IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 53-58

Feature Selection Methods on Genomic Data

Marrynal S. Eastaff[1], Dr. V. Saravanan[2]
[1]*Ph.D Research Scholar, Department of Computer Science, Hindusthan College of Arts and Science*
[2]*Head Department of Information Technology, Hindusthan College of Arts and Science*

Abstract: *The are various ways of performing dimensionality reduction on high-dimensional microarray data. Many different feature selection and feature extraction methods exist and they are being widely used. All these methods aim to remove redundant and irrelevant features so that classification of new instances will be more accurate. A popular source of data is microarrays, a biological platform for gathering gene expressions. analysing microarrays can be difficult due to the size of the data they provide. This paper presents some of the most popular methods for selecting significant features.*
Keywords: - *Big Data, Feature selection, Filters, Wrappers.*

I. Big Data

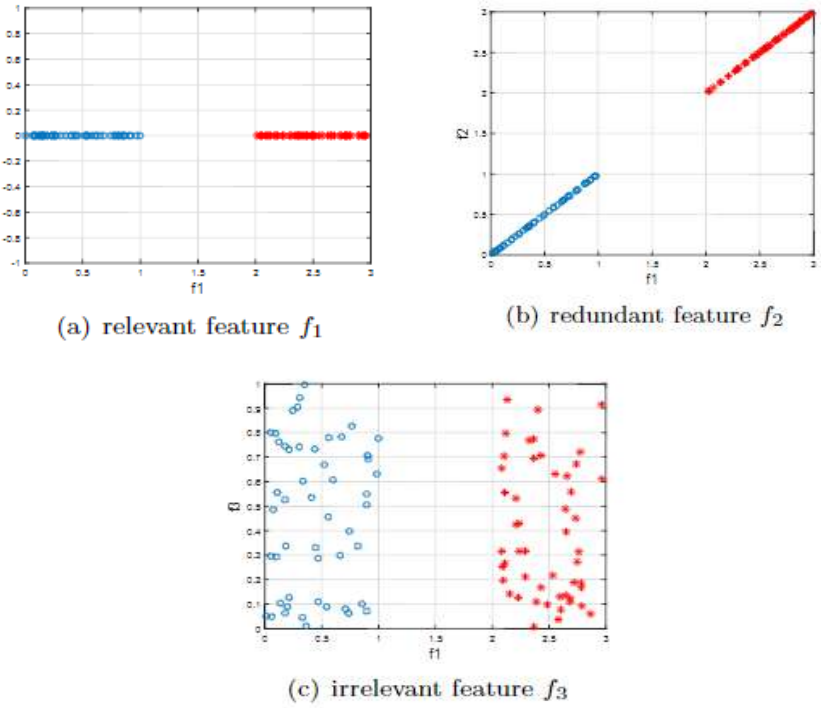
We have entered an era of big data. Data are becoming bigger not only in terms of the abundance of patterns (data instances or tuples), but also the dimensionality of features (or data attributes). This can significantly degrade the accuracy and efficiency of most learning algorithms, especially when there exist irrelevant or redundant features. Sometimes, the sheer size of the data even reNders the data mining algorithms completely useless. The situation is particularly acute in bioinformatics [1,2,3,4,5,6,7]

Laney characterized big data with 3Vs, i.e., volume (enormous size of data sets), variety (many sources and types of data) and velocity (fast pace at which data flows in from sources). Later, Normandeau added the 4th V, i.e. veracity, to describe biases, noise and abnormality in data. A large variety of bioinformatics data includes genomics, proteomics, biomedical imaging, clinical trial data, etc. Langley et al. pointed out that the predictive accuracy of the learning algorithms are reduced in the presence of irrelevant features. Koller et al. proved that the distribution of truly relevant features for the main task are blurred by irrelevant or redundant features [15].

