International Journal of Soft Computing 11 (5): 334-342, 2016
ISSN: 1816-9503
© Medwell Journals, 2016

Review of the Effect of Feature Selection for Microarray Data on the Classification Accuracy for Cancer Data Sets

Naeimeh Elkhani and Ravie Chandren Muniyandi
Center for Software Technology and Management,
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

Abstract: DNA microarrays can be used to monitor the expression level of thousands of genes simultaneously and gene microarray data can be used in cancer diagnosis and classification. Many machine learning techniques have been developed for computational analyses of microarray data. A common difficulty for all techniques is the large number of genes compared to the small sample size which has a negative impact on their speed and accuracy. To overcome these limitations, feature selection techniques are applied to distinguish between significant and redundant or irrelevant genes. Feature selection methods are used for two main goals. The first is to identify the relationship between specific diseases and genes. The second is to examine a compact set of discriminative genes to develop a pattern classifier with good generalizability and limited complexity. Here, we review different feature selection methods for cancer microarray data sets and analyze their accuracy. We describe methods commonly used for selecting significant features including filters, wrappers and embedded methods, categorized according to their experimental methodology. We then compare the classification accuracy of the methods for various cancer data sets and their time complexity to make some suggestions regarding the use of suitable methods for cancer data sets.

Key words: Microarray cancer data sets, feature selection methods, classification accuracy, wrappers, experimental

INTRODUCTION

Analysis of microarray data for thousands of genes in arrays of abnormal and normal cells is an effective approach for investigating gene expression in cancer. Microarray data trials involve small samples (as small as a few dozen) and gene expression of high dimensionality (as high as a few thousand). A very high number of genes are found to be irrelevant for analysis which may hinder correct prediction (Li, 2006; Li and Yang, 2005; Nguyen and Rocke, 2002; Tan *et al*., 2004).

Cancer is a major public health issue because of difficulties in early diagnosis and treatment. Although cