The new baseline for high dimensional dataset by ranked mutual information features

Fung Yuen Chin[1*], and *Yong Kheng* Goh[2]

[1]Department of Physical and Mathematical Science, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia
[2]Department of Mathematical and Actuarial Sciences, Universiti Tunku Abdul Rahman, Kajang, Selangor, Malaysia

Abstract. Feature selection is a process of selecting a group of relevant features by removing unnecessary features for use in constructing the predictive model. However, high dimensional data increases the difficulty of feature selection due to the curse of dimensionality. From the past research, the performance of the predictive model is always compared with the existing results. When attempting to model a new dataset, the current practice is to benchmark for the dataset obtained by including all the features, including redundant features and noise. Here we propose a new optimal baseline for the dataset by mean of ranked features using a mutual information score. The quality of a dataset depends on the information contained in the dataset, and the more information contains in the dataset, the better the performance of the predictive model. The number of features to achieve this new optimal baseline will be obtained at the same time, and serve as the guideline on the number of features needed in a feature selection method. We will also show some experimental results that the proposed method provides a better baseline with fewer features compared to the existing benchmark using all the features.

1 Introduction

High dimensional datasets commonly come with noise, irrelevant features and redundant features, which will decrease the usability in classification tasks. Microarray data is an example of high dimensional data with small sample size and disproportion of high dimensionality and small sample size poses a great challenge to the research community on the analysis of microarray data.

A variety of statistical methods and machine learning tools used in the classification task in microarray data analysis can be classified into supervised learning and unsupervised learning. For more detailed reviews on supervised and unsupervised learning, please see [1]. Since "noise" always exists in a dataset, and the assumption of the normal distribution for a microarray dataset may not be suitable as the low number of sample, statistical approaches such as the regression model may not be suitable as normality is assumed. On the other hand,

* Corresponding author: chinfy@utar.edu.my

a non-parametric method is preferred as it does not require the data to satisfy a pre-assumed distribution.

One of the controversial issues in features selection for microarray analysis is the number of genes needed to interpret the predictive model. As the smallest number of genes is preferred, when going through the clinical tests or validation tests in the lab [2]. Current approaches are usually do not consider the number of features involved in building a predictive model. Most only show that the more relevant features are added to the built predictive model, the better the accuracy of the prediction. Therefore, when the number of features is added into the predictive model increased, this does not guarantee the newly added feature will provide new information to the predictive model. From the previous study, the accuracy of the predictive model was frustrated when more features were added to the predictive model, causing this to overfit. If the newly added feature cannot provide new information to the predictive model, then the meaning of adding this new feature is lost.

The objectives of this research are: (1) to develop a robust algorithm to handle the high dimension with low samples dataset in a classification task; and (2) use ranked features to obtain optimal baseline with minimal features, instead of using all the features. The number of ranked features, *k* needed to obtain the optimal baseline plays an essential role in feature selection where any features selection method should not use more than *k* features to achieve this optimal baseline. The *k* features serve as the unique cutoff number of features to obtain the optimal baseline. This is a significant breakthrough because the past research does not indicate the number of features needed to achieve the baseline and most likely the baseline is established using all the features.

2 Related work

For over the past 20 years, information theory was widely applied in various feature selection algorithm. Lewis has proposed a feature selection algorithm using Mutual Information Maximization (MIM) on text categorisation. In his paper, the features were ranked according to the expected mutual information and selection of the different size of features had been investigated. Besides, Lewis found that the optimal feature set size in his dataset should be in between 10 to 15 features, which is significantly low as compared with other feature selection methods during that time [3]. Subsequently, Battiti (1994) proposed another algorithm utilising mutual information and a greedy selection named Mutual Information based Feature Selection (MIFS) [4]. MIFS showed that mutual information is capable of measuring the dependency between variables, including both linear and non-linear dependencies, in contrast to many feature selection methods are only considering linear dependency between variables.

