A Comparative Study on Bioinformatics Feature Selection and Classification

Amal Tamer
Computer Science Dept.
Cairo University
Giza, Cairo, 12613, Egypt

Amr Badr
Computer Science Dept.
Cairo University
Giza, Cairo, 12613, Egypt

ABSTRACT
This paper presents an application of supervised machine learning approaches to the classification of the colon cancer gene expression data. Established feature selection techniques based on principal component analysis (PCA), independent component analysis (ICA), genetic algorithm (GA) and support vector machine (SVM) are, for the first time, applied to this data set to support learning and classification. Different classifiers are implemented to investigate the impact of combining feature selection and classification methods. Learning classifiers implemented include K-Nearest Neighbors (KNN) and support vector machine. Results of comparative studies are provided, demonstrating that effective feature selection is essential to the development of classifiers intended for use in high dimension domains. This research also shows that feature selection helps increase computational efficiency while improving classification accuracy..

General Terms
Hold Out, PCA, SVM, KNN, ICA , Features Classification, Feature Selection, Accuracy, Colon Cancer.

Keywords
PCA, SVM, KNN, ICA

1. INTRODUCTION
Microarray datasets, characterized by a limited number of samples due to the cost of acquisition and usually a larger number of gene expressions, have highlighted the importance of the dimensionality reduction task [1].
Reducing dimensionality of a problem, in many real world problems, is an essential step before any analysis of the data. The general criterion for reducing the dimensionality is the desire to preserve most of the relevant information of the original data according to some optimality criteria. Dimensionality reduction or feature selection has been an active research area in pattern recognition, statistics and data mining communities. The main idea of feature selection is to choose a subset of input features by eliminating features with little or no predictive information. In particular, feature selection removes irrelevant features, increases efficiency of learning tasks, improves learning performance and enhances comprehensibility of learned results [2].
In pattern recognition and general classification problems, methods such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Fisher Linear Discriminate Analysis (LDA) have been extensively used. These methods find a mapping from the original feature space to a lower dimensional feature space [3].
In this paper, we present an integrated framework for gene expression data analysis. Section II describes three techniques for feature selection and dimensionality reduction and then section III studies two techniques for feature classification. Section IV discusses the simulation results and the paper is concluded in section V.

2. FEATURE SELECTION AND DIMENSIONALITY REDUCTION
Feature selection refers to the process of selecting descriptors that are most effective in characterizing a given domain. It addresses the specific task of finding a subset of given features that are useful to solve the domain problem, without disrupting the underlying meaning of the selected features. In this regard, it is related to, but different from, the processes of variable dimensionality reduction and of parameter pruning, although its effect to data set dimensionality is the same as that of the latter [4].
As many pattern recognition techniques were originally not designed to cope with large amounts of irrelevant features, combining them with FS techniques has become a necessity in many applications. The objectives of feature selection are manifold, the most important ones being: (a) to avoid over fitting and improve model performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the case of clustering, (b) to provide faster and more cost-effective models and (c) to gain a deeper insight into the underlying processes that generated the data [5].

2.1 Principal Component Analysis (PCA)
Principal component analysis is a standard statistical technique that can be used to reduce the dimensionality of a data set. This is done by projecting the data of a dimensionality N onto the eigenvectors of their covariance matrix with, usually, the largest M eigenvalues taken (M < N). More formally, each so-called principal component PC_i, i = 1,2,……,M is obtained by linearly combining the original attributes (or features) such that

$$PC_i = \sum_{j=1}^{M} b_{ij} X_j \qquad (1)$$

where Xj is the jth original attribute, and bij are the linear factors (eignvectors) which are chosen so as to make the variance of the corresponding PCi as large as possible. In implementation, the transformation from the original attributes to principal components is carried out through a process by first computing the covariance matrix of the original attributes and then, by extracting its eigenvectors to act as the principal components. The eigenvectors specify a linear mapping from the original attribute space of dimensionality N to a new space of size M in which attributes are uncorrelated. The resulting eigenvectors can be ranked according to the amount of variation in the original data that they account for. Typically, the first few transformed

attributes account for most of the variation in the data set and are retained, while the remainder are discarded.
Note that in contrast with the information gain-based feature ranking, PCA is an unsupervised method which makes no use of information embodied within the class variable. Also, what the PCA returns are linear combinations of the original features. Therefore, the meaning of the original features is not preserved. As opposed to this, selecting a subset of top-ranked features based on information gain ranking will preserve the original meaning of those features selected, performing feature selection in its strict sense.