II. Feature Selection

Feature selection "Feature selection is the process of selecting the relevant features and discarding the irrelevant and redundant ones". We are surrounded by huge amounts of large-scale high dimensional data. It is desirable to reduce the dimensionality of data for many learning tasks due to the curse of dimensionality. Feature selection has shown its effectiveness in many applications by building simpler and more comprehensive model, improving learning performance, and preparing clean, understandable data. Recently, some unique characteristics of big data such as data velocity and data variety present challenges to the feature selection problem

Feature selection, as a type of dimension reduction technique, has been proven to be effective and efficient in handling high dimensional data [8, 9]. It directly selects a subset of relevant features for the model construction. Since feature selection keeps a subset of original features, one of its major merit is that it well maintains the physical meanings of the original feature sets, and gives better model readability and interpretability. Due to this particular reason, it is more widely applied in many real world applications such as gene analysis and text mining. Feature selection obtains relevant features by removing irrelevant and redundant features. The removal of these irrelevant and redundant features reduces the computational and storage costs without significant loss of information or negative degradation of the learning performance.



(a) relevant feature f_1

(b) redundant feature f_2

(c) irrelevant feature f_3

Figure 1: Example of relevant, irrelevant and redundant features

Taking Figure 1 as an example, feature f1 is a relevant feature which can separate two classes (clusters) in Figure 1(a); while in Figure 1(b), feature f2 is considered as a redundant feature w.r.t feature f1 since feature f1 already can discriminate two classes (clusters) well; in Figure 1(c), feature f3 is an irrelevant feature as it does not contain useful information to separate two classes (clusters). According to the availability of class labels, we can categorize feature selection algorithms into supervised and unsupervised methods. Supervised feature selection is usually taken as a preprocessing step for the classification/regression task. It chooses features that can discriminate data instances from different classes or regression targets. Since the label information is known a priori, relevance of a feature is normally assessed by its correlation with class labels. On the other hand, unsupervised feature selection is generally applied for the clustering task. Without class labels to guide feature selection, it evaluates feature importance by some alternative criteria such as data similarity, local discriminative information and data reconstruction error. With regard to search strategies, feature selection algorithms can be divided into wrapper methods, filter methods and embedded methods. Wrapper methods typically use the learning performance of a predefined model to evaluate the feature relevance.

III. Feature Selection Repository

The open source feature selection repository called scikit-feature. The feature selection repository effectively assists researchers to achieve more reliable evaluation in the process of developing new feature selection algorithms. Currently, scikit-feature consists of popular feature selection algorithms in the following categories:

- similarity-based feature selection,
- information theoretical-based feature selection,
- statistical-based feature selection,
- sparse learning-based feature selection,
- wrapper-based feature selection,
- structural feature selection, and
- streaming feature selection.

Among these different categories of feature selection methods, similarity-, information theoretical-, and statistical-based methods correspond to the filter methods discussed above. Wrapper- and sparse learning-based methods correspond to the wrapper methods and embedded methods, respectively. We also include structural features, linked data, multiview, and multisource data to the category of structural feature selection, and streaming data and features to the streaming feature selection category. In addition, scikit-feature provides many

benchmark feature selection datasets and examples of how to evaluate feature selection algorithms via classification or clustering tasks.

When the dimensionality of a dataset grows significantly there is an increasing difficulty in proving the result statistically significant due to the sparsity of the meaningful data in the dataset in question. Large datasets with the so-called "large p, small n" problem (where p is the number of features and n is the number of samples) tend to be prone to overfitting. An overfitted model can mistake small fluctuations for important variance in the data which can lead to classification errors. This difficulty can also increase due to noisy features. Noise in a dataset is defined as "the error in the variance of a measured variable" which can result from errors in measurements or natural variation [10]. Machine learning algorithms tend to be affected by noisy data. Noise should be reduced as much as possible in order to avoid unnecessary complexity in the inferred models and improve the efficiency of the algorithm [11]. Common noise can be divided into two types [12]:

1. Attribute noise.
2. Class noise.

Attribute noise is caused by errors in the attribute values (wrongly measured variables, missing values) while class noise is caused by samples that are labelled to belong in more than one class and/or misclassifications.

