Analysing Big Data to Build Knowledge Based System for Early Detection of Ovarian Cancer

P. Yasodha* and N. R. Anathanarayan\textsuperscript{2}

\textsuperscript{1}Department of Computer Science, Pachaiyappa’s College for Women, Kanchipuram - 631501, Tamil Nadu, India; yasodhaphd@gmail.com

\textsuperscript{2}Department of CSA, SCSMV University, Kanchipuram - 631561, Tamil Nadu, India

Abstract

Big data analysis plays a crucial role in the health care for early diagnosis of fatal disease. The data mining techniques are widely used for data analysis problem to discover valuable knowledge from a large amount of data. This paper uses the data mining methods such as feature selection and classification to provide a predictive model for ovarian cancer detection. A huge amount of dataset is gathered to build knowledge based system. Rough set theory is utilized to find the data reliance and reduce the feature set contained in the data set. The Hybrid Particle Genetic Swarm Optimization (PGSO) is used to optimize the selected features to efficiently classify the ovarian cancer, either normal or early or different stages of ovarian cancer. Multi class SVM is adopted as the classifier to classify normal or different stages of ovarian cancer using the optimized feature set. The experiment is done on different ovarian cancer dataset and the proposed system has obtained better results for all datasets.

Keywords: Big Data Analysis, Genetic Algorithm (GA), Multi Class Support vector Machine (SVM), Particle Swarm Optimization (PSO), Rough Set Theory, Ovarian Cancer

1. Introduction

Ovarian cancer is common disease for women and it ranks fourth among other cancer for high death rate. The death rate of ovarian cancer is due to the most of the women were unaware of the cancer until the disease has advanced to Stage III or Stage IV\textsuperscript{1}. Early detection is a key to reduce the death rate of Ovarian Cancer. An accurate and reliable diagnosis process might followed by the physician for early diagnosis.

Big data analysis plays a crucial role in the health care for early diagnosis of fatal disease\textsuperscript{2}. Analysing a large amount of data, which is generated from various real time patient records produce a lot of potential information for creating quality health care at reduced costs\textsuperscript{3}. Traditional data analysis techniques have become inadequate for processing such huge volume of data\textsuperscript{3}. Knowledge discovery of data is a new technique that encompasses a variety of pattern recognition, statistical analysis and machine learning techniques, which exploit the knowledge from huge amount of recorded data\textsuperscript{4,5}. Data Mining is commonly alluded as pattern extractions strategy, which assembles and formulates the knowledge from tremendous information\textsuperscript{4,5}. Data mining strategies have officially demonstrated to wide range of medicine, including prognosis, diagnosis and treatment.

This paper uses the data mining technique to obtain a knowledge based system for early ovarian cancer detection from an unorganized data set. The major commitment of the paper is as per the following: 1. Initially, the rough set theory is applied to the dataset to determine the data dependencies and to minimize the feature set. 2. The Hybrid Particle Genetic Swarm Optimization (PGSO) is used to optimize the rough set feature reduction to effectively classify the ovarian cancer tumors either normal or abnormal and 3. Multi Class Support Vector Machine (SVM) is used for classifying the different stages of ovarian cancer and non-ovarian cancer.

The rest of the paper is organized as follows: Section II describes the recent related works about the system. Section III describes proposed knowledge base system. Experimental results and analysis of the proposed system.
is described in Section IV. Finally, Section V renders the conclusions.

2. Related Work

A model is proposed in⁹ for early detection and correct diagnosis of lung cancer disease to save the life of patient using the data mining classification methods. The historical lung cancer database is used to extract the concealed information to predict the patients with lung cancer or not. Three different classifier are utilized such as Naive Bayes took after by if-then rule, Neural Network and Decision trees and the performance of the Naive Bayes shows that it is obviously better than the other two.

The data mining techniques is used in¹⁰ to improve the breast cancer diagnosis and prognosis. The different data mining techniques discussed to find the effective classifier for the breast cancer diagnosis. The UCI machine learning data set and SEER data set used for the research and the decision tree attained better accuracy than other classifiers to predict the disease.

