Cancer Detection Methodology using Adaptive Neuro Fuzzy Inference System and Support Vector Machine Classification

A. Samuel Vijay
Research Scholar (M. Phil. (SSP))
Department of Computer Applications
Alagappa University - Karaikudi India

Abstract—Data mining is a process of extracting hidden knowledge from large volumes of data. It is used intensively in the field of medicine to predict diseases such as heart diseases, lung cancer, breast cancer and more. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain; such patterns are utilized for medical diagnosis. Medical images play an important role in assisting diagnosis and treatment of healthcare management systems. The advancements and large volumes of medical image data become major challenges. In this paper, a novel method to enhance the performance of classifiers Adaptive Neuro-Fuzzy Inference System (ANFIS) and Support Vector Machine (SVM) through feature selection is proposed. The feature selection methods Genetic Algorithm (GA) and Rough Set (RS) are used to select the features. This research work mainly focuses on selecting the prominent features to improve the accuracy of the classification algorithms. Experimentation has been made on various medical images. The performance of the classification algorithms is estimated in terms of increase in accuracy after feature selection.

Key words: Cancer Images, Feature Selection, Classification, Genetic Algorithm, Rough Set, Support Vector Machine, Adaptive Neuro-Fuzzy Inference System

I. INTRODUCTION

The data mining techniques are widely used in the detection of the disease occurrence. The data mining techniques is a very popular research tool for medical researchers to predict the outcome of a disease using the datasets (2). There have been a large number of data mining algorithms rooted in these fields to perform different data analysis tasks. Feature selection is a promising application that is used to increasing the speed of classification technique. The main goal of feature selection is to find a feature subset that produces higher classification accuracy. Feature selection in medical prediction plays a major role as it identifies a prominent feature that influences prediction. The performance of both feature selection and classifiers are applied and tested at liver and breast cancer dataset images.

Cancer is a disease of the human cells. In commonly, cells grow and divide in an ordered way. Occasionally, some cells reproduce themselves in an uncontrolled way and these abnormal cells may grow into a lump that is called the tumour. If cancer has spread, treatment becomes more difficult and a person’s chance of survival is less(1). The earlier cancer is found, the greater the chance of survival. With the rapid advancements in information technology, detection of cancer at an early stage is made possible.

The rest of the paper is structured as follows; section 1 provides the brief introduction in data mining in the field of cancer medical images. Section 2 embraces the various existing papers which are based on cancer detection and classification methods. It is followed by section 3 includes the proposed model of feature selection and classification. The next section 4 contains the experimental results and discussions of the proposed method. Finally, section 5 brings conclusion of the accuracy of cancer classification.

II. RELATED WORK

The following papers are motivated to propose the Cancer Detection Methodology using Adaptive Neuro Fuzzy Inference System and Support Vector Machine Classification mechanism.

In 2015, Hamza Turabieh (3) focused on breast cancer recurrence problem, hybridizing two methodologies, Genetic Algorithm (GA) and Adaptive Neuro Fuzzy Inference System (ANFIS), to develop a good diagnosis system. The author examined their proposed hybrid methodology using classification accuracy, sensitivity, and specificity. The literature survey is also classified in this paper to discuss the classification results.

In 2016, Radhanath Patra and Shankha Mitra Sunani (4) has reviewed of diagnosis of breast cancer database, some of following points were much emphasized. Some of the review paper already predicted about the accuracy of different machine learning algorithm. Due to limitation of ANN various modified form of machine learning algorithm as well as hybrid process were adopted to improve accuracy with least time span .some of paper clearly presented that accuracy was above 95% to 99% . Various data mining process also considered predicting the diagnosis of breast cancer with high accuracy and result showed that decision tree, KNN (K-Nearset Neighbor) algorithm and PSO (Particle Swarm Optimization) is better in terms of accuracy calculation. Feature extraction with some good algorithm should be carried out with certain modified form of machine learning algorithm or data mining process to a large volume of data in a small time span with much higher accuracy.

