Journal of Computer Science

Original Research Paper

A Novel Distinguishability Based Weighted Feature Selection Algorithms for Improved Classification of Gene Microarray Dataset

[1]Jeyachidra, J. and [2]M. Punithavalli

[1]Department of Computer Science and Applications, Periyar Maniammai University, Vallam-613 403. Thanjavur, Tamilnadu, India
[2]Department of MCA, Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India

Article history
Received: 03-01-2014
Revised: 06-02-2015
Accepted: 16-01-2015

Corresponding Author:
Jeyachidra, J.,
Department of Computer Science and Applications, Periyar Maniammai University, Vallam-613 403. Thanjavur. Tamilnadu, India
Email: j_chidra72@yahoo.com

Abstract: Data mining played vital role in comprehending, analyzing, understanding and interpreting microarray technology expression data. That includes search for genes that had similar or correlated patterns of expression. For that, the feature selection was one of the frequently used important techniques for data preprocessing. Many feature selection algorithms had been developed. Yet the persisting problem was in selecting optimal subset of features from the colon tumor dataset. The use of feature selection reduced the number of features, removed irrelevant, redundant or noise data thereby improving the accuracy, efficiency, applicability and understandability of the learning process. Dimensionality reduction and feature subset selection were important components of classification techniques. In this study, the authors presented a comparative study of existing six feature selection methods and the proposed two algorithms of their own.

Keywords: Feature Selection, Microarray Data, Classification, C4.5, Bayes

Introduction

The authors examined various feature selection methods for handling dataset with many features and found that many learning techniques were proved to be useful. The existing learning techniques worked well for most instances. However, when the number of samples or the number of features in the data was very large, the performance of the learning methods got degraded. The samples might become noisy and unclassifiable or the features might become irrelevant to the classifications. In order to overcome the problem, this study presented a " Novel Feature Selection Method Called A Novel Distinguishability Based Weighted Feature Selection Algorithms for Improved Classification of Gene Microarray Dataset". The researchers proposed A Novel Combinations of Two Level and Multilevel Dimensionality Reduction Methods which were based on the Feature Selection Column Wise K Neighborhood (DWFS-CKN) which was published in Jeyachidra and Punithavalli (2014) and Distinguishability Based Weighted Feature Selection Algorithms for classification of gene microarray dataset Jeyachidra and Punithavalli (2013). Features were selected from the reduced set based on quality measure of the individual features. In the proposed approach, to evaluate the performance of the feature selection methods, several experiments were conducted with the standard dataset and the results obtained were analyzed and found that they were better than the existing six feature selection methods taken up for comparison.

The feature selection was a process, in which no new set of features generated but only a subset of original features were selected with reduction in feature space. Dimensionality reduction was an active research area in the field of pattern recognition, machine learning, data mining and statistics. The purpose of dimensionality reduction was to improve the classification performance. It could be achieved in two ways namely feature selection and feature transformation.

Microarray technology had provided biologists with the ability to measure the expression levels of thousands of genes in a single experiment. The vast

amount of raw gene expression data led to statistical and analytical challenges including the classification of the dataset into right classes. The goal of classification was to identify the differentially expressed genes that might be used to predict class membership for new samples. The classification of gene expression data samples involved feature selection and classifier design. But the real problem in handling microarray data was its dimension. The increase of data size in terms of number of instances and number of features became a great challenge for the feature selection algorithms. Many of the features were irrelevant and redundant that increased the search space size resulting in difficulty to process the data further. This curse of dimensionality was a major problem in machine learning and data mining applications.

In this study, the authors addressed two simples, fast and efficient feature selection algorithms which would select more distinguishable features from gene expression profiles of microarray dataset. The performance of the proposed algorithms had been compared with the selected six popular feature selection techniques and the accuracy had been tested with two different classification algorithms Bayes and C4.5. The performance had been validated using k-fold validation and Leave-one-out cross validation by considering accuracy as metrics. The obtained results proved that the proposed algorithms performed better in accuracy as well as speed.

Problem Statement

Microarray technology provides a lot of data and it could be used for monitoring the gene expression levels of thousands of genes simultaneously. The feature selection plays a major role particularly in the microarray dataset. The feature selection is a task of crucial importance for the application of machine learning in various domains. Additionally, the recent increase of data dimensionality poses a severe challenge to many existing feature selection approaches with respect to efficiency and effectiveness. As an example, the proposed DWFS-AED is an effective algorithm which extracts most distinguishable features. It is hindered by the recent increase in data dimensionality. Therefore, adopting a new algorithm to cope with the high dimensionality of the data becomes a compelling necessity. This research work was to mainly focused on the selection of the feature subset using DWFS-AED. It was to improve and refine the machine learning and also to improve the optimization decisions. The objective of this research comprises (i) to study historical background using existing six methods by analyzing the performance and its impact (ii) to propose a new, simple, effective and efficient algorithm in order to improve the classification accuracy further.