The data mining concept is used in¹¹ to predict the head and neck cancer. The data set is collected from the various diagnostic centres with both the cancer and non-cancer patient. A different classification technique has been applied for the data set and C4.5 obtained higher accuracy rate than other classifiers.

The performance of the Decision tree classifier-CART is analysed in¹² with and without feature selection on different breast cancer data sets. The experimental result shows that classification accuracy is improved with feature selection by removing the irrelevant features in the datasets.

The feature selection approach is analysed in¹³ for classification and a new approach is proposed for feature selection. The database contains a large number of features, so dimensionality reduction is used using the association rules and correlation attributes features selection. After removal of unwanted attributes the accuracy of different classifiers is checked using the reduced feature set.

A modified REF-SVM based feature selection method is proposed in¹⁴ for classification problem. It is greedy method to find the best possible combination for classification to improve the quality of the final classifier. The of experiments is done on dataset from UCI Machine Learning Repository and the promising results has been obtained by contributing local search on the classification process.

A new gene selection technique is proposed in¹⁵, which combines Techniques for Order Preference by Similarity to an Ideal Solution and F-Score method to select relevant gene. The selected gene is fed in to four different classifiers such as K-Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Machine (SVM) and Naive Bayes (NB). The SVM obtained better classification accuracy with the reduced feature set.

A Multi objective Firefly Algorithm technique is proposed in¹⁶ for Multiclass Gene Selection. The method optimizes the multiple fire flies in the multiple class-specific statistics to select the genes. The method is compared with the existing gene selection methods and the results show that the method achieves high classification accuracy with less complexity than the existing methods.

3. Knowledge based System Utilizing Data Mining Concepts for Early Detection of Ovarian Cancer

The goal of this paper is to provide a knowledge base system to classify the ovarian cancer and normal case. The dataset contains N number of features, to determine a relevant feature to portray the dataset. The rough set theory is well known method for feature subset selection. The Hybrid Particle Genetic Swarm Optimization (PGSO) is used to optimize the rough set feature reduction to effectively classify the data into normal or abnormal. The Multi class SVM is used in the classification stage to classify the data into normal and abnormal case. The Figure 1 sows the proposed knowledge based system for early detection of ovarian cancer.

![Figure 1. Knowledge based system for early detection of ovarian cancer.](image-url)
3.1 Rough Set Theory

The rough set technique is a mathematical model introduced by Pawlak for dealing with uncertainty\textsuperscript{17,18}. The rough set theory minimizes the feature set and reducts a significant feature set, which is capable of detecting the visible points by the original of Information System (IS).

Let IS = (U, A), where A(\neq \emptyset) be the set of attributes and U(\neq \emptyset) is termed as Universe and it is a non-empty finite set of all objects. A reduct of A is a reduced set of attribute B \subseteq A, such that non empty subset of A.

Indiscernibility Relation (IND (B)) is a relation on U. Consider two objects x\textsubscript{i} and x\textsubscript{j} \in U are viewed to be indiscernible by the set of attributes B in A, if and only if a(x\textsubscript{i}) = a(x\textsubscript{j}), a \in A-B. i.e.

\[ \text{IND}(B) = \{ (x\textsubscript{i}, x\textsubscript{j}) \in U \mid \forall a \in Ba(x\textsubscript{i}) = a(x\textsubscript{j}) \} \] (1)

A subset is obtained without containing any dispensable attributes and that subset is known as a reducts. The set of all reducts in IS is indicated as RED.

3.2 Particle Genetic Swarm Optimization (PGSO) based Rough Set Theory

The GA and PSO is widely used Evolutionary algorithm because of its simple process with optimized solution. In this paper, the hybrid PGSO procedure is followed by combining the PSO and GA for optimized feature selection. The GA is embedded within the PSO to improve the PSO by serving as a local optimizer at each iteration.

