ACTA SCIENTIFIC MEDICAL SCIENCES

Volume 3 Issue 7 July 2019

Research Article

Classification of Cancer Micro-Array Data with Feature Selection using Swarm Intelligence Techniques

R Kaja Nisha[1] and A Sheik Abdullah[2]*

[1]Department of Computer Applications, Madurai Kamaraj University, Madurai, India

[2]Department of Information Technology, Thiagarajar College of Engineering, Madurai, India

***Corresponding Author:** A Sheik Abdullah, Department of Information Technology, Thiagarajar College of Engineering, Madurai, India.

Received: April 09, 2019; **Published:** June 12, 2019

Abstract

Data mining is the process of using computational algorithms and tools to automatically discover useful information in large data archives. Data mining techniques are deployed to score large databases in order to find novel and useful patterns that might otherwise remain unknown. They also can be used to predict the outcome of a future observation or to assess the potential risk in a disease situation. Recent advances in data generation devices, data acquisition, and storage technology in the life sciences have enabled biomedical research and healthcare organizations to accumulate vast amounts of heterogeneous data that is key to important new discoveries or therapeutic interventions. Extracting useful information has proven extremely challenging however. Traditional data analysis and mining tools and techniques often cannot be used because of the massive size of a data set and the non-traditional nature of the biomedical data, compared to those encountered in financial and commercial sectors.

Data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining involves many different algorithms to accomplish different tasks. All of these algorithms attempt to fit a model that is closest to the characteristics of the data being examined. In recent years, data mining has been used widely in the areas engineering suchas bioinformatics, genetics, medicine, education and electrical power engineering. It aims to find out how the changes in an individual's DNA sequence affects the risks of developing common diseases such as cancer, which is of great importance to improving methods of diagnosing, preventing, and treating these diseases.

Keywords: Cancer; Swarm; Techniques; Data

Introduction

Medical data analysis plays a significant role in data prediction and classification. Determining the significant risk and classifying the predominant risk factors is considered to be the major challenge and need in medical domain. Region based analysis of risk factors and its co-morbidities make the medical experts to have knowledge on data driven decision-making [11].

Upon considering various diseases and its risk factors such as diabetic disease, heart disease and cancer disease the most significant need is for classifying the data attributes to determine the most significant risk factors [12]. According to the statistical analysis by World Health Organization (WHO) there are about 1,92,370 new subjects were added during the year of 2009 which were found to be malignant specifying women were prone to the disease.

Treatment specific analysis has been carried out by various health care organizations to diagnose the disease through popular screening and treatment strategies [13]. Technology in medical domain seems to be an important factor for effective analysis of disease specific risk factors which then makes for suitable decision making process. However early detection and treatment analysis makes some curable phenomenon for cancer classification and prediction [14-16].

Meanwhile, microarray cancer data makes the process of visualizing and analyzing thousands of genes in parallel which then maps enormous gene selection paradigms for classification and prediction. Machine learning techniques in data classification and prediction for medical data provides the significant way of analyzing the genes to at most level of data classification process. Thereby, the subset of relevant features gets classified with the most significant classification algorithm which produces the maximum level of accuracy.

The mechanism of feature selection process provides the following benefits:

- Interpretability
- Reduced training
- Reduced time in over fitting

Subset selection in the given set of features also needs to be considered for data classification and prediction process. Mining and analyzing the feature subsets need to be considered for predictive analysis and the most predominant risk selection [18,19]. The following are the methods corresponding to feature subset selection:

- Filter method
- Wrapper method
- Embedded method.

Classification in data mining is of two types

1. Supervised learning
2. Unsupervised learning

Supervised learning deals with the mechanism of classifying the data if the data contains the class label. Class label plays an important role in data classification process. In the case of unsupervised learning the class label is not known for doing the classification process. Some of the data classification algorithms are

- ID3 Algorithm
- Naïve Bayes Classification
- C4.5 Algorithm
- Random Forest
- Decision tree Learning

This paper mainly deals with analysis of cancer data using feature selection and classification process to determine and classify the best set of genes for predictive analysis.