Approach

Two new algorithms called DWFS-AED and DWFS_CKN were proposed in this study for improving the classification accuracy. The main goal in the analysis of colon tumor dataset was to identify the most important and relevant features of the gene that get expressed in a set of experimental conditions. It was simulated by MATLAB coding. The proposals were applied for feature selection in high dimensional gene expression colon tumor dataset. It was mainly based on the feature weights. To assess the effectiveness and efficiency of the algorithms, feature selection and accuracy were taken up for study. The impact of features were validated with k fold validation and the evaluated results were compared. Based on the accuracy a new method was proposed after comparing the results with that of the existing six methods taken up for study.

Objectives, Purpose and Motivation

The following paragraphs spelled out the objective and purpose for which this study was carried out along with the motivating factors.

Objectives

The objective of the research was an application of algorithms in data mining and bioinformatics for better understanding, diagnosis and treatment of colon tumor patients; that would mitigate their sufferings to a great extent; if not completely eliminate it.

Purpose

The purpose of this study was to develop and apply an appropriate algorithm with the data mining on colon tumor detection and cure; which was not hitherto attempted to.

Motivation

The suffering of the cancer patients who needed a better diagnosis and treatment for more comfortable life with less sufferings had been the motivating factor of this research. The earlier work also provided impetus and encouragement for developing new algorithms for better approach.

Scope of the Research

The research study was confined to the application of feature selection algorithms for improving the classification accuracy. The literature review revealed that the feature selection algorithms were widely based in numbers and applications. But the authors of study had taken only six existing feature selection algorithms based on its feature weightage and ranking.

The scope of study was restricted to the following:

- To provide an overview of existing feature selection techniques

- To eliminate or minimize the redundant and inappropriate data for improving the quality and accuracy
- To reduce the feature space and time by increasing the speed, for cost effectiveness
- To explore the possibilities of finding and then implementing an improved feature selection method
- To implement the proposed feature selection algorithm and validating their performance with suitable dataset
- The present study was concerned with the adoption of only six algorithms for analyzing the already available dataset. In this study, the feature selection problem was approached under two phases, because the extraction of information from massive data was difficult, time consuming and costly

The expected end results of the study were:

- Feature selection was important to select the relevant feature subset
- The scholar reviewed and analyzed the existing six feature selection methods and appreciated their strengths and weaknesses, before introducing a new algorithm called DWFS-AED and DWFS_CKN
- The two proposed methods were addressed based on the following backgrounds
    - To improve the quality of the data
    - To get better classification accuracy

So, in order to handle larger training dataset on high-dimensional dataset, the data mining required a simple, effective, efficient and yet a fast new algorithm to enhance classification accuracy.

The following research study had been carried out:

- Dataset selection. In that, feature subset selection was done using existing six methods
- Performance evaluation
- The impact of the existing six feature selection methods, on performance
- Proposal of two new feature selection algorithms
- A comprehensive and consolidated comparison of the tecniques under study for eight methods. (6 existing +2 proposed)

Step By Step Approach

The following steps were adopted in carrying out the research:

- An elaborate background study carried out on datamining and bioinformatics had been presented
- A complete literature survey on feature selection method had been incorporated
- The performance and characteristics of feature selection methods using gene microarray dataset had been evolved and the results were recorded Jeyachidra and Punithavalli (2012)
- The impact of the feature subset selection methods for classification of gene expression profiles of microarray dataset had been investigated and the results were compared Jeyachidra and Punithavalli (2013)
- Distinguishability Based Weighted Feature Selection using Attribute wise Euclidean Distance (DWFS-AED) for the classification of gene microarray dataset had been presented and its improvement in performance had been validated with suitable metrics and the comparative results were depicted pictorially Jeyachidra and Punithavalli (2013)
- Distinguishability Based Weighted Feature Selection using Column wise K Neighborhood (DWFS_CKN) for the classification of gene microarray dataset had been adopted and its validated improvement in performance had been incorporated Jeyachidra and Punithavalli (2013)
- A Novel Distinguishability Based Weighted Feature Selection Algorithms for Improved Classification of Gene Microarray Dataset were studied and its improvement in performance were compared with others
- A comprehensive and consolidated comparison of the methods under study had been carried out as results analysis in this paper

About the Research

As stated earlier, microarray experiments produced vast of thousands of attributes making it difficult for a classification algorithm to handle. Further, due to the lot of irrelevant/unimportant attributes present in the data, the accuracy of the classification algorithm also suffered considerably. To handle that, numerous algorithms had been experimented with for feature selection. The feature selection algorithms would find the most important features among the various features in the microarray data attempting to minimize the feature space and to maximize classification accuracy.

This research had addressed two newly proposed algorithm methods for feature selection in order to improve the classification accuracy.

Previous Works

Gene selection is of vital importance in classification of cancer using high dimensional gene expression data.