3.2.1 Particle Swarm Optimization (PSO)

In PSO, each particle represents a candidate solution of a population, simultaneously coexist and evolve based on knowledge sharing with neighbouring particles\textsuperscript{19}. Each particle flies on the problem search space and based on the directed velocity vector, it will generate a solution. Each particle changes its velocity to determine the better position by using its own flying knowledge for the best position memory found in the earlier flights and experience of neighbouring particles as the best determined solution of the population. The best position determined by the particles is represented as P\textsubscript{best} and the tendency to move forwards the previous best position of the neighbourhood's P\textsubscript{best}. The velocity of the particle is updated using the following equation

\[ v'_{i\text{maxcount}} = w v_i' + c_r (p'_i - x'_i) + c_r (p^g_i - x'_i) \] (2)

Where x\textsubscript{t} represents the current position of particle i, p\textsubscript{i} is the current best position determined by particle i, p\textsubscript{t} is the global best position determined among all particles in the problem space up to iteration count i, c\textsubscript{r} and c\textsubscript{r} represents the cognitive and social scaling parameters, r\textsubscript{1} and r\textsubscript{2} are random numbers distributed uniformly in the interval (0,1).

w\textsubscript{t} is the particle inertia, which minimize the search area dynamically,

\[ w\textsubscript{t} = (w\textsubscript{max} - w\textsubscript{min}) \times \frac{\text{maxcount} - t}{\text{maxcount}} + w\textsubscript{min} \] (3)

Where w\textsubscript{max} and w\textsubscript{min} indicates the maximum and minimum of w\textsubscript{t}, respectively max count represents the maximum iteration and t represents the current iteration number.

The particle position updated according to the following equation,

\[ x'_{i\text{maxcount}} = \begin{cases} 1.\text{rand}() < \text{sig}(v'_i) & \text{if} \text{rand}() \geq \text{sig}(v'_i) \\ 0.\text{rand}() \geq \text{sig}(v'_i) & \text{otherwise} \end{cases} \] (4)

\[ \text{sig}(x) = \frac{1}{1 + e^{-x}} \] (5)

Particle i fly towards the new position according to equation 1 and 2. In this way all particles determines the new positions and update their individual best position p\textsubscript{i} and global best position p\textsubscript{g} of the swarm. This process will be continued until maximum iteration reached.

Consider an Information System (IS) = (U, A) and A = (X \cup Y) where X is a non-empty finite set of condition attributes and Y is a non-empty finite set of decision attributes, such that RED IS \subseteq X. The objective function of particle i at position x is determined by the following equation,

\[ f(x'_i) = a \times y^c(x'_i, Y) + \beta \times \frac{|x'_i - x^c_i|}{|x|} \] (6)

Where y^c(x'_i, Y) is the classification quality of particle condition attribute set x'_i, which contains the RED, and relative to decision table Y, and it is shown in the following equation,

\[ y^c(x'_i, Y) = \frac{d_{\text{sup}}}{d_i} \] (7)
Where $d_{\text{RED}}$ represents a dependency degree of RED on $Y$ and $d_Y$ represents the dependency degree of $X$ on $Y$. \( x_i \) is the '1' number of length of selected feature subset for particle $X_i^t$, while population of solutions $P$ (number of particles in the population) is at iteration count $t$. $|X|$ is the total number of condition attributes. The parameter $\alpha \in [0,1]$ and $\beta = 1 - \alpha$ represents importance of classification quality and subset length.

### 3.2.2 Genetic Algorithm

Genetic Algorithm (GA) is one of the computational models used widely because of its evolution. GA optimization technique contains selection, crossover and mutation operations to a population of completing problem solutions. After the three operations are applied a new generation of the populations will be generated and at the same time the GA will generate a set of chromosomes randomly at the space. The fitness value will be calculated for the chromosomes and the chromosome with a higher fitness value will be kept and the same operation will be performed until a fixed number of iterations are reached.