MIFS-U and MIFS-ND are two variants of the MIFS. MIFS-U [5] overcomes the issue that MIFS does not perform well in nonlinear cases dependent variables. MIFS-ND [6] selects features based on an optimisation algorithm named as Non-dominated Sorting Genetic Algorithm II (NSGA-II). NSGA-II will select features based on comparison not only feature-feature mutual information but also feature-class mutual information. In the event of a tie during feature selection, feature with higher feature-class mutual information will be selected. Peng et al. (2005) proposed a feature selection method based on the criteria of maximising the relevance between feature and class and, at the same time, minimising the redundancy between the feature known as “Minimal-Redundancy-Maximal-Relevance” (mRMR) [7]. The difference between mRMR and MISF-ND is mRMR uses a two-stage feature selection that involved filter method and wrapper method, while MISF-ND only uses the filter method in feature selection.

Normalised Mutual Information Features Selection (NMIFS) is an enhanced edition of MIFS, MIFS-U, and mRMR [8]. NMIFS does not depend on any parameters like MIFS,

MIFS-U, and mRMR, and in practice, there is no clear guidance on how to select the value of the parameter. Joint Mutual Information (JMI) considers the joint mutual information between the features and the selected features with the class [9]. Bennasar, Hicks, and Setchi (2015) proposed the Joint Mutual Information Maximisation (JMIM) and Normalised Joint Mutual Information (NJMIM) [10]. Max-Relevance and Max-Independence (MRI) consider the relevancy between the features and class and features classification independence among the features [11].

3 Methodology

3.1 Mutual information

Mutual information measures the amount of information contributed by the presence of one event to the occurrence of another event. The mutual information of two discrete random variables X and Y is defined as $I(X;Y) = \sum_y \sum_x p(x,y) log \left[\frac{p(x,y)}{p(x)p(y)}\right]$. Intuitively, mutual information measured the shared information between random variables X and Y. Besides, mutual information is non-negative, $I(X;Y) \geq 0$ and symmetrical, $I(X;Y) = I(Y;X)$. The amount of information obtained from another event was at most as much entropy as another event, and it will not exceed the amount of information contained in itself, i.e. $I(X;Y) \leq H(X)$ and $I(Y;X) \leq H(Y)$. When the random variables X and Y correspond one to one, this implies that $I(X;Y) = H(X)$ and $H(X|Y) = 0$. When the random variables X and Y were stand-alone of each other, $H(X|Y) = H(X)$ implies that $I(Y;X) = 0$.

Mutual information measures the amount of information contained between features and the class label. If the feature correlates with the class label, then the value of mutual information will be high. The mutual information does not require any assumptions about the nature of the relationship between the features and the class label, and it is well suited for the feature selection in bioinformatics. The mutual information calculation is similar to information gain, and the average of mutual information is also the information gain. The mutual information also can be written as:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \quad (2)$$

$$= \sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \quad (3)$$

$$= \sum_{x,y} p(x,y) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x|y)\right) \quad (4)$$

$$= H(X) - H(X|Y) \quad (5)$$

By symmetry, $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. Since $H(X,Y) = H(Y|X) + H(X)$, therefore $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

3.2 The optimal baseline based on ranked features

In this section, we will present our method of determining an optimal baseline of the number of selected feature by using the ranked features based on mutual information. The past research has proven that the mutual information can be used as a measure for the relevancy between two elements regardless of their distributions, whether linear or nonlinear. When the mutual information is of most significant, it means these two elements are closest to each other. The optimal baseline can be obtained using the measurement of mutual information of all the features to the class label. When all the features are ranked according to the mutual

information, the most significant feature is the most relevance to the class label, and this feature is sufficient to represent the class label. When more significant features are selected, the compact subset of features will well present the class label. By measuring the performance of the ranked features in a classifier, the more ranked features added into the classifier, the better is the performance. From this point of view, there will be a point or a few points on the number of ranked features which will indicate the highest performance in the classification.