One of the major problem in applying gene expression profiles to tumor classfication. Many adoptable feature selection algorithms had been devised that could be found in (Guyon and Elisseeff, 2003; Molina *et al*., 2002). Some researchers (Dash and Liu, 2003; Djatna and Morimoto, 2008; Gheyas and Smith, 2010; Kohavi and John, 1997; Liu and Setionoo, 1998; Yang and Honavar, 1998; Yu and Liu, 2003) were involved in the study of goodness of a feature subset while determining an optimal one. The basic feature selection was indeed an optimization process.

Wang and Gotoh (2010) highlighted the differences in the behaviours of feature selection algorithms.In this paper they used eight different datasets particularly colon tumor was one of the gene analysis dataset by applying four different features selection algorithms and four different classifiers such as NB, DT, SVM and k.NN used and the results were shown clearly.

Ben-Dor *et al*., (2000) reported in their study, used colon tumor dataset and described how gene selection could be done. The minimum error was calculated and validated by LOOCV and used two different classifiers SVM and RLS. The same dataset was used by Renya *et al*. (2005) and used three different datasets including colon tumor dataset. The researchers improved the classification accuracy but the accuracy calculated time was not reduced.

Furey *et al*. (2000) reported in their work about support vector machine classification and validation of cancer tissue samples using microarray expression data. Leng and Muller (2006) explained the classification using functional data analysis for temporal gene expression data. Mohamad *et al*. (2007) developed a model for gene selection and classification of gene expression data.

Wang and Makedon (2004) described the application of relief feature filtering algorithms to select the informative genes for classification using microarray data. Xiong *et al*. (2000) used two methods called –principal componenet analysis and discriminat analysis methodsof tumor classification using expression profile data. Using this methods, the percentage of correctly classified normal and tumor tissues was 87 % in the colon tumor dataset.

Xing *et al*. (2001) explained how to apply data mining techniques for cancer classification from gene expression data.

Zhenyu and Palade (2011) developed an intrepretable fuzzy models for high dimensional data analysis in cancer diagnosis Previously, the same author Zhenyu in 2007- developed a framework for cancer microarray data gene expression analysis. That method had used three microarray cancer datasets namely Leukemia, Colon Cancer and Lymphoma Cancer. A novel fuzzy based system was used for both gene selection and classification by applying the microarray gene expression data. The performance achieved by that method was more practicable.

Osareh and Shadgar (2010) for cancer classification. An automated system was developed for consistent cancer analysis based on gene microarray expression data. The researchers used the microarray datasets which included both binary and multi-class cancer problems. Santanu *et al*. (2011) offered a Nonparallel Plane Proximal Classifier (NPPC) ensemble for cancer classification based on microarray gene expression data. A hybrid CAD method was introduced based on filters and wrapper approaches.

Dudoit *et al*. (2000) detailed that since tumors were normally rich in epithelial cells, while normal tissues contained variety of cells including large fraction of muscle cells. These different cell composition might lead to genes that had different expression value in normal and tumor tissues. Yeh *et al*. (2007) explained the nine classes of samples were identified-colon, breast and central nervous system.

By considering the above problems- speed, accuracy and optimization in order to improve the classifictaion accuracy and background study were made. Then, we decided to use only one dataset-colon tumor dataset. But the real survey by the author would not have been truthfully success. Since, most of the people would not reveal that they have cancer due to fear, shyness and social stigma associated with. Also, some of the previous researchers used the colon tumor dataset and highlighted the complexity of that dataset. Hence, it was decided to select and use the publicly available standard colon tumor dataset for this study. The source of the dataset is given below. http://www.molbio.princeton.edu/colondata

The Colon Tumor Microarray Data Set

This dataset contained 62 samples collected from colon tumor patients. Among them, 40 tumor biopsies were from tumors (labeled as “negative”) and 22 normal (labeled as “positive”) biopsies were from healthy parts of the colons of the same patients. Each sample was described by 2000 genes. So, the data set contains 62×2000 continuos variables and 2000 class ids (The negative was assigned as 1 and positives as 2 for handling with MATLAB code) Jeyachidra and Punithavalli (2012).

The Proposed Distinguishability Based Weighted Feature Selection Algorithms

The authors used two classification algorithms Bayes and C4.5 and validation had been done by LOOCV and k-Fold which were used for evaluating the performance of the feature selection algorithms. Also, the authors had selected similar dataset for the proposed two algorithms.

The Proposed Distinguishability based Weighted Feature Selection using Column wise K Neighborhood (DWFS_CKN)

In the proposed algorithm, feature weights were calculated based on the classifiable/distinguishable nature of the corresponding member points of that features using a column wise k-neighborhood method. In it, for a particular column of a feature, most of the points were definitely belonging to any one of the class and distinguishable from the other classes based on k-neighborhood of each value, then the feature weight of that particular column was high. So, a feature which had the highest feature weight of the most important attribute of the data and a feature which had lowest feature weight was the least important attribute of the data. So, the researchers had selected a small set of first few features which had high feature weights for classification tasks.