To improve the PSO performance, the GA is used as a local optimizer at each iteration\(^\text{20,21}\). After the initial population are created, the operation such as selection, crossover and mutation will be applied to the initially created particles. Choose two particles randomly and determine the relative difference for those two particles by the following equation,

$$d = \frac{|f_1 - f_2|}{f_1 + f_2}$$

(8)

Where $f_1$ and $f_2$ are the fitness value of particle 1 and particle 2. According to the relative difference value the cross over operation is chosen and it is defined in the following equation,

$$\begin{cases} 
\text{If } f_1 < f_2 \text{ and } R < d, \text{ then crossover operation is done on particle 2} \\
\text{If } f_1 \text{ and } R > d, \text{ normal crossover operation is chosen} \\
\text{If } f_1 > f_2 \text{ and } R < d, \text{ then crossover operation is done on particle 1} \\
\text{if } f_1 > f_2 \text{ and } R > r, \text{ normal crossover operation is chosen}
\end{cases}$$

(9)

Mutation operation is done on the two selected particles and it is set to probability of $1/n$ ($n$= number of particles).

The process will be done until the maximum iteration is reached. Then the PSO based optimization is performed for the particles. The PGSO based rough set algorithm is given in the Figure 2.

**Figure 2.** Algorithm for optimized feature set.

Input: C, the set of all condition attributes, max count
Output: Reduct (RED)

Step 1: Initializing position, velocity, $c_1,c_2$.
Step 2: For $i=1$ to $P$
  - Obtaining the interia weight for particle$i$ by using equation (3)
  - Calculate the fitness (objective) function for particle$i$ by using the equation (6)
End For
Step 3: Initialize $P_i = P_i^{min}$ and $Gbest = f_{best}^t\left(p_i^{best}\right)$
Step 4: For $i=1$ to Maxcount
  - Select two random particle and calculate the relative difference using equation 8
  - Select the cross over operation according to the equation (9)
  - Mutuate (particle$x$), Mutuate (particle$y$)
End For
Step 5: while ($t<$ maxcount)
  - For $i=1$ to $P$
    - Compute the fitness function for particle$i$ using eq (6)
    - If $pbest_i < f(x_i)$
      - $pbest_i < f(x_i)$
      - $p_i = x_i$
    End if
  - Find the $f^{max}_{p_i^t}\left(p_i^{min}\right) = \max\left\{p_1^t, p_2^t, \ldots p_i^t\right\}$
    - if $pbest_i < f_{p_i}^{min}$
      - $p_i = p_i^{min}$ and $Gbest = f_{p_i}^{min}\left(p_i^{min}\right)$
    End if
  - Evaluate the velocity for particle$i$ by equation (2)
  - Update particle position by using equation (4)
End For
End while
Step 6: Redact RED

---

Indian Journal of Science and Technology
3.2.3 Classification

The effective feature set is built by eliminating the noisy feature and it improves the classification accuracy. In this paper, the Support Vector Machine is used as a classifier and it is originally designed for binary classification. Currently there exist two types of approaches for multiclass SVM. The outline to illuminate multiclass SVM problems in one step has parameters proportional to the k classes. Therefore, for multiclass SVM techniques, either a number of binary classifiers have to be constructed\(^2\). There are four approaches for multiclass classification based on binary classification: One Against One (OAO), One Against All (OAA), Fuzzy Decision Function (FDF) and Decision Directed Acyclic Graph (DDAG) SVM.

The One Against All (OAA) method is used in this paper to classify different stages of ovarian cancer cases and the normal cases. In OAA, k SVM models are constructed. The \(i^{th}\) SVM is learned with all of the training examples in the \(i^{th}\) class with positive labels and all other classes’ negative labels. The final output of the OAA is the class that relates to the SVM with the highest output value\(^3\). Hence, by illuminating the optimization problem of SVM using all the training examples in the dataset, \(i^{th}\) SVM decision function is

\[
D_i(x) = w'_i \phi(x) + b_i
\]  

(10)

Where \(x\) is the input vector, which is assigned to the \(i^{th}\) class corresponds to the largest value of the decision function; \(w\) is the vector in the feature space, \(\phi(x)\) is mapping function and \(b_i\) is a scalar.

Sample \(x\) is classified into the class as defined in the following equation,

\[
x = \arg\max_{i=1,...,k} (D_i(x))
\]  

(11)

4. Experimental Results and Analysis

The gene and blood test are normally used indicators for the ovarian cancer, so in this paper two data sets are considered.