Literature Review

In this paper they present the usage of Integer –Coded Genetic Algorithm (ICGA) and Particle Swarm Optimization (PSO) coupled with Neural Network based Extreme Learning Machine (ELM) for Gene Selection and Cancer Classification. Further the performance of ICGA-PSO-ELM has been evaluated. The selected gene set is a randomly selected set of genes related to tumor cells [1]. This thesis presents a new Feature selection method which uses backward elimination process [20]. The proposed approach is used to evaluate the feature ranking score from the statistical analysis of weights of vectors of multiple linear SVM on the training data. The dataset used here is Gene Expression Dataset for cancer classification [2].

This paper introduces a novel and efficient feature selection approach based on statistically defined effective range of features for every class termed as ERGS (Effective Range based Gene Selection) which is used in ranking the genes and also helps in identifying the most relevant genes responsible for the diseases like lung cancer, colon tumor etc. The basic principle behind ERGS is that higher weight is given to the feature that discriminates the classes clearly [21,22]. The dataset used here is Gene Expression Data. The classifiers used here are Nave Bayes Classifier (NBC) and Support Vector Machine (SVM) [3].

In this paper, a novel approach for ranking features based on the predictive quality using properties unique to learning algorithms based on the group method of data handling (GMDH) has been proposed. This method helps in determining optimum feature subset. The datasets used here were breast cancer and heart disease dataset. The comparisons were made with other feature ranking and selection methods which lead to the improvement of classification accuracy [4].

This paper portrays the major challenges in microarray analysis especially in gene expression data. Supervised machine learning techniques were used with microarray datasets to build classification models for improving the diagnostic of different diseases [23]. A comparison of the classification accuracy among nine decision tree methods was made. In addition to the comparative analyses, the evaluation of the behaviors of the methods with/without applying attribute selection (A.S.) was carried out [5].

Proposed system

Problem statement

An efficient feature selection method that can be used for handling extremely high dimensional gene expression data for finding optimal feature subset.

Problem description

The problem of classifying and analyzing the cancer detection mainly focused on the determination of cancer and its subtypes with a detailed interruptions and interpretation on the observed data. Advancements in the field of medicine with improved methods in cancer classification serves as a benchmark for the analysis of microarray data. It also provides molecular level analysis of gene in which it corresponds to a specific domain of action. Recent studies in literature prove that the overlapping of gene analysis can be classified with the set of available sample and its corresponding attributes.

Algorithm description

Let S be the number of particles in the swarm, each having a position x^i ∈^{n} in the search-space and a velocity vi ∈n. Let pi be the best known position of particle i and let g be the best known position of the entire swarm [24,25]. A basic PSO algorithm is given in the following procedure 1 and flow is illustrated in Figure 1: Procedure 1. PSO Search Algorithm.

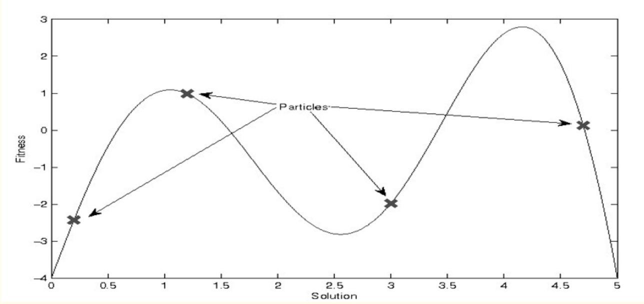


Figure 1: Initial PSO State.

```
For each particle i = 1, ..., S do:
    • Initialize the particle's position with a uniformly distributed random vector: xi ~ U(blo, bup), where blo and bup are the lower and upper boundaries of the search-space.
    • Initialize the particle's best known position to its initial posi tion: pi ← xi
    • If (f(pi) < f(g)) update the swarm's best known position: g ← pi
    • Initialize the particle's velocity: vi ~ U(-|bup-blo|, |bup-blo|)
```