The following algorithm detailed the proposed Data Distinguishability based Weighted Feature Selection (DWFS_CKN) algorithm Jeyachidra and Punithavalli (2014).

Algorithm: DWFS_CKN

Let D be the set of Microarray Data of m rows of n features

T be the corresponding class id's of m records of D.

The dataset D could be grouped in to c number of sub groups based on the class membership as follows

$D = \{g_1, g_2, .. g_c,\}$

Where

$g_1, g_2, .. g_c$, were the c number of sub sets of data belonging to c classes.

$\overline{g_1}, \overline{g_1}, \ldots \overline{g_c}$ were the colum-wise average of $g_1, g_2, .. g_c$,

W - array of size of 1×n to hold the feature weights

Dist- array of size of 1×n to hold the minimum distance.

```
for i = 1 to n        //for every feature in the data do this
{
 for j = 1: m         //for every row in the data do this
 {
          //k-neighbor Detection
          for k = 1: m // again for every row in the data do this {
          // calculate the distance between
 // the selected attribute point
 // and other points
          d(k) = |D(j,i)^2 - D (k)^2 | ^1/2
          }
          // we would have the set of distances of size m×1
          d = { d1, d2,….. dm}
          //sort the distances in ascending order
          idx = sort(d)
          //Now we would find top k neighbors
          Neighbors = T(idx(1: kn))
          //find the index of neighbors which were in the same class T(j)
          Idx = find(Neighbors == T(j))
          /If there were at least k/2 neighbors belong to the class T(j)
          //then that data point was a classifiable one-increase weight
          If size(idx) >k/2 {
 W(i) = W(i)+1;
 }
 }
 }
 Features = sort(W,'descend' );
```

Now, the first n features could be used as the primary features.

The Proposed Distinguishability Based Weighted Feature Selection using Attribute-wise Euclidean Distance (DWFS-AED)

In the proposed algorithm, feature weights were calculated based on the classifiable/distinguishable nature of the corresponding member points of that features. In it, for a particular column of a feature, most of the points were definitely belonging to any one of the classes and distinguishable from the other classes, then the feature weight of that particular column was high. So, a feature which had highest feature weight was the most important attribute of the data and a feature which had lowest feature weight was the least important attribute of the data. So, the researchers had selected a small set of first few features which had high feature weights for classification tasks. The authors had also selected similar dataset for proposed two algorithms.

The following algorithm explained the proposed Data Distinguishability based Weighted Feature Selection using Attribute-wise Euclidean Distance (DWFS-AED) algorithm Jeyachidra and Punithavalli (2013).

Algorithm: DWFS-AED

Let

D be the set of Microarray Data of m rows of n features

T be the corresponding class id's of m records of D.

The dataset D could be grouped in to c number of sub groups based on the class membership as follows

$D = \{g_1, g_2, .. g_c,\}$

Where

$g_1, g_2, .. g_c$, were the c number of sub sets of data belonging to c classes.

$\overline{g_1}, \overline{g_1}, \ldots \overline{g_c}$ were the colum-wise average of $g_1, g_2, .. g_c$,

```
for i = 1 to n        //for every feature in the data do this
 {
 for j = 1:m          //for every row in the data do this
 {
          for k = 1: c          //for each class of data
 {
          // calculate the distance between
 // the selected attribute point
```

```
// and the class average of that attribute
            d(k) = |D(j,i)^2 - g (k)^2 | 1/2
            }
            // we would have the set of c distances
            d = {d1, d2,….. dc}
            //find the index of the smallest distance
            idx = Idx(min(d))
            if (T(j) == idx)
                        {
//if the calculated class of the attribute
// was equal to the original class ID
// then increase the weight of that feature.
                        W(i) = W(i)+1;
                        }
 }
 }
}
Features = sort(W,'descend');
```

Now, the first n features could be used as the primary features

Metrics Used for Performance Evaluation-Classifiers and Validation

Bayes Classifier and C4.5 Classifier

Bayes classifier and C4.5 algorithms were used to verify the accuracy Quinlan (1993). C4.5 was quite popular and efficient algorithm in Decision tree-based approach. The accuracy and error rate were calculated using the following foumulae. Validation had been done by using LOOCV and k-fold cross validation:

Accuracy = (TP+TN) / (TP + FP + TN + FN)

Error rate = (FP+FN) / (TP + FP + TN + FN)

Results and Discussion

About the Implementation

The authors used the feature selection tool box called ‘fspackage‘ provided by Arizona State University for doing the experiments. A MATLAB application based on this tool box for this evaluation was developed. The results of six other previous feature selection methods Gini Index, Chi Square, MRMR, T-Test, Relief-F and Information Gain along with the results of the two proposed algorithms and implemented the proposed DWFS_CKN and DWFS-AED algorithms under MATLAB, were compared.