Due to the high dimensionality, microarray data will always contain irrelevant features and noise. It is essential to find a better baseline involving as many as relevant features rather than using all the features. Since the relevant features ranked according to the mutual information, likewise the ranked features are ranked according to their relevancy to the class label. Using these relevant features hope this will outperform methods that using all the features in classification, and hence an optimal baseline can be obtained using those ranked features. At the same time, the number of features, *k* that is needed to achieve this optimal baseline, can also be known. The number of features, *k* plays a vital role in a classification problem, as current research does not indicate how many features are needed to achieve a baseline or the baseline is obtained by using all the features. Earlier research on feature selection have only focused on getting some selected features but there is no standard guideline on the number of selected features. With an emphasis on the same number of features, *k*, one could now make a direct comparison of the performance of other features selection method on equal footage.

3.3 Algorithm on ranked features

Mutual information can be used as a metric to measure the similarity between the features and class. When the mutual information score is more significant, it means that the feature is closer to the class label. The mutual information score for all the features between the class label can be computed using equation (1). The microarray dataset was a $N \times M$ matrix where *N* represents the number of attributes, and *M* represents the number of samples. For a feature set $X = \{x_1, x_2, \dots, x_N\}$ of a dataset D with N dimension and M sample, the mutual information for each x_i between the class label will be computed.

The real experimental data should be normalised into [–1, 1] before calculating the mutual information score. The advantage of remapping the feature into three equal bins is that no data pre-processing is needed when the data has a missing value, as the calculation of mutual information only depends on the remapped frequency count and not based on the original value. Again, mutual information did not consider the nature of the relationship between the features and the class label; therefore, mutual information can be applied to the data either linear or nonlinear. Each feature will be divided into three equal bins, and these three bins represent the expression of the microarray data in low, normal and high categories. The class label will be divided into p equal bins, where p is the number of classes of the class label.

The features with three equal bins and the class label with p equal bins will then be remapped into a frequency count and form a probability mass function (pmf) and the joint probability mass function (pdf) of each x_i and the class label. The mutual information score for each feature and class label will be computed using the equation (1). Then, the features will be ranked according to the mutual information score and the graph of accuracy versus the cumulative of ranked features will be plotted. The highest accuracy from the graph will represent the optimal baseline, and the cutoff number of the feature can be obtained at the same time. Fig. 1 shows the flowchart of finding the optimal baseline and the unique cutoff number of features.

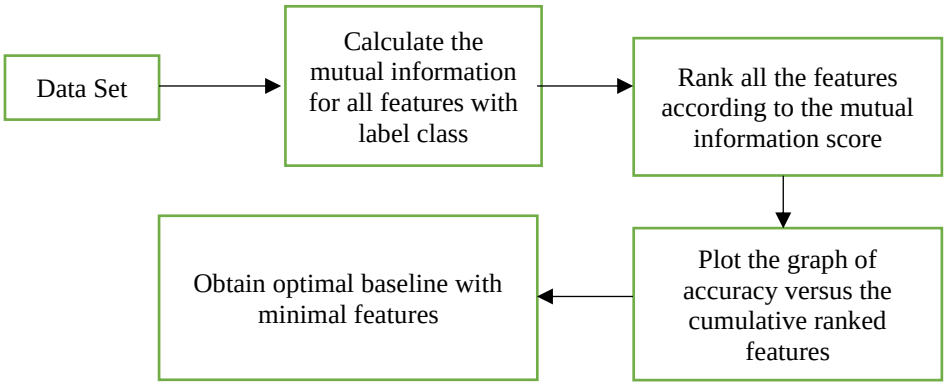

Fig 1. Flowchart of finding the optimal baseline and the unique cutoff number of features.

The performance of the ranked features measure using the support vector machine classifier and the accuracy of the research was defined as follow:

$$accuracy = 1 - \frac{falsenegative + falsepositive}{falsenegative + falsepositive + truepositive + truenegative}$$

or

$$accuracy = \frac{truepositive + truenegative}{falsenegative + falsepositive + truepositive + truenegative}$$

The next section shows the algorithm on calculating the mutual information score for each feature in a microarray dataset.