Until a termination criterion is met (e.g. number of iterations performed, or a solution with adequate objective function value is found), repeat:

```
For each particle i = 1, ..., S do:
    • For each dimension d = 1, ..., n do:
        • Pick random numbers: rp, rg ~ U(0,1)
        • Update the particle's velocity: vi,d ← ω vi,d + φp rp (pi, d-xi,d) + φg rg (gd-xi,d)
    • Update the particle's position: xi ← xi + vi
    • If (f(xi) < f(pi)) do:
        • Update the particle's best known position: pi ← xi
        • If (f(pi) < f(g)) update the swarm's best known position: g ← pi
```

Now g holds the best found solution.

Once the velocity for each particle is calculated, each particle's position is updated by applying the new velocity to the particle's previous position:

$$xi(t+1) = xi(t) + vi(t+1)$$

This process is repeated until a stopping condition is met. The stopping conditions include

- A preset number of iterations of the PSO algorithm.
- A number of iterations since the last update of the global best candidate solution.
- A predefined target fitness value.

Classification using decision trees (J48)

Classification using J48 method is selected since it showed a higher accuracy when compared to the other methods. This was performed without feature selection mechanism. The accuracy was about 60.84%. This technique is performed using WEKA. In order to improve the accuracy; the classification is performed on selected features, derived from both Particle Swarm Optimization algorithm and Ant colony Optimization algorithm [26,27].

Data set

To evaluate the performance of the proposed method, publicly available gene microarray data sets were selected from literatures. These data sets are often used to validate the performance of classifiers and gene selectors.

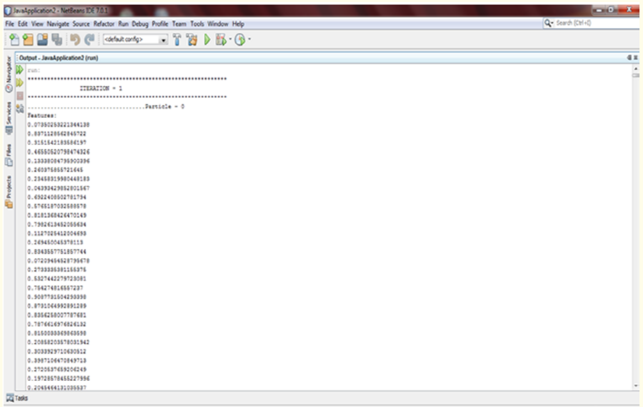
Breast cancer

The quantities of genes and samples in this dataset are 24,481 and 97, respectively. Among these samples, 46 of which are from patients who had labeled as relapse, the rest 51 samples are from patients who remained healthy and regarded as non-relapse.

Results and Discussion

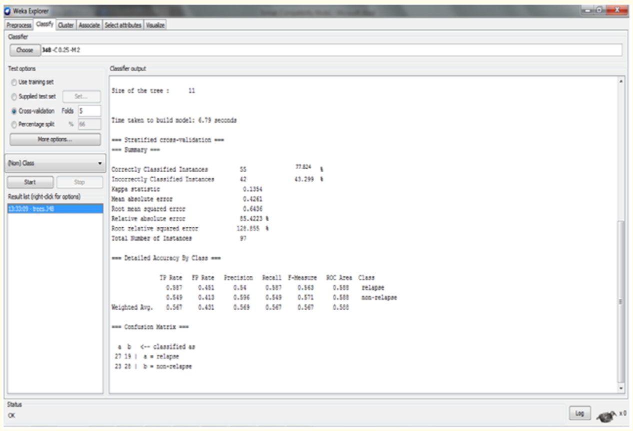
Results for the cancer dataset have been evaluated and analyzed using PSO and ACO algorithm with decision trees for classification. The following are the metrics used for evaluating the performance of the model:

- Accuracy
- Sensitivity
- Specificity
- Precision
- Recall


JavaApplication2 - NetBeans IDE 7.0.1
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
<default config>
Output - JavaApplication2 (run)
Navigator
Services
Files
Projects
Tasks

```
run:
************************************************************
              ITERATION = 1
************************************************************
..................................Particle - 0
Features:
0.07350253221344138
0.8971128562845722
0.3151542183586197
0.46550520798474326
0.13338084796900396
0.260375855721645
0.23458319980448183
0.04393429852801567
0.6922408502781794
0.5765187032588578
0.8181368426470149
0.7982613452055634
0.1127025412004693
0.269450045378113
0.8343557751857744
0.07209464528795678
0.2733335381155375
0.5327442279723081
0.754274816557237
0.9087731504293398
0.8731064992891289
0.8356258007787681
0.7876616976826132
0.8150088369868598
0.20858203578031942
0.3033929710630512
0.3987106470849713
0.2720537459204249
0.19728578455227996
0.2065464131035537
```

Figure 2: Feature Selection using PSO.



Choose J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 5
○ Percentage split % 66
More options...

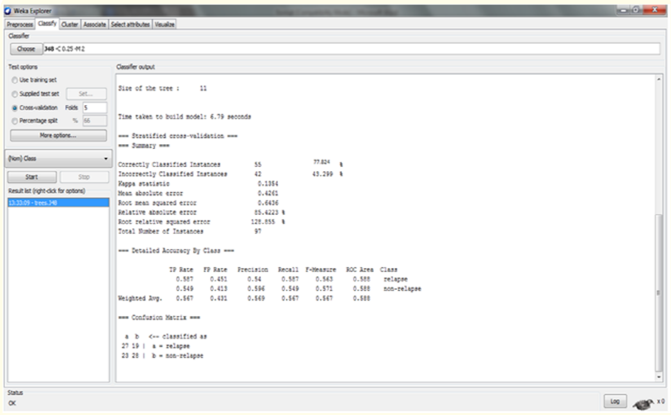
(Nom) Class

13:23:09 - trees.J48

OK

```
Size of the tree :      11


Time taken to build model: 6.79 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        55               77.824  %
Incorrectly Classified Instances      42               43.299  %
Kappa statistic                        0.1354
Mean absolute error                    0.4261
Root mean squared error                0.6436
Relative absolute error               85.4223 %
Root relative squared error          128.855  %
Total Number of Instances             97

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.587     0.451      0.54      0.587     0.563      0.588    relapse
                 0.549     0.413      0.596     0.549     0.571      0.588    non-relapse
Weighted Avg.    0.567     0.431      0.569     0.567     0.567      0.588

=== Confusion Matrix ===

  a  b   <-- classified as
 27 19 |  a = relapse
 23 28 |  b = non-relapse
```

Figure 3: Classification Accuracy without Feature Selection.



Choose J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 5
○ Percentage split % 66
More options...

(Nom) Class

13:23:09 - trees.J48

OK

```
Size of the tree :      11


Time taken to build model: 6.79 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        55               77.824  %
Incorrectly Classified Instances      42               43.299  %
Kappa statistic                        0.1354
Mean absolute error                    0.4261
Root mean squared error                0.6436
Relative absolute error               85.4223 %
Root relative squared error          128.855  %
Total Number of Instances             97

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.587     0.451      0.54      0.587     0.563      0.588    relapse
                 0.549     0.413      0.596     0.549     0.571      0.588    non-relapse
Weighted Avg.    0.567     0.431      0.569     0.567     0.567      0.588

=== Confusion Matrix ===

  a  b   <-- classified as
 27 19 |  a = relapse
 23 28 |  b = non-relapse
```

Figure 4: Classification accuracy with feature selection using PSO.