2.2 Independent Component Analysis ICA and the FastICA Algorithm

ICA is a useful extension of PCA that has been developed in context with blind separation of independent sources from their linear mixtures. Such blind separation techniques have been used in various applications such as auditory signal separating and medical signal processing. In a sense, the starting point of ICA is the uncorrelatedness property of the standard PCA. Roughly speaking, rather than requiring that the coefficients of a linear expansion of the data vectors be uncorrelated, in ICA they must be mutually independent (or as independent as possible). This implies that higher-order statistics are needed in determining the ICA expansion [6]. Considering an n * p data matrix X, whose rows r_i,(i = 1,2,…..,n) correspond to observational variables and whose columns c_j ; (j = 1,2,…., p) are the individuals of the corresponding variables, the ICA model of X can be written as:

$$X=AS \quad (2)$$

Without loss of generality, A is an n * n mixing matrix, and S is an n * p source matrix subject to the condition that the rows of S are as statistically independent as possible.
Those new variables contained in the rows of S are called independent components, that is, the observational variables are linear mixtures of independent components. The statistical independence between variables can be quantified by mutual information $I = \sum_k H(s_k) - H(S)$, where $H(s_k) = p(s_k) \log p(s_k) ds_k$ is the marginal entropy of the variable s_k , $p(s_k)$ is the probabilistic density function, and H(S) is the joint entropy. Estimating the independent components can be accomplished by finding the right linear combinations of the observational variables, since we can invert the mixing as [6]:

$$U = S = A^{-1}X = WX \quad (3)$$

So far there have been a number of algorithms for performing ICA In this paper, we employ the FastICA algorithm to address the problems of tumor classification. In this algorithm, the mutual information is approximated by a contrast function:

$$J(s_k) = (E\{G(s_k)\} - E\{G(v)\})^2 \quad (4)$$

where G is any no quadratic function and v is a normally distributed variable. Like PCA, ICA can remove all linear correlations and only take into account higher-order dependencies in the data. Yet, ICA is superior to PCA since PCA is just sensitive to second-order relationships of the data. In addition, the ICA model usually leaves some freedom of scaling and sorting by convention; the in- dependent components are generally scaled to unit deviation, while their signs and orders can be chosen arbitrarily. In general, the number of independent components equals to the number of the observational variables [6].

2.3 Genetic Algorithm GA

The presented method uses a genetic algorithm for feature selection. Genetic algorithms (GAs), a form of inductive learning strategy, are adaptive search techniques. Genetic algorithms derive their name from the fact that their operations are similar to the mechanics of genetic models of natural systems.
Genetic algorithms typically maintain a constant-sized population of individuals which represent samples of the space to be searched. Each individual is evaluated on the basis of its overall fitness with respect to the given application domain. New individuals (samples of the search space) are produced by selecting high performing individuals to produce ”offspring” which retain many of the features of their ”parents”. This eventually leads to a population that has improved fitness with respect to the given goal.
New individuals (offspring) for the next generation are formed by using two main genetic operators, crossover and mutation. Crossover operates by randomly selecting a point in the two selected parents gene structures and exchanging the remaining segments of the parents to create new offspring. Therefore, crossover combines the features of two individuals to create two similar offspring. Mutation operates by randomly changing one or more components of a selected individual. It acts as a population perturbation operator and is a means for inserting new information into the population. This operator prevents any stagnation that might occur during the search process.
Genetic algorithms have demonstrated substantial improvement over a variety of random and local search methods. This is accomplished by their ability to exploit accumulating information about an initially unknown search space in order to bias subsequent search into promising subspaces. Since GAs are basically a domain independent search technique, they are ideal for applications where domain knowledge and theory is difficult or impossible to provide.
The main issues in applying GAs to any problem are selecting an appropriate representation and an adequate evaluation function. For detailed description of both of these issues for the problem of feature selection.

Representation: The natural representation for the feature selection problem is precisely the one described earlier, namely a binary string of length N representing the presence or absence of each of the N possible features. The advantage of this representation is that the classical GAs operators as described before (binary mutation and crossover) can easily be applied to this representation without any modification. This eliminates the need for designing new genetic operators, or making any other changes to the standard form of genetic algorithms.

Evaluation function: Selecting an appropriate evaluation function is an essential step for successful application of GAs to any problem domain. Evaluation functions provide GAs with the feed-back about the fitness of each individual in the population. GAs then use this feed-back to bias the search process so as to provide an improvement in the populations average fitness. The evaluation function is solely based on the performance of the classification process used, in order to select the appropriate feature set, without attempting to bias the search toward small feature subsets.

3. FEATURES CLASSIFICATION

3.1 SVM Classifier

Support Vector Machines, a technique derived from statistical learning theory, is used to classify points by assigning them to one of two disjoint half spaces. So, SVM performs mainly a (binary) 2-class classification. For linearly separable data, SVM obtains the hyperplane which maximizes the margin (distance) between the training samples and the class boundary. For non linearly separable cases, samples are mapped to a high dimensional space where such a separating hyperplane can be found. The assignment is carried out by means of a mechanism called the kernel function [7].