The algorithm on calculating mutual information score:

Input:
A training sample D with a full feature set $X = \{x_1, x_2, \dots, x_N\}$ and the class label C with N dimension.

Output:
1. (Initialisation) Set $X \leftarrow$ "initial set of N features"
2. (Normalising) For $\forall x_i \in X$, normalise each x_i in [– 1 , 1]
3. (Remapping) For $\forall x_i \in X$, remapped each x_i into three equal bins and for remapped the class label C into p equal bins
4. (Forming) For $\forall x_i \in X$, formed a frequency count and a joint pdf the each x_i and class label C
5. (Calculating) For $\forall x_i \in X$, calculate the mutual information for each pair of x_i and class label C using the joint pdf
6. (Ranking) For $\forall x_i \in X$, ranked the mutual information score for each x_i in descending order
7. (Plotting) Plotted a graph of accuracy versus cumulative ranked features
8. (Identifying) Identify the highest accuracy and the number of features

4 Result and discussion

4.1 Dataset

The microarray data used in this research were public microarray data downloaded from the National Center for Biotechnology Information (NCBI) and UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/index.php. The Gene Expression Omnibus (GEO) at NCBI was the largest fully public repository for molecular data and gene expression data. The GEO database was publicly accessible via the www at http://www.ncbi.nlm.nih.gov/geo.

These microarray data have also been used by other authors and publishers in many journal publications. The summary of the four data had been shown in table 1. There were two binary classification datasets and two multiclass classification datasets. The reason for choosing these four datasets was that these microarray datasets are high dimensional datasets with low sample sizes that are widely used in published journals. Thus, the proposed algorithm can be applied here to obtain the optimal baseline.

Table 1. Summary of the downloaded dataset.

Dataset	No. of attribute	No. of sample	Type of classification
Colon cancer	1988	62	Binary
Leukaemia	7128	72	Binary
Skin cancer	22215	15	3 classes
Lymphoma	4026	96	9 classes

A full dataset will first be divided into two random subsets: a training set and test set according to the ratio of 7:3. The training sample *D* with a full feature set $X = \{x_1, x_2, \dots, x_N\}$ and the class *C* with *N* dimension as an input in the algorithm. The training set then will be normalised into a range of [-1, 1], then for each feature will be divided into three equal bins and the class label will be classified according to the number of classes. The joint probability mass functions (pdf) of a feature and the class label will then be calculated. The optimal baseline will be computed for all datasets using the ranked features, as described in section 3. This process will be repeated for 50 times to obtain the average mutual information score for each feature. Each time, a new training set and the test set will be obtained.

4.2 Optimal baseline

The optimal baseline for the data will be obtained by using the proposed algorithm provided in section 3. Fig. 2-5 show the average accuracy of the ranked features for colon cancer dataset, the leukaemia dataset, skin cancer dataset, and lymphoma dataset.

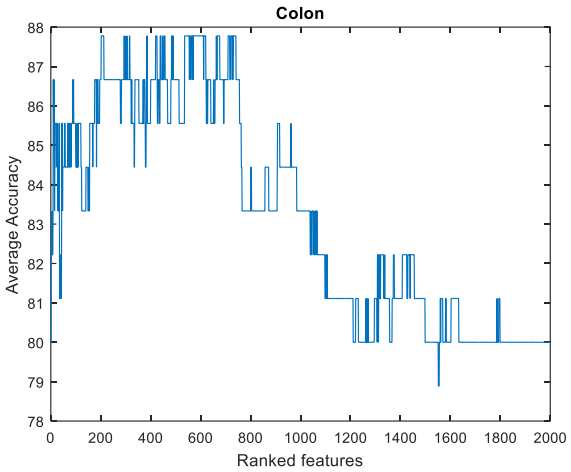

Fig. 2. The average accuracy of the ranked features for colon cancer dataset.