The Table 1 shows the accuracy and error rate of classification by Bayes and J48 (C4.5) with respect to first 50 features selected by different feature selection algorithms. The metrics were calculated Jeyachidra and Punithavalli (2013) by doing Leave-One Out Cross Validation (LOOCV).

The Fig. 1 shows the accuracy of classification by Bayes and J48 (C4.5) with respect to first 50 features selected using LOOCV by different feature selection algorithms. As seen from the Fig. 1, the performance of the proposed DWFS_CKN and DWFS-AED were better than the existing six methods, pertaining to the parameters of accuracy and error.

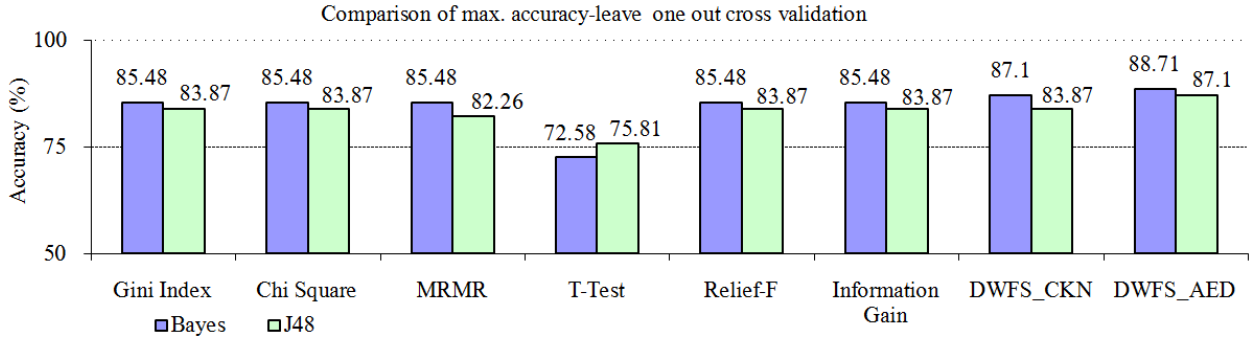

Fig. 1. The accuracy found through leave one out cross validation

Table 1. LOO cross validation using 50 features

	Bayes(%)		J48 (%)	
Feature Selection Methods	Accuracy	Error	Accuracy	Error
Gini Index	85.48	14.52	83.87	16.13
Chi Square	85.48	14.52	82.26	17.74
MRMR	85.48	14.52	82.26	17.74
T-test	72.58	27.42	75.81	24.19
Relief-F	85.48	14.52	83.87	16.13
Information gain	85.48	14.52	83.87	16.13
Proposed DWFS_CKN	87.10	12.90	83.87	16.13
Proposed DWFS-AED	88.71	11.29	87.10	12.90

The Table 2 shows the average accuracy, average error rate, minimum achieved error and maximum accuracy achieved with respect to 50 features by the algorithms. This was calculated by repeating the 10 fold cross validation for 25 times (each time, the data were kept in a random order) Jeyachidra and Punithavalli (2014).

The Fig. 2 shows the average accuracy of the 25 iterations of 10 fold cross validation. As shown in the Fig. 2, the average accuracy and error of the proposed DWFS were better than the existing six algorithms.

The Table 3 shows the time taken by the different algorithms with respect to first 10 index of the selected features Jeyachidra and Punithavalli (2012).

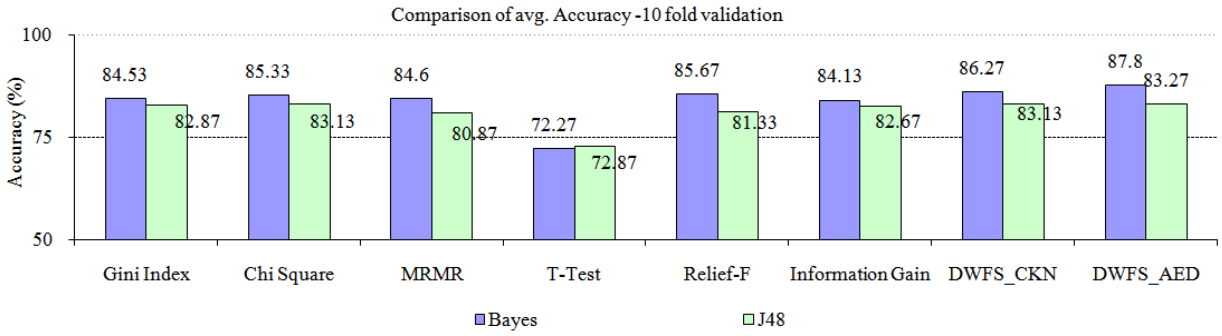

Fig. 2. The average accuracy found through the average of 25 runs of k fold cross validation (k = 10)