As the dimensionality increases the computational cost also increases, usually exponentially. To overcome this problem it is necessary to find a way to reduce the number of features in consideration. Two techniques are often used:

1. Feature subset selection.
2. Feature extraction.

Cancer is among the leading causes of death worldwide accounting for more than 8 million deaths according to the World Health Organization. It is expected that the deaths from cancer will rise to 14 million in the next two decades. Cancer is not a single disease. There are more than 100 known different types of cancer and probably many more. The term cancer is used to describe the abnormal growth of cells that can, for example, form extra tissue called mass and then attack other organs [6].

Microarray databases are a large source of genetic data, which, upon proper analysis, could enhance our understanding of biology and medicine. Many microarray experiments have been designed to investigate the genetic mechanisms of cancer, and analytical approaches have been applied in order to classify different types of cancer or distinguish between cancerous and noncancerous tissue. In the last ten years, machine learning techniques have been investigated in microarray data analysis. Several approaches have been tried in order to

i. distinguish between cancerous and noncancerous samples,
ii. classify different types of cancer, and
iii. identify subtypes of cancer that may progress aggressively.

IV. Feature Subset Selection In Microarray Cancer Data

Feature subset selection works by removing features that are not relevant or are redundant. The subset of features selected should follow the Occam's Razor principle and also give the best performance according to some objective function. In many cases this is an NP-hard (nondeterministic polynomial-time hard) problem [13, 14]. Unlike feature extraction methods, feature selection techniques do not alter the original representation of the data [19]. One objective for both feature subset selection and feature extraction methods is to avoid overfitting the data in order to make further analysis possible. The simplest is feature selection, in which the number of gene probes in an experiment is reduced by selecting only the most significant according to some criterion such as high levels of activity. Feature selection algorithms are separated into three categories [20, 21]:
(i)The filters which extract features from the data without any learning involved.
(ii)The wrappers that use learning techniques to evaluate which features are useful.
(iii)The embedded techniques which combine the feature selection step and the classifier construction.

4.1 Filters

In general, feature selection refers to the process of applying statistical tests to inputs, given a specified output, to determine which columns are more predictive of the output. The Filter Based Feature Selection module provides multiple feature selection algorithms to choose from, including correlation methods such as Pearsons's or Kendall's correlation, mutual information scores, and chi-squared values. When you use the Filter Based Feature Selection module, you provide a dataset, identify the column that contains the label or dependent variable, and then specify a single method to use in measuring feature importance. The module outputs a dataset that contains the best feature columns, as ranked by predictive power. It also outputs the names of the features and their scores from the selected metric.

Filters work without taking the classifier into consideration. This makes them very computationally efficient. They are divided into multivariate and univariate methods. Multivariate methods are able to find relationships among the features, while univariate methods consider each feature separately. Gene ranking is a popular statistical method. The following methods were proposed in order to rank the genes in a dataset based on their significance [22]:
(i)(Univariate) Unconditional Mixture Modelling assumes two different states of the gene on and off and checks whether the underlying binary state of the gene affects the classification using mixture overlap probability.
(ii)(Univariate) Information Gain Ranking approximates the conditional distribution, where is the class label and is the feature vector. Information gain is used as a surrogate for the conditional distribution.
(iii)(Multivariate) Markov Blanket Filtering finds features that are independent of the class label so that removing them will not affect the accuracy.In multivariate methods, pair -scores are used for evaluating gene pairs depending on how well they can separate two classes in an attempt to identify genes that work together to provide a better classification [23].

4.2 Wrappers

Wrapper feature selection alternatives are usually combined with machine learning classifiers to develop a heuristic mechanism that aims to provide an optimal input for targeting optimization functions by considering the options available within a search space boundary. This is performed by the renowned genetic algorithm (GA) [26, 27], particle swarm optimization (PSO) [24, 25], the ensemble learning algorithm [12], extreme learning machines (ELM) [13], ant colony optimization (ACO) [14,15], the imperialist competitive algorithm (ICA) [16], and the harmony search (HS) algorithm [17,18], among others. This distinctive characteristic gives the wrapper method a much-needed robustness and accuracy, especially with regard to massive, multidimensional data processing, which requires a highly sophisticated classification [19].