SVM is widely used in the domain of cancer studies, protein identification and specially in Microarray data. Unfortunately, in many bioinformatics problems the number of features is significantly larger than the number of samples. For this reason, tools for decreasing the number of features in order to improve the classification or to help to identify interesting features in noisy environments are necessary. In addition, SVM can treat data with a large number of features, but it has been shown that its performance is increased by reducing the number of features [7].

3.2 K-Nearest Neighbor (KNN) Classifier

KNN is one of the simplest learning/classification algorithms, and has been successfully applied to a broad range of problems. To classify an unclassified vector X, the KNN algorithm ranks the neighbors of X amongst a given set of N data (X_i, c_i), i = 1,2,…..,N , and uses the class labels c_j,(j = 1,2,….,K) of the K most similar neighbors to predict the class of the new vector X. In particular, the classes of these neighbors are weighted using the similarity between X and each of its neighbors, where similarity is typically measured by the Euclidean distance metric (though any other distance metric may also do). Then, X is assigned the class label with the greatest number of votes among the K nearest class labels [4].

The KNN classifier works based on the intuition that the classification of an instance is likely to be most similar to the classification of other instances that are nearby to it within the vector space. Compared to other classification methods such as Support Vector Machine, KNN does not rely on prior probabilities, and it is computationally efficient if the data set concerned is not very large. If, however, the data sets are large (with a high dimensionality), each distance calculation may become quite expensive. This reinforces the need for employing PCA, ICA, SVM and GA feature ranking to reduce data dimensionality, in order to reduce the computation cost [4].

4. EXPERIMENTAL RESULTS

4.1 Data Set

The colon cancer data set was taken from Kent Ridge Biomedical Data Repository. It has gene expression samples that were analyzed with an Affymetrix Oligonucleotide array complementary to more than 6500 human genes. The data set contains 62 samples collected from colon-cancer patients. Among them, 40 tumor biopsies are from tumors (labeled as ”negative”) and 22 normal (labeled as ”positive”) biopsies are from healthy parts of the colons of the same patients. Two thousand out of around 6500 genes were selected based on the confidence in the measured expression levels [8].

The data set can be expressed in the form of a matrix composed of 62 column vectors

$$X = [x_1, x_2, \ldots\ldots\ldots\ldots\ldots x_N] \quad (5)$$

Each vector x_i has 2000 elements, so that the dimension of the matrix X is 2000 x 62.

4.2 Holdout Cross Validation

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. To avoid over-fitting, an independent test set is preferred. A natural approach is to split the available data into two non-overlapped parts: one for training and the other for testing. The test data is held out and not looked at during training. Hold-out validation avoids the overlap between training data and test data, yielding a more accurate estimate for the generalization performance of the algorithm. The downside is that this procedure does not use all the available data and the results are highly dependent on the choice for the training/test split [9].

4.3 Data PreProcessing

A data preprocessing phase is usually performed prior to the application of any learning algorithm [10]. Instance-wise normalization to mean zero and variance one, Then feature-wise normalization to mean zero and variance one was applied to the colon cancer data on each of the gene expression values, in order to normalize the data. In addition, data were scaled to a range from 0 through 1 for use in the SVM.

4.4 Results and Comparisons

A large corpus of experiments has been carried out. Different classifiers are used to accomplish classification by mapping feature patterns of a different dimensionality onto their underlying functional class types. There are a total of two output classes for the present problem case. The classification performance is measured using hold out cross validation. That is, the gene expression vectors are partitioned into two equally sized subsets, and one subset is used as a train set for a classifier and the another is used for a test. The empirical accuracy is given by the number of error of classifier. A specific point worth noting is that for the KNN classifiers, the results on the use of a varying number K of nearest neighbors are obtained, with K set to 1, 3, 5, 8, 10, 12 and 15. The actual K value of a certain classifier is selected amongst those which leads to the best classification performance. Figure 1 and Table I illustrate the classification performance using KNN and SVM classifiers, in conjunction with the use of PCA, ICA, GA and SVM selection method. Each value indicates the classification accuracy using a different classifier and a different number of selected original features. In this figure, for example, PCASVM means the use of PCA for feature selection , and SVM for classification. We apply the classification across different number of feature selection [10 20 30 50 70 90 120 150 200].

Clearly, different feature sets significantly affect the classification performance. The best performance for each type of classifier with its corresponding feature set is summarized in Table I. It is very interesting to note that the classification performance using KNN can be better than that of using the SVM. In particular, for KNN, much better results can be obtained if careful selection of a subset of features is carried out.