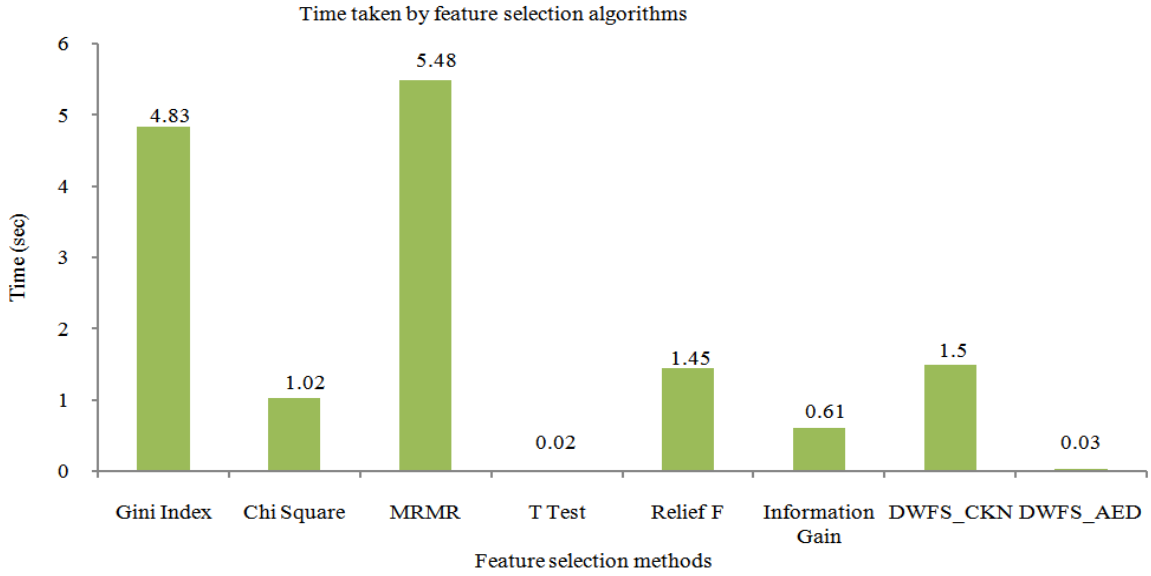

Fig. 3. The time taken for feature selection methods

Table 2. 10-fold cross validation with respect to first 50 features-the average, max and min of 25 iterations

	Bayes (%)				J48 (%)			
Feature selection methods	Avg. Acc.	Avg. Error	Max. Acc.	Min. Error	Avg. Acc.	Avg. Error	Max. Acc.	Min. Error
Gini index	84.53	15.47	86.67	13.33	82.87	17.13	86.67	13.33
Chi square	85.33	14.67	88.33	11.67	83.13	16.87	90.00	10.00
MRMR	84.60	15.40	86.67	13.33	80.87	19.13	88.33	11.67
T-Test	72.27	27.73	75.00	25.00	72.87	27.13	83.33	16.67
Relief-F	85.67	14.33	86.67	13.33	81.33	18.67	90.00	10.00
Information gain	84.13	15.87	86.67	13.33	82.67	17.33	88.33	11.67
Proposed DWFS_CKN	86.27	13.73	88.33	11.67	83.13	16.87	87.13	12.87
Proposed DWFS-AED	87.80	12.20	90.00	10.00	83.27	16.73	88.33	11.67

Table 3. The Top 10 Index of the primary features according to eight different methods

Feature selection methods	Time taken (sec)	Index of the first 10 selected features
Gini Index	4.83	1671, 249, 493, 765, 1423, 513, 1771, 245, 267, 1772
Chi Square	1.02	1671, 249, 493, 765, 1423, 513, 1771, 245, 267, 1772
MRMR	5.48	1671, 249, 493, 765, 1772, 625, 1042, 1423, 513, 1771
T-Test	0.02	1772, 1582, 513, 1771, 780, 138, 515, 625, 1325, 43
Relief-F	1.45	267, 245, 249, 1423, 822, 765, 1892, 66, 493, 897
Information Gain	0.61	1671, 249, 493, 765, 1772, 625, 1042, 1423, 513, 1771
Proposed DWFS_CKN	1.50	249, 1671, 1423, 513, 765, 245, 267, 493, 1892, 415
Proposed DWFS-AED	0.03	249, 765, 245, 267, 1423, 415, 1892, 66, 822, 897

The Fig. 3 shows performance of the feature selection algorithms in terms of run time. As shown the Fig. 3, the time taken of the DWFS-AED was better than the other compared algorithms except the one of T-Test. But T-test provided poor performance in terms of accuracy as seen in Fig. 1 and 2.

Conclusion

In this study the authors had proposed two simple, fast and efficient feature selection algorithms called DWFS_CKN and DWFS-AED and compared their performance with already existing six classical feature selection methods using a complex microarray dataset.