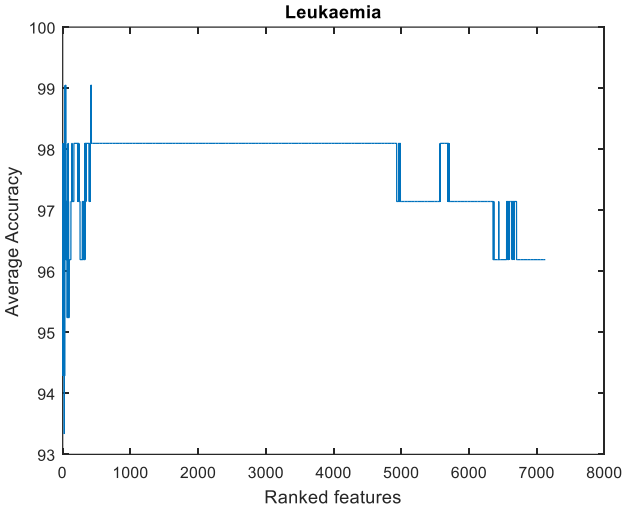

Fig. 3. The average accuracy of the ranked features for the leukaemia dataset.

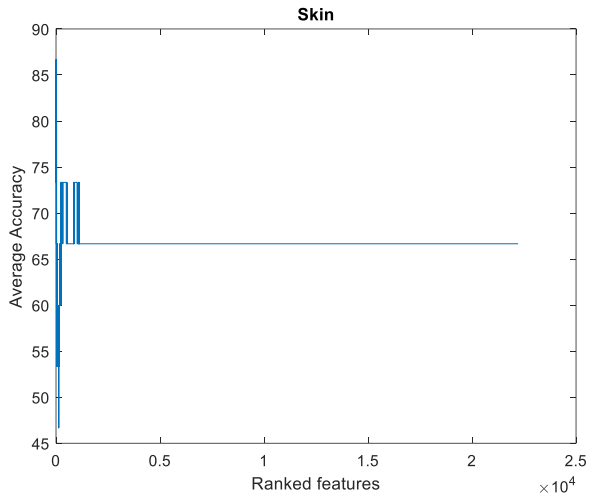

Fig. 4. The average accuracy of the ranked features for skin cancer dataset.

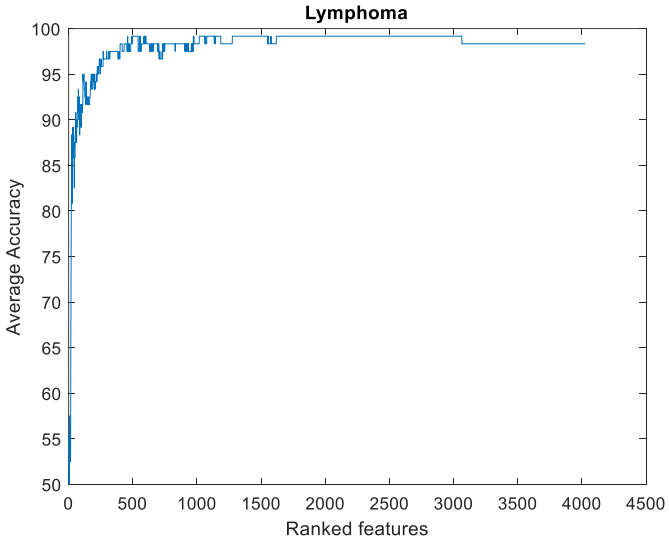

Fig. 5. The average accuracy of the ranked features for the lymphoma dataset.