Table 1. Accuracy Performance of Different Feature Selection and Classification Methods

Sel-Class	Number of Features								
	10	20	30	50	70	90	120	150	200
PCA-SVM	68	68	74	84	81	77	87	87	84
ICA-SVM	60	56	54	61	70	73	71	76	69
SVM-SVM	68	83	83	83	83	83	80	81	82
GA-SVM	83	89	91	90	88	93	90	91	90
PCA-KNN	96	97	94	98	95	95	96	100	100
ICA-KNN	68	87	91	94	94	93	95	96	99
SVM-KNN	98	100	97	97	95	98	97	100	100
GA-KNN	78	89	91	92	91	89	87	94	95

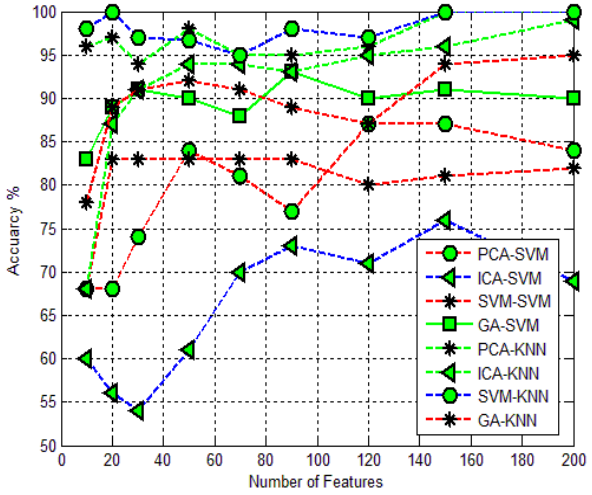

Fig 1: The Accuracy performance with different feature selections and classification.

5. CONCLUSION

This paper has presented an experiment-based comparative study of two classification methods applied to the colon cancer gene expression data set. The work is itself employed in conjunction with the learning process of each classifier considered to address the difficulties in handling real problems represented by this data set. It has shown that in general, attribute selection is beneficial for improving the performance of these common learning algorithms. It has also shown that, as with the learning algorithms, there is no single best approach for all situations involving dimensionality reduction or feature selection. This investigation has helped to reinforce the fact that when building a practical classifier, what is needed is not only an understanding of how different learning algorithms work, but also when they work the best with what kind of support attainable from feature selection, as well as what background knowledge is available about the data in the given domain. In particular, this work has investigated the following two classification algorithms: K-Nearest Neighbors, Support Vector machine; and the following four methods for making choice of features: PCA, ICA, GA and SVM. Comparative studies have been performed between the use of different feature set and that of a subset. Amongst a large corpus of systematic experimental studies carried out, the best classification accuracy is achieved by using a subset of features chosen PCA and SVM method for KNN classifiers. It is very interesting, and would be potentially very beneficial to the relevant research communities, to investigate such alternatives and compare their performance with the ones studied here, including the use of most recently developed feature selection techniques (e.g. [11]). In addition, the present work is focused on a highly specific and unbalanced data set. It would also be useful to examine this data set more carefully, in terms of missing values and inconsistent relationships. This will help extend the useful findings to other problem domains.

6. REFERENCES

[1] Canul Reich, J. L. Hall, O. , Goldgof , D., and Eschrichy, S. A. 2008 "Feature selection for microarray data by auc analysis," in IEEE SMC.

[2] Sheikhi, N. , Rahmani A., and Veisisheikhrobat, R. 2011 "An unsupervised feature selection method based ongenetic algorithm," International Journal of Computer Science and Information Security.

[3] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, 2007 "Feature selection using principal feature analysis," in ACM Multimedia, Augsburg, Germany.

[4] C. Shang and Q. Shen, 2005. "Aiding classification of gene expression data with feature selection: A comparative study," International Journal of Computational Intelligence Research., vol. 1, no. 1, p. 68-76.

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

[6] C.-H. Zheng, D. Huang, X.-Z. Kong, and X.-M. Zhao,2008 "Gene expression data classification using consensus independent component analysis," Genomics Proteomics & Bioinformatics, vol. 6, no. 2, pp. 74–82.

[7] E. Alba, J. Garcia-Nieto, L. Jourdan, and E.-G. Talbi, 2007 "Gene selection in cancer classification using pso/svm and ga/svm hybrid algorithms," in IEEE Congress on Evolutionary Computation, pp. 284–290.

[8] S. Mahadevi Alladi, S. Santosh P., V. Ravi, and U. Suryanarayana Murthy,2008 "Colon cancer prediction with genetic profiles using intelligent techniques," Biomedical Informatics Publishing Group, vol. 3, no. 2, pp. 130–133.

[9] P. Refaeilzadeh, L. Tang, and H. Liu, 2008 "Cross-validation," Arizona State University, Tech. Report

[10] D. Pyle, 1999, Data preparation for data mining. California,USA: Morgan Kaufmanns.

[11] R. Jensen and Q. Shen, 2004 "Semantics-preserving dimensionality reduction: rough and fuzzy-rough approaches." IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 12, pp. 1457–1471.