Wrappers tend to perform better in selecting features since they take the model hypothesis into account by training and testing in the feature space. This leads to the big disadvantage of wrappers, the computational inefficiency which is more apparent as the feature space grows. Unlike filters, they can detect feature dependencies. Wrappers are separated in 2 categories: Randomised and Deterministic. A comparison is shown in Table 1.

Table 1: Deterministic versus randomised wrappers.

Deterministic	Randomised
Small over fitting risk	High overfitting risk
Prone to local optima	Less prone to local optima
Classifier dependent	Classifier dependent
—	Computationally intensive

4.2.1 Deterministic Wrapper

A number of deterministic investigations have been used to examine breast cancer such as a combination of a wrapper and sequential forward selection (SFS). SFS is a deterministic feature selection method that works by using hill-climbing search to add all possible single-attribute expansions to the current subset and evaluate them. It starts from an empty subset of genes and sequentially selects genes, one at a time, until no further improvement is achieved in the evaluation function. The feature that leads to the best score is added permanently [28]. For classification, support vector machines (SVMs), -nearest neighbours, and probabilistic neural networks were used in an attempt to classify between cancerous and noncancerous breast tumours [29]. Very accurate results were achieved using SVMs. The contribution factor, based on minimal error of the support vector machine, of each gene is calculated and ranked. The top ranked genes are chosen for the subset. LOOCSFS is expected to be an accurate estimator of the generalization error while GLGS scales very well with high-dimensional datasets.

4.2.2 Randomised Wrappers

Most randomised wrappers use genetic algorithms (GA) (Algorithm 1) and simulated annealing (Algorithm 2).Best Incremental Ranked Subset (BIRS) [30] is an algorithm that scores genes based on their value and class label and then uses incremental ranked usefulness (based on the Markov blanket) to identify redundant genes. Linear discriminant analysis was used in combination with genetic algorithms. Subsets of

genes are used as chromosomes and the best 10% of each generation is merged with the previous ones. Part of the chromosome is the discriminant coefficient which indicates the importance of a gene for a class label.

A genetic algorithm is run as a first step before the simulated annealing in order to get the fittest individuals as inputs to the simulated annealing algorithm. Each solution is evaluated using Fuzzy -Means (a clustering algorithm that uses coefficients to describe how relevant a feature is to a cluster [31, 32]). The problem with genetic algorithms is that the time complexity becomes , where is the number of samples, is the dimension of the data sets, represents the population size, and is the number of generations. In order for the algorithm to be effective the number of generations and the population size must be quite large. In addition like all wrappers, randomised algorithms take up more CPU time and more memory to run.

V. Conclusion

This paper has presented different ways of reducing the dimensionality of high-dimensional microarray cancer data. The increase in the amount of data to be analysed has made dimensionality reduction methods essential in order to get meaningful results. Different feature selection and feature extraction methods were described. Their advantages and disadvantages were also discussed.

References

Journal Papers:

[1]. Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517.

[2]. S. Mitra, R. Das, Y. Hayashi, Genetic networks and soft computing, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (1) (2011) 94–107, http://dx.doi.org/10.1109/ TCBB.2009.39.

[3]. J.H. Phan, C.F. Quo, M.D. Wang, Cardiovascular genomics: a biomarker identification pipeline, IEEE Trans. Inf. Technol. Biomed. 16 (5) (2012) 809– 822, http://dx.doi.org/10.1109/TITB.2012.2199570.

[4]. C.C.M. Chen, H. Schwender, J. Keith, R. Nunkesser, K. Mengersen, P. Macrossan, Methods for identifying SNP interactions: a review on variations of logic regression, random forest and bayesian logistic regression, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (6) (2011) 1580–1591, http://dx.doi.org/10.1109/ TCBB.2011.46.