4.1 Dataset

4.1.1 Micro Array Gene Expression

A normalized gene expression data and it consists of 493 instances and it is taken from the TCGA portal (http://tcga-data.Nci.nih.gov/). The dataset composed of 12,042 genes without any missing values. The dataset classified 39 cases as Early (Stage I and Stage II) and 454 cases as late stage (Stage III and Stage IV).

4.1.2 Blood Assays

The Data set 2 has been taken from the Singapore National University Hospital (NUH) (http://www.nuh.com.sg/#) and the dataset is based on the blood test results of 172 instances and it consists of 28 features. The dataset contains 23 normal, 10 of borderline, 19 of early (Stage I and Stage II) and 42 of late stage (Stage I and Stage II).

4.2 Performance Metrics

The system performance is computed in terms of mean absolute error, classification accuracy, sensitivity, root mean square error and specificity respectively.

Sensitivity is the ability of the test to identify the class correctly among other classes belonging to the classes of the ovarian cancer dataset.

\[
\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}
\]  

(12)

Where True positive is correctly identified and False Negative is incorrectly rejected.

Specificity is ability of the tests to exclude a condition correctly.

\[
\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}
\]  

(13)

Where True negative is correctly rejected and False positive is incorrectly identified.

The classification accuracy is directly proportional to the correctly classified objects and it is determined by the following formula,

\[
\text{Classification Accuracy} = \frac{\text{Number of instances classified correctly}}{\text{Total number of instances}}
\]  

(14)

The Mean absolute error is the average of the difference between the predicted and actual value in all test cases.

\[
\text{MAE} = \frac{|a_1 - c_1| + |a_2 - c_2| + \ldots + |a_n - c_n|}{n}
\]  

(15)

The root mean squared error yields the error value the same dimensionality as the actual and predicted values.
\[ RMSE = \sqrt{\frac{1}{n} \left( (a_1 - c_1)^2 + (a_2 - c_2)^2 + \ldots + (a_n - c_n)^2 \right)} \] (16)

Where the \( c_i \) is the computed value and \( a_i \) is the corresponding corrected value.

### 4.3 Discussion

Figure 3 shows the comparison of the Multiclass SVM, ANN and Naive Bayes in terms mean absolute error for two ovarian datasets. The multiclass SVM classifier incurred a minimum error rate by using the optimized feature set for the two data set when compared to ANN and Naive Bayes.

Figure 4 shows the comparison of the Multiclass SVM, ANN and Naive Bayes in terms of root mean square error for two datasets. The multiclass SVM class obtained a minimum error rate by using the optimized feature set for the two data set when compared to ANN and Naive Bayes.

### Table 1. Classification Accuracy for different Ovarian cancer dataset

| Classifier     | Dataset I    | Dataset II   |
|----------------|--------------|--------------|
|                | Sensitivity  | Specificity  | Accuracy | Sensitivity  | Specificity  | Accuracy |
| Multiclass SVM | 97           | 95.4         | 96       | 99           | 96.7         | 98       |
| ANN            | 95           | 91.6         | 93       | 97           | 92.9         | 95       |
| Naive Bayes    | 93           | 87           | 90       | 96           | 91.4         | 93       |
The classification accuracy of the proposed system is compared with the ANN and Naive Bayes and it has achieved 96 percent accuracy for dataset I and 98 percent accuracy for dataset II and it is graphically represented in the Figure 5. The Table 1 gives the Sensitivity, specificity, and accuracy values for classifying normal, early stage and late stage for different ovarian cancer dataset and the proposed system achieves better accuracy than the ANN and Naive Bayes.

5. Conclusion

Ovarian cancer diagnosis is an important research by considering significant increase in death rate, so early detection and accurate staging will help the physician for correct diagnosis procedure. The paper proposed a knowledge based system using the data mining concepts such as feature selection and classification. The optimized feature selection process is achieved by the hybrid PGSO based rough set theory. The Multi class SVM is used for classifying the normal, early stage and late stage cases using the optimized attribute set. The system performance has been tested using the two different ovarian cancer data sets and the proposed Knowledge based system attain better accuracy than other classifiers using the reduced features. The system has tendency to provide better suggestion to the physician for diagnostic process.