Discussion

We analyzed the results from various classification techniques like Random Forest, J48 etc. and selected J48 as our classifier since it showed a higher accuracy when compared to others [10]. In addition the accuracy was evaluated by using the proposed feature selection algorithms, PSO (Particle Swarm optimization) and ACO (Ant Colony Optimization) on a well-known breast cancer microarray dataset. The selected subset of features was given as the input for the J48 classifier. We also analyzed that each time the samples selected were different and also the performance improved by using k-fold cross validation. The entire process was repeated several times and a higher accuracy of about 74.226% and 77.82% was obtained by using each of those algorithms respectively [9].

The efficiency of the proposed PSO and ACO algorithms, the classification accuracy obtained using these algorithms is compared with the accuracy obtained without performing feature selection [8]. Fivefold cross validation has been used as the validation strategy to give a relatively comprehensive comparison on the performances. Table 1 represents the classification accuracy obtained by different algorithms without doing feature selection. J48 algorithm showed highest classification accuracy when compared to other methods. For this reason we have chosen J48 as our classification method.

PSO based gene selection and classification results

PSO is used to select a set of attributes from the original 24,482 attributes using fivefold cross validation scheme on the 97 samples. PSO is executed several times and each time different set of genes is selected [7]. Using the selected attributes classification accuracy is calculated using J48algorithm considering 1000 attributes. Table 2 represents the best classification accuracy obtained during each iteration of PSO.

Without performing feature selection for 1000 features we got an accuracy of 60.8247%. By doing feature selection using PSO we obtained a classification accuracy of 74.2268% by selecting 474 genes. From these results, we can say that PSO algorithm improves the classification accuracy by selecting minimum lesser number of genes [6].

Future enhancements

Future work would explore applying these feature selection approach to other learning algorithms such as ACO, forward selection, backward elimination, and to various other datasets.

Conclusion

Feature selection technique plays a significant role in prediction and data classification. The impact of data analysis and classification has been widely used in medical data analysis for the development and analysis of decision support system. Feature reduction is useful with high dimensional data like microarray characterized by large number of features and relatively few instances. This experimental study is used to analyze the classification performances using PSO. We have described important features (genes) and the classification mechanism uses 5 fold cross validation technique. These selected feature improved the overall classification accuracy from 60.82% to above 74%. This can be still improved using the various combinations of feature selection and classification techniques for determination of relevant subset of features.

Bibliography

1. Saras Saraswathi., *et al*. "ICGA-PSO-ELM Approach For Accurate Multiclass Cancer Classification Resulting in Reduced Gene Sets In Which Genes Encoding Secreted Proteins Are Highly Represented". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8.2 (2011).

2. Kai–Bo Duan., *et al*. "Multiple SVM-RFE For Gene Selection in Cancer Classification With Expression Data". *IEEE Transaction on Nano bioscience* 4.3 (2005).

3. Chandra and Manish Gupta. "An efficient statistical feature selection approach for classification of gene expression data". *Journal of Biomedical Informatics* 44.4 (2011). 529-535.

4. RE Abdel Aal. "GMDH-based feature ranking and selection improved classification of medical data". *Journal of Biomedical Informatics* 38.6 (2005). 456-468.

5. MohmadBadr Al Snousy., *et al*. "Suite of decision tree-based classification algorithms on cancer gene expression data". *Egyptian Informatics of Journal*12.2 (2010). 73-82.

6. A.Sheik Abdullah., *et al*. "Data Classification: Its Techniques and Big data". Handbook of Research on Advanced Data Mining Techniques and Applications for Business Intelligence, IGI Global.

7. A.Sheik Abdullah., *et al*. "Comparing the Efficacy of Decision Tree and its Variants using Medical Data". *Indian Journal of Science and Technology* 10.18 (2017).