[5]. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17, http://dx.doi.org/10.1016/j. csbj.2014.11.005.

[6]. U.B. Neto, Fads and fallacies in the name of small-sample microarray classification – a highlight of misunderstanding and erroneous usage in the applications of genomic signal processing, IEEE Signal Process. Mag. 24 (1) (2007) 91–99, http://dx.doi.org/10.1109/MSP.2007.273062.

[7]. M.P. Liang, O.G. Troyanskaya, A. Laederach, D.L. Brutlag, R.B. Altman, Computational functional genomics, IEEE Signal Process. Mag. 21 (6) (2004) 62–69, http://dx.doi.org/10.1109/MSP.2004.1359143.

[8]. A. L. Blum and R. L. Rivest, "Training a 3-node neural network is NP-complete," Neural Networks, vol. 5, no. 1, pp. 117–127, 1992.

[9]. S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in Proceedings of the 18th International Conference on Machine Learning (ICML '01), pp. 74–81, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 2001.

[10]. J. Han, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 2005.

[11]. D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context," Communications of the ACM, vol. 40, no. 5, pp. 103–110, 1997.

[12]. X. Zhu and X. Wu, "Class noise vs. attribute noise: a quantitative study of their impacts," Artificial Intelligence Review, vol. 22, no. 3, pp. 177–210, 2004.

[13]. A. L. Blum and R. L. Rivest, "Training a 3-node neural network is NP-complete," Neural Networks, vol. 5, no. 1, pp. 117–127, 1992.

[14]. T. R. Hancock, On the Difficulty of Finding Small Consistent Decision Trees, 1989.

[15]. H. George John, Enhancing Learning using Feature and Example Selection (Master thesis), Department of Computer Science, Texas A&M University, 2003.

[16]. X.J. Fu, L.P. Wang, Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance, IEEE Trans. Syst. Man Cybern. Part B Cybern 33 (2003) 399–400.

[17]. L.P. Wang, X.J. Fu, Data Mining with Computational Intelligence, Springer- Verlag, 2005.

[18]. S. Halgamuge, L.P. Wang (Eds.), Classification and Clustering for Knowledge Discovery, Springer, 2005.

[19]. Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007

[20]. A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artificial Intelligence, vol. 97, no. 1-2, pp. 245–271, 1997.

[21]. S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in Proceedings of the 18th International Conference on Machine Learning (ICML '01), pp. 74–81, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 2001

[22]. E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," in Proceedings of the 18th International Conference on Machine Learning, pp. 601–608, Morgan Kaufmann, 2001

[23]. T. Bø and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," Genome biology, vol. 3, no. 4, 2002.

[24]. Zhang Y, Wang S, Phillips P, Ji G. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based Syst. Elsevier B.V.; 2014;64: 22–31.

[25]. Tsai CY, Chen CJ. A PSO-AB classifier for solving sequence classification problems. Appl Soft Comput J. 2015;27: 11–27.

[26]. Soufan O, Kleftogiannis D, Kalnis P, Bajic VB. DWFS: A wrapper feature selection tool based on a parallel Genetic Algorithm. PLoS One. 2015;10. pmid:25719748

[27]. Ma B, Xia Y. A tribe competition-based genetic algorithm for feature selection in pattern classification. Appl Soft Comput. 2017;58: 328–338.

[28]. P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," Pattern Recognition Letters, vol. 15, no. 11, pp. 1119–1125, 1994.
[29]. A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in Proceedings of the 5th International Symposium on Health Informatics and Bioinformatics (HIBIT '10), pp. 114–120, April 2010.
[30]. R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification," Pattern Recognition, vol. 39, no. 12, pp. 2383–2392, 2006.
[31]. J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," Journal of Cybernetics, vol. 3, no. 3, pp. 32–57, 1973.
[32]. J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, Norwell, Mass, USA, 1981.