6. References

1. Tan TZ, Quek C, See Ng G, Razvi K. Ovarian Cancer diagnosis with complementary learning fuzzy neural network. Artificial Intelligence System. 2008; 43:207–22.
2. Peisker A, Dalai S. Data analytics for rural development. Indian Journal of Science and Technology. 2015; 8:50–60.
3. Radhakrishnan A, Kalmadi K. Big data medical engine in the cloud (BDMeIC): Your new health doctor. Infosys Lab Briefings. 2013; 11(1).
4. Lim CS, Lee W, Joon YY, Shon K. Creating values from a noisy accumulated contents based on data analysis. Indian Journal of Science and Technology. 2015; 8:58–63.
5. Mohammadzadeh N, Safdari R, Mohammadzadeh F. Using intelligent data analysis in cancer care: Benefits and challenges. Journal of Health Informatics in Developing Countries. 2014; 8(2):66–72.
6. Noh K-S. Plan for vitalisation of application of big data for e-learning in South Korea. Indian Journal of Science and Technology. 2015; 8:149–55.
7. Ramesh Kumar KK, Anbumani A. Medical image segmentation using multifractal analysis. International Journal of Inventions in Computer Science and Engineering. 2014; 1(3):2348–431.
8. Sharma R, Sharma P. A novel approach towards big data challenges. International Journal of Innovative Computer Science and Engineering. 2014; 1(2):28–34.
9. Krishnaiah V, Narsimha G, Subhash Chandra N. Diagnosis of lung cancer prediction system using data mining classification techniques. IJCSDIT. 2013; 4(1):39–45.
10. Kharya S. Using data mining techniques for diagnosis and prognosis of cancer disease. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT). 2012; 2(5):55–66.
11. Suji RJ, Rajagopalan SP. An automatic oral cancer classification using data mining techniques. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE). 2013; 2(10):3579–3765.
12. Lavanya D. Analysis of feature selection with classification: Breast cancer datasets. Indian Journal of Computer Science and Engineering (IJCSE). 2011; 2(5):756–63.
13. Rajeswari K, Vaithiyananthan V, Pede SV. Feature selection for classification in medical data mining. International Journal of Emerging Trends and Technology in Computer Science (IJETITCS). 2013; 2(2):492–7.
14. Samb ML, Camara F, Ndiaye S, Slimani Y, Esseghi MA. A novel RFE-SVM-based feature selection approach for classification. International Journal of Advanced Science and Technology. 2012; 43:27–36.
15. Abd-El Fattah M, Khedr WI, Sallam KM. A TOPSIS based method for gene selection for cancer classification. Int J Comput Appl. 2013; 67(17):39–44.
16. Manoharan GV, Shanmugakshmi R. Multi-objective firefly algorithm for multi-class gene selection. Indian Journal of Science and Technology. 2015; 8:27–34.
17. Nizar Banu PK, Hannah Inbarani H. Performance evaluation of hybridized rough set based unsupervised approaches for gene selection. International Journal of Computational Intelligence and Informatics. 2012; 2(2):132–41.
18. Srimani PK, Koti MS. Knowledge discovery in medical data by using rough set rule induction algorithms. Indian Journal of Science and Technology. 2014; 7(7):905–15.
19. Arafat H, Barakat S, Goweda AF. Using intelligent techniques for breast cancer classification. International Journal of Emerging Trends and Technology in Computer Science (IJETITCS). 2012; 1(3):26–36.
20. Rouhi F, Efatnejad R. Unit commitment in power system T by combination of Dynamic Programming (DP), Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). Indian Journal of Science and Technology. 2015; 8(2):134–41.
21. Chaung L-Y, Yang C-H, Li J-C, Yang C-H. A hybrid BP-SO-CGA approach for gene selection and classification of microarray data. Journal of Computational Biology. 2012; 19(1):68–82.
22. Kohli N, Verma NK. Arrhythmia classification using SVM with selected features.IJEST. 2011; 3(8):122–31.
23. Bhat HF, Wani MA. Modified one-against-all algorithm based on support vector machine. Int J Adv Res Comput Sci Software Eng. 2013; 3(12):972–5.