In 2015, Rajamani.R and Rathika.M (5) provides the overview on liver cancer analysis using Adaptive Neuro Fuzzy Inference System (ANFIS) data mining technique. The author’s consider the input is 2-D CT (Computed Tomography) images. In data preprocessing step, the noise removal in the CT image, segmentation process, morphological operation and the feature extraction techniques has been discussed. The authors have also discussed the study of Adaptive Neuro Fuzzy Inference System for early detection of Liver Cancer in human. They discussed the implementations of this technique or combination of ANFIS with other data mining techniques can be made to help the medical field at early diagnosis of liver cancer.
III. RESEARCH CONTRIBUTIONS

The aim of this paper is to reduce the features using Rough Set and Genetic Algorithm to improve the accuracy of the classification process. The experimentation of the entire processing was performed with two datasets namely, breast cancer dataset and liver dataset. This dataset contains the information about benign and malignant tumours. The entire system model of the RS-GA and ANFIS-SVM proposed methodology is followed.

![Workflow of the Combined RS-GA and ANFIS-SVM](Image)

A. Data Pre-processing

In the pre-processing method, the raw data is first partitioned into three groups like, i) A finite set of objects, ii) the set of attributes such as features and variables, iii) The domain of attribute. For each group in the dataset, a decision system is constructed. Each decision system is subsequently split into two parts: the training dataset and the testing dataset. Each training dataset uses the corresponding input features and falls into normal (+1) and abnormal (-1) classes.

B. Feature selection using Rough Set Technique

The rough set based feature selection technique as the process of finding a subset of features, from the original set of patterns, optimally according to defined criteria. Rough Sets theory is based on the concept of lower and upper approximation of a set, the approximation space and models of sets. An information system can be represented as,

\[ S = (U, A, V, f) \]

Where U is the universe, a finite set of N objects \( (x_1, x_2, \ldots, x_N) \) a nonempty set, A is a finite set of attributes, \( V = \bigcup_{a \in A} V_a \) (where \( V_a \) is a domain of the attribute \( a \)), f: \( U \times A \rightarrow V \) is the total decision function also called as the information function such that \( f(x, a) \in V_a \) for every \( a \in A, x \in U \). B subset of attributes \( B \subseteq Q \) defines an equivalence relation (called an unnoticeable relation) on U.

\[ (A) = \{ (x, y) \in U \; \text{for all} \; a \in B; \; f(x, a) = f(y, a) \} \]

Denoted also by A.

The straightforward feature selection procedures are based on an evaluation of the predictive power of separate features, followed by a ranking of such evaluated features eventually the choice of the first best m features. A criterion applied to an individual feature could be either of the open-loop or closed-loop type. It can be expected that a single feature alone may have a very low predictive power, whereas when put along with others, it may demonstrate a significant predictive power.

C. Feature selection using Genetic Algorithm

A GA is starting by generating a large set of possible solutions to a given problem which means a solution to a problem corresponds to a genome or chromosome in genetics; a large set of possible solutions to a given problem corresponds to a population (6). It evaluates each of those solutions and decides on a “Fitness level” for each solution set. In generally the genetic algorithm contains the following steps:

- Generate an initial population: An initial population is created from a random selection of solutions.
- Evaluation Fitness: A value for fitness is assigned to each solution (chromosome) depending on how close it actually is to solving the problem, thus arriving at the answer of the desired problem.
- Reproduce, Selection, Mutate and crossover: Those chromosomes with a higher fitness value or more likely to reproduce offspring which can mutate/inverse after reproduction.
- Control Next Generation: If the new generation contains a solution that produces an output that is close enough or equal to the desired answer then the problem has been solved. If this is not the case, then the new generation will go through the same process as their parents did. This will continue until a solution is reached.

D. Classification using ANFIS and SVM

1) Adaptive Neuro-Fuzzy Inference System Classifier

ANFIS is based conventional mathematical tool. This tool is also known as fuzzy modelling or fuzzy identification. Generally, it was developed in 1990 (7). ANFIS is a combination of fuzzy logic and neural network which is used to form a hybrid intelligent system that extends the ability to learn automatically and adapt. It can be trained as hybrid learning algorithm, like this kind of hybrid system are used by the following criteria,

- There are no standard methods exist for transforming human knowledge to experience in the rule base and data base of a Fuzzy Inference System (FIS).
- Need for effective methods from tuning the membership functions so as to minimize the output error measure or maximize performance index.