8. S Selvakumar., *et al*. "Decision Support System for Type II diabetes and its risk factor prediction using Bee based harmony search and Decision tree Algorithm". *International Journal of Biomedical Engineering and Technology, Interscience Publishers* 29.1 (2019).

9. Sheik Abdullah A., *et al*. "Heart Disease Prediction using Data Mining". Fifth International Conference on Biosignals, Images and Instrumentation (ICBSII 2019), SSN College of Engineering, Chennai, Springer Innovations in Communications and Computing Book Series (2019).

10. A Sheik Abdullah and S Selvakumar. "An Improved Medical Informatic Decision Model by Hybridizing Ant colony Optimization Algorithm with Decision Trees for Type II Diabetic Prediction". International Conference on Business Analytics and Intelligence, Indian Institute of Science (2018).

11. A Sheik Abdullah., *et al*. "Data Classification: Its Techniques and Big data". Handbook of Research on Advanced Data Mining Techniques and Applications for Business Intelligence, IGI Global, ISBN.

12. A.Sheik Abdullah., *et al*. "An Ensemble model for Predictive Analytics in Clinical data sets". Conference on Research Issues in Computing PSG College of Technology, Coimbatore (2015).

13. A Sheik Abdullah., *et al*. "An Efficient Prediction Model Using Multi Swarm Optimization Empowered By Data Classification For Type Ii Diabetes". Third International Conference on Business Analytics and Intelligence, Indian Institute of Management, (2015).

14. A Sheik Abdullah., *et al*. "Estimating the Predictive Performance Analysis of Medical Data Using Weight Based Decision Trees". Fourth International Conference on Business Analytics and Intelligence". IISC Bangalore (2016).

15. A Sheik Abdullah., *et al*. "Classification in medical data using type 2 fuzzy logic system with adaptive swallow swarm optimization". Fourth International Conference on Business Analytics and Intelligence". IISC Bangalore (2016).

16. A Sheik Abdullah., *et al*. "A Hybrid Preddictive model using Feed forward Neural Network with swarm intelligence technique in medical data". ICBAI IISC, Bangalore. (2016).

17. R Suganya., *et al*. "Classification of Cardiac disease using hybrid Support Vector Machine". Fourth International Conference on Business Analytics and Intelligence, IISC Bangalore. (2016).

18. R Suganya., *et al*. Application of Big Data in Intelligent Transportation system using modified Kruskal's Algorithm, Fourth International Conference on Business Analytics and Intelligence, IISC Bangalore. (2016).

19. A Sheik Abdullah., *et al*. "A Survey on Evolutionary Techniques for Feature Selection". ICEDSS during (2017).

20. A.Sheik Abdullah., *et al*. "Predictive Performance analysis of Type II Diabetic data based Support Vector Machine Classifiers upon varied kernel functions Springer - Lecture Notes in Computational Vision and Biomechanics [Scopus Indexed].

21. A Sheik Abdullah., *et al*. "An Introduction to Data Analytics: Its Types and Its Applications". Handbook of Research on Advanced Data Mining Techniques and Applications for Business Intelligence, IGI Global, ISBN.

22. A Sheik Abdullah., *et al*. "Descriptive Analytics". Applying Predictive analytics within the service sector, IGI Global.

23. A Sheik Abdullah., *et al*. Big data Analytics in Healthcare Sector for the book Machine Learning Techniques for Improved Business Analytics, IGI Global.

24. A Sheik Abdullah., *et al*. "An Introduction to Survival Analytics, Types, and Its Applications", Biomechanics, Intech Open Publishers, UK. (2019).

25. Marco D and Thomas S. "Ant Colony Optimization". Massachusetts Institute of Technology, (2004).

26. Wong KC. Computational biology and Bioinformatics: gene regulation CRC Press, Taylor and Francis group, ISBN. (2016).

27. Myers JL., *et al*. "Research Design and Statistical Analysis". second edition, Lawrence Erlbaum. (2003).

Volume 3 Issue 7 July 2019