Based on the above figures, we use the highest average accuracy as the optimal baseline of the data set and the number of features to obtain this optimal baseline was also achieved at the same time. From the figures, it was noticed that when using all the features as a baseline, it usually does not produce an optimal one, as these baselines are lower than the proposed optimal baseline. When all the features are included, meaning that at the same time, the redundancy and noise are included as well. In the case of using all the features, the accuracy of the predictive model shows inconsistent up and down movements, due to the addition of redundant features into the model. Besides, the accuracy of the predictive model is decreased. From the figures, the optimal baseline and the number of features to achieve this optimal baseline can be obtained.

Table 2 shows the baseline using full features and the optimal baseline obtained using the proposed algorithm in section 3 with the number of features to achieve this optimal baseline. The proposed algorithm shows that the proposed model yield optimal baselines are better

than these of using full features. Also, the number of features required to obtain the optimal baseline is significantly lower than the number of all the features. The features obtained using the proposed algorithm are ranked features where information contained in the features are the most relevant. Therefore, this number of features obtained from the proposed algorithm should provide a more robust prediction power in the predictive model as the number of dependent variables is less.

Table 2. The baseline using full features and the optimal baseline with the number of features.

Dataset	Baseline	Full features	Optimal baseline	No. of features
Colon	80%	1988	87.78%	202
Leukaemia	96.19%	7128	99.05%	38
Skin	66.67%	22215	86.67%	2
Lymphoma	98.33%	4026	99.17%	460

The number of features obtained using the proposed algorithm plays a vital role in features selection, and the features selection method should not take more than this number of features to achieve the same accuracy. Therefore, we established a new guideline on the cut off value of the maximum number of features that are allowed in a feature selection. For the past research, they are using the all the features to obtain the baseline, therefore no clear guideline on how many features should be selected in a predictive model and this is the advantage of the currently proposed method.

5 Conclusion

The mutual information has been used in this study. There are advantages in using mutual information in feature selection, as mutual information can evaluate both linear and nonlinear dependencies among the features and class label. Besides this, the mutual information score can be easily calculated regardless of the linearity of the features and even the data have some missing values. Besides this, no normal assumption was applied here, unlike some statistical method, where the data set must fulfil the normal assumption.

The features were ranked according to the mutual information, and the higher the mutual information score, the more information are contained in that feature. The result shows that the optimal baseline is better than the current benchmark that involved all the features as this will also include the irrelevant, redundant features and noise. At the same time, the number of features to achieve this optimal baseline can be obtained. The number of features will serve as a guideline on how many features are needed in a predictive model.

This study only compares the baseline using all the features as the current research is focused on feature selection methods to compare the performance of the feature selection methods with existing features selection method or to compare with the baseline using all the features. This study is focused on developing an algorithm to determine the optimal baseline for a data set before feature selection takes place in building the predictive model.

References

1. N.X. Vinh, J. Bailey, Pattern Recognit. **46**, 4 (2013)
2. V. Elyasigomari, D.A. Lee, R.C, Screen, M.H. Shaheed, J. Biomed. Inform. **67** (2017)
3. D.D. Lewis, *Feature selection and feature extraction for text categorisation*, in Proceedings of the workshop on Speech and Natural language (1992)
4. R. Battiti, IEEE Trans. Neural Netw. **5** (1994)
5. N. Kwak, C. Choi, IEEE Trans. Neural Netw. **13** (2002)
6. N. Hoque, D.K. Bhattacharyya, J.K. Kalita, Expert Syst. Appl. **41**, 14 (2014)
7. H. Peng, F. Long, C. Ding, IEEE Trans. Pattern. Anal. Mac. Intell. **27** (2005)
8. P.A. Estévez, M. Tesmer, A. Perez, J.M. Zurada, IEEE Trans Neural Netw. **20** (2009)
9. H. Yang, J. Moody, *Feature selection based on joint mutual information*, in Proceedings of international ICSC symposium on advances in intelligent data analysis (1999)
10. M. Bennasar, Y. Hicks, R. Setchi, Expert Syst. Appl. **42** (2015)
11. J. Wang, J. Wei, Z. Yang, S. Wang, IEEE Trans Knowl. Data Eng. **29** (2017)