From the results of the experiments, the following were concluded in terms of.

Speed

Time taken by the different feature selection algorithms were given in the Table 3. The only exception was T-Test, in which the time taken by T-Test was lower (0.02) compared to the proposed DWFS-AED (0.03). But the remaining algorithms as per Table 3, the time taken was higher than the DWFS-AED.

As far as the run time is concerned, the proposed DWFS-AED algorithm consumed only negligible time compared to all other feature selection methods but performed better accuracy compared to other algorithms.

Accuracy

From Table 1 showed it could be concluded that the accuracy was improved to 88.71 with respect to 50 features by Bayes classifier whereas 87.10 by J48 classifier compared to the seven feature selection algorithms by adopting the proposed DWFS-AED.

Also, the average accuracy (87.80%) and maximum accuracy (90%) by Bayes classifier for 25 iterations with respect to 50 features when compared to the seven algorithms as per Table 2 whereas the average accuracy (83.27%) and the maximum accuracy (88.33%) by J48 classifier for 25 iterations with respect to 50 features when compared to the remaining seven algorithms as per the Table 2.

The Fig. 2 and 3 shows the comparative analysis of the accuracy with respect to 50 features by Bayes and J48 classifiers.

Hence, it could safely be concluded that the proposed DWFS-AED algorithm-the proposed one-was better that the seven feature selection algorithms in terms of accuracy and speed.

Limitations of this Research

Feature selection had been an active field of research for decades in data mining and had been widely applied to many fields such as genomic analysis, text mining and image retrieval etc. Data mining involved the use of data analysis tools to discover previously not -known, valid patterns and relationships in large datasets. Yet, there were some limitations to its capability.

The first limitation of data mining was that it could help to reveal patterns and relationships. It needed the support of the following - A well trained user who could supply the perfect data. Incorrect and insufficient information, could lead to inaccurate and often wrong diagnosis leading to inappropriate treatment.

Yet, data mining was truly an innovative process that if it was used in the proper way, could yield amazing results in spite of its shortcoming and limitations. It could be well differentiated, calculated the accuracy and validated k fold and LOOCV knowledge-discovery. The validation of the proposed feature selection algorithm was based on the actual implementation of the dataset.

Initially, the author had taken existing six feature selection methods and compared their accuracies, with the application of the same microarray dataset. The feature selection methods used to select only first 10 features and extracted features were shown in the Table 1. The accuracy of the features varied depending on the selection algorithms and when applied to one dataset, it might or might not result with the same accuracy. Similarly, when applied to any other dataset, its ranking might vary from one dataset to another dataset depending on the features selected.

Computation with existing methods made it difficult to directly handle the higher dimensional dataset, since computational power, accuracy and speed were being challenged. To address those challenges, future researchers need to develop efficient and effective algorithm to handle the higher dimensional dataset for improving the classification accuracy.

Future Work

Data mining would continue to be a valuable tool both in hospital and pharmacy sector to protect and cure patient diseases.

This development is an unending with never ending learning. The research is not finite with more and more unknown are made known by exploring more and more unexplored areas by continuous search. Hence, the scope of this present study was restricted to the following:

- The new algorithms could be expanded with other large microarray datasets to improve the classification accuracy
- Future work could deal with presenting better results in an accessible way and assessing the feasibility of methods to increase classification performance
- Based on the observation, the behavior and characteristics of the feature selection algorithms, to find more and more distinguishable features, the performance of the proposed DWFS-AED algorithm, further, it could be enhanced by using suitable distance calculation technique

Funding Information

The authors have no support or funding to report.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article contains original research work.The data and text extracted and presented/inserted in this paper are acknowledged with citation reference at appropriate places. Though it is own work, it has been properly recorded in the reference section for the comparative study. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Ben-Dor, A., L. Bruhn, N. Friedman, I. Nachman and M. Schummer *et al*., 2000. Tissue classification with gene expression profiles. J. Comput. Biol.,7: 559-583. PMID: 11108479

Dash, M. and H. Liu, 2003. Consistency-based search in feature selection. Artificial Intell., 151: 155-176. DOI: 10.1016/S0004-3702(03) 00079-1

Djatna, T. and Y. Morimoto, 2008. A novel feature selection algorithm for strongly correlated attributes using two-di mensional discriminant rules. Hiroshima University.

Dudoit, S., J. Fridlyand and T.P. Speed, 2000. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc., 97: 77-87. DOI: 10.1198/016214502753479248

Furey, T., N. Cristianini, N. Duffy and D. Bednarski, 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16: 906-914. DOI: 10.1093/bioinformatics/16.10.906

Guyon, I. and A. Elisseeff, 2003. An introduction to variable and feature selection. J. Mach. Learn. Res., 3: 1157-1182.