2) Support Vector Machine (SVM) Classifier

Support vector machines are called as supervised learning methods and it is used for classification and regression a task that is from statistical learning theory (8). They belong to a family of generalized linear classification. Normally, a classification task involves training and test sets which consist of data instances. The original idea of SVM was developed for linearly separable data. SVM tunes the capacity of the classification function by maximizing the margin between the training patterns and decision boundary. In the high dimensional feature space, simpler and linear hyper plane classifiers that have a maximal margin between the classes can be obtained.

E. Description of the Algorithm

The proposed algorithm is implemented with the following seven steps.

1) Step 1: Find the unique rows in the medical data set.
2) Step 2: Find the differing features in each instance by comparing it with all other instance.
3) Step 3: Find the overall concatenated comparability matrix
4) Step 4: Find the lower and upper bounds of the data set.
   Lower Bound=\text{Union}(Y \in \text{Ind(Data Set)}; Y \text{ is subset of } X)
   \text{Where } X \text{ is the given data set, Ind is the indiscernibility relation Ind (Data Set) = }
   \{\{1\},\{2,5\},\{3\},\{4,6\}\}
   \text{ X Yes= } \{1,2,3,6\} \text{ X No=} \{4,5\}
   \text{ Lower Bound= } \{1,3,6\}
   \text{ Upper Bound= Union}(Y \in \text{Ind(Data Set)}; Y \text{ intersection } X != \text{empty})
   \text{ Difference set between lower and upper bounds: }
   \{2,5\}
   \text{ Therefore, it is a rough set since difference set is nonempty.}
5) Step 5: Find the sum of occurrence of each feature in the comparability matrix. Here, Feature 1’s occurrence=6
   Feature 2’s occurrence=4 Feature 3’s occurrence=9
6) Step 6: Finding the significant features. Since the comparability matrix is designed based on the features that vary in value, the maximum varying feature gets a greater value.
7) Step 7: Obtain classification accuracy for new dataset with significant features. The new dataset is the one that contains the selected significant features. This data set is passed to the classifiers such as support vector machine (SVM). The original unprocessed data set is also evaluated using the same classifiers for comparison.

IV. RESULTS AND DISCUSSIONS

In this methodology, the breast cancer and liver cancer dataset were used to process feature selection and classification. The description of WBCD and BUPA as follows,

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Number of Dataset</th>
<th>Number of Samples</th>
<th>Number of Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>Wisconsin Breast Cancer Dataset (WBCD)</td>
<td>569</td>
<td>32</td>
</tr>
<tr>
<td>Liver</td>
<td>BUPA Liver Disorder</td>
<td>345</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 1: Cancer Dataset Details

From the datasets, the features are reduced by using the Rough set method and Genetic Algorithm. The reduced number of features gives better classification accuracy. The accuracy was calculated using classification techniques SVM and ANFIS.

The performance of the classifiers was evaluated by percentage of accurately. The accuracy indicates the performance of the classifier. The accuracy calculation formula is derived as follows,

\[
\text{Accuracy (\%) } = \frac{(TP+TN)}{(TP+FN+TN+FP)} \times 100
\]

The classification accuracy of cancer and liver datasets using SVM and ANFIS,

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Datasets</th>
<th>ANFIS (%)</th>
<th>SVM (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver</td>
<td>BUPA</td>
<td>98.20</td>
<td>66.86</td>
</tr>
<tr>
<td>Breast</td>
<td>WBCD</td>
<td>98.92</td>
<td>93.02</td>
</tr>
</tbody>
</table>

Table 2: Classification Accuracy for Liver and Breast cancer using ANFIS and SVM

From the above table, it is evident that the proposed novel method provides better results.

V. CONCLUSION

The perfect accuracy of cancer classification is important for human’s life. Normally, using diagnosis tool, most of the researchers are interested in Artificial Intelligence (AI) classification techniques to classify cancer. The main scope of this research is conducted in order to compare the performance of two feature selection methods and two AI classification techniques namely, SVM and ANFIS in classifying cancer data. This paper focuses on selecting the main features to improve the accuracy of the classification algorithms. Both the SVM and ANFIS technique are effective in order to classify cancer data.

REFERENCES