corresponding association or correlation rules characterize interesting relationships between attribute conditions and class labels, and thus have been recently used for effective classification. Association rules show strong associations between attribute-value pairs (or *items*) that occur frequently in a given data set. Association rules can also be generated and analyzed for use in classification. The general idea is that strong associations can be searched between frequent patterns (conjunctions of attribute-value pairs) and class labels. Because association rules explore highly confident associations among multiple attributes, this approach may overcome some constraints introduced by decision-tree induction, which considers only one attribute at a time. In many studies, associative classification has been found to be more accurate than some traditional classification methods.

IV. PROPOSED MODEL

The proposed model to be designed needs a collection of data set where each sample is a set of attributes collected from people suspected with a particular disease. We have to collect the samples by interviewing the suspected patients and their relatives or by questionnaires. This collected data set is not suitable for processing. It needs to be normalized and standardized so that the attributes are evenly spread over a range of processable values. At this stage the data set still contains many redundant and irrelevant attributes which needs reduction before actual processing. After reduction the data set is ready for further processing.

A suitable clustering technique will be applied on the preprocessed data to distinguish the data set to two primary clusters i.e. diseased and disease free. The disease free dataset will be discarded for further analysis. The data set clustered as diseased will undergo a further clustering technique. This clustering will be applied to segregate the dataset into further clusters depending on the categorization of the pronounced disease. Those clusters would be labeled as per the known types of the disease. After labeling a class attribute is added to every sample of the dataset. Those classes will become the basis of further classification.

The next task would be to design an optimized multiclass classifier for a, to be tested sample of symptoms for a particular disease. Once the classifier is built it would be able to classify a sample at first as diseased or free from disease and subsequently if diseased then particular type of the disease. Once the classification is done accuracy of the classifier is to be measured and the result is to be validated both technically and biologically.



V. EXPLANATORY ANALYSIS

We have to collect the samples by interviewing the suspected patients and their relatives or by questionnaires. Many diseases vary in their symptoms greatly. Even many diseases show similar symptoms. So, identification of a disease and its symptoms are important. Symptoms may be clinical or expressional. Hence, proper selection of attributes capable of diagnosing the disease is difficult. This collected data set is not suitable for processing. The unprocessed dataset spreads over a diverged range. It needs proper normalization and standardization, followed by reduction as it contains redundant and irrelevant attributes. With a close observation and thorough analysis of the collected data, it can be concluded that *z*-score normalization is most convenient for the data set because as it considers mean and standard deviation, it evenly spreads the data through the available range. For feature reduction, removal of irrelevant attributes can be done and for reducing the attributes, clustering may be applied as it can group the attributes in to relevant and irrelevant clusters. In our current model, two levels of clustering is proposed. Initially, a partitional clustering technique can be applied which will broadly partition the diseased and disease free samples. Followed could be a hierarchical divisive clustering technique which will analyze the symptoms from more general to specific orders, depending on categorization of the diseases. After evaluation of all the classification techniques, we suggest SVM technique for classification as it gives most accurate classification result with smaller pre classified data sets. Also, execution time cannot be a constraint for symptom analysis and classification.

Different approaches may be employed to solve the problem of multiclass classification. At first binary classification problems may be extended to handle the multiclass case directly. This can include neural networks, decision trees, support vector machines, naive bayes, and *k*-nearest neighbours. The second approach can be to decompose the problem into several binary classification tasks. Several methods are used for this decomposition: one versus- all, all-versus-all, error-correcting output coding, and generalized coding. The third approach can be to arranging the classes in a tree, usually a binary tree and utilizing a number of binary classifiers at the nodes of the tree till a leaf node is reached.

VI. CONCLUSION & FUTURE DIRECTIONS

The outcome of the proposed work will be a classifier which takes in a set of symptoms to be analyzed for a particular disease and outcome would be to successfully conclude whether the sample is diseased or free from disease, if diseased then type of the disease(if any). Purpose of this work is pre analysis of a patient's symptoms for pronouncement of a disease and its type, prior to adoption of medication under a physician's supervision. This work will certainly save price less time and effort of physicians and will bring to an end sufferings of diseased people waiting for diagnosis.

In future we can test the model for various types of diseases and substitute the theoretical model with an implementation model. As and when required, in every sub-step of the model, optimization techniques can be embedded to improve the overall performance of the multi-class classifier. Also, comparisons can be done among alternative data mining techniques adoptable in

every stage of the classifier. Further, once the classifier is built and validated, the next challenge is to find the biological significance of the classified symptoms for further analysis.

REFERENCES

[1] Maglogiannis I., Loukis E., Zafiropoulos E., Stasis A., "Support Vectors Machine-based identification of heart valve diseases using heart sounds", Computer Methods and Programs in Biomedicine, Vol. 95, pp. 47–61 (2009)

[2] Yu W., Liu T., Valdez R., Gwinn M., Khoury M J., "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes", BMC Medical Informatics and Decision Making, Vol. 10(16) (2010)

[3] Son Y J.,Kim H G., Eung-Hee Kim E H., Choi S, Lee S K., "Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients", Healthcare Informatics Research (HIR), Vol.16(4), pp. 253-259 (2010)

[4] Babaoglu I., Findik O., Bayrak M., "Effects of principle component analysis on assessment of coronary artery diseases using support vector machine", Expert Systems with Applications, Vol. 37, pp. 2182–2185 (2010)

[5] Babaoglu I., Findik O., Ulker E., "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine", Expert Systems with Applications, Vol. 37, pp. 3177–3183 (2010)

[6] Candelieri A., Conforti D, "A Hyper-Solution Framework for SVM Classification: Application for Predicting Destabilizations in Chronic Heart Failure Patients", The Open Medical Informatics Journal, Vol. 4, pp. 136-140 (2010)

[7] Xie J., Wang C., "Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases", Expert Systems with Applications, Vol. 38, pp. 5809–5815 (2011)

[8] Hu1 V W.,Steinberg M E., "Novel clustering of items from the Autism Diagnostic Interview-Revised to define phenotypes within autism spectrum disorders", Autism Res, Vol. 2(2), pp. 67–77 (2009)

[9] K. Srinivas, B. Kavihta Rani, A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering (IJCSE), Vol. 02(2), pp. 250-255 (2010)

[10] Thakkar B A., Hasan M I., Desai M A., "Healthcare Decision Support System for Swine Flu Prediction using Naïve Bayes Classifier", International Conference on Advances in Recent Technologies in Communication and Computing, pp. 101-105 (2010)

[11] Jiao Y., Chen R., Ke X., Chu K, Lu Z., Herskovits E H., "Predictive models of autism spectrum disorder based on brain regional cortical thickness", Neuroimage, Vol. 50(2),pp. 589–599 (2010)

[12] A. Vecsei, G. Amann, S. Hegenbart, M. Liedlgruber, A. Uhl, "Automated Marsh-like classification of celiac disease in children using local texture operators", Computers in Biology and Medicine, Vol.41, pp.313–325 (2011)

[13] Chen H L., Liu D Y., Yang B., Liu J., Wang G., "A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis", Expert Systems with Applications, Vol. 38, pp. 11796–11803 (2011)

[14] Mostafa F G., Mohammad S A., "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis", Expert Systems with Applications, Article in Press (2011)

❑❑❑