Gheyas, I.A. and L.S Smith, 2010. Feature subset selection in large dimensionality domains. Patt. Recognit., 43: 5-13. DOI: 10.1016/j.patcoq.2009.06.009

Jeyachidra, J. and M. Punithavalli, 2012. A comparative study of feature selection methods for cancer classification using gene expression dataset. Eur. J. Scientific Res., 93: 214-225.

Jeyachidra, J. and M. Punithavalli, 2013a. An investigation into the impact of the feature subset selection methods for classification of gene expression profiles of microarray dataset. Int. J. Scientif. Eng. Res., 4: 2395-2402.

Jeyachidra, J. and M. Punithavalli, 2014. Distinguishability based weighted feature selection using columnwise K neighborhood for the classification of gene microarray dataset. Am. J. Applied Sci., 11: 1-7. DOI: 10.3844/ajassp.2014.1.7

Jeyachidra, J. and M. Punithavalli, 2013b. Distinguishability based weighted feature selection algorithms for classification of gene microarray dataset. Int. J. Eng. Sci. Innovat. Technol., 2: 173-181.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. 1st Edn., Morgan Kaufmann, San Mateo, ISBN-10: 1558602380, pp: 302.

Kohavi, R. and G.H. John, 1997. Wrappers for feature subset selection. Artificial Intell., 97: 273-324. DOI: 10.1016/s0004-3702(97)00043-x

Leng, X. and H. Muller, 2006. Classification using functional data analysis for temporal gene expression data. Bioinformatics, 22: 68-76. DOI: 10.1093/bioinformatics/bti742

Liu, H. and R. Setionoo, 1998. Incremental feature selection. Applied Intell., 9: 217-230. DOI: 10.1023/A:1008363719778

Molina, L.C., L. Belanche and A. Nebot, 2002. Feature selection algorithms: A survey and experimental evaluation. Proceedings of IEEE International Conference on Data Mining, Dec. 9-12, IEEE Xplore Press, pp: 306-313. DOI: 10.1109/ICDM.2002.1183917

Osareh, A. and B. Shadgar, 2010. Microarray data analysis for cancer classification. Proceedings of the 5th International Symposium on Health Informatics and Bioinformatics, Apr. 20-22, IEEE Xplore Press, Antalya, pp: 125-132. DOI: 10.1109/HIBIT.2010.5478893

Renya, H., C. Qiansheng, W. Lianwen and Y. Kehong, 2005. New feature extraction in gene expression data for tumor classification. Prog. Nat. Sci., 15: 861-864. DOI: 10.1080/10020070512331343040

Mohamad, M.S. S. Omatu, S. Deris and S.Z.M. Hashim, 2007. A model for gene selection and classification of gene expression data. Artificial Life Robot., 2: 219-222. DOI: 10.1007/s10015-007-0432-1

Santanu, G., A. Mukherjee, S. Sengupta and P.K. Dutta, 2011. Cancer classification from gene expression data by NPPC ensemble. IEEE/ACM Trans. Comput. Biol. Bioinform., 8: 659-671. DOI: 10.1109/TCBB.2010.36

Wang, Y. and F. Makedon, 2004. Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data. Proceedings of the IEEE Conference on Computational Systems Bioinformatics, Aug. 16-19, IEEE Xplore Press, pp: 497-498. DOI: 10.1109/CSB.2004.1332474

Xing, E.P., M.I. Jordan and RM. Karp, 2001. Feature selection for high-dimensional genomic microarray data. Proceedings of the 18th International Conference on Machine Learning, (CML' 01), Morgan Kaufmann Publishers Inc. San Francisco, pp: 601-608.

Xiong, M., L. Jin and W. Li, 2000. Computational methods for gene expression-based tumor classification. Bio Techniques, 29: 1264-1270. PMID: 11126130

Wang, X. and O. Gotoh, 2010. A robust gene selection method for microarray-based cancer classification. Cancer Informat., 9: 15-30.

Yang, J. and V. Honavar, 1998. Feature subset selection using a genetic algorithm. IEEE Intell. Syst. Applic., 13: 44-49. DOI: 10.1109/5254.671091

Yeh, J.Y., T.S. Wu, M.C. Wu and D.M. Chang, 2007. Applying data mining techniques for cancer classification from gene expression data. Proceedings of the International Conference on Convergence Information Technology, Nov. 21-23, IEEE Xplore Press, Gyeongju, pp: 703-708. DOI: 10.1109/ICCIT.2007.153

Yu, L. and H. Liu, 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. Proceedings of the 12th International Conference on Machine Learning, (CML' 03), San Francisco, CA, Morgan Kaufmann, Washington, D.C., pp: 856-863.

Zhenyu, W. and V. Palade, 2011. Building interpretable fuzzy models for high dimensional data analysis in cancer diagnosis. BMC Genom., 12: S5-S5. DOI: 10.1186/1471-2164-12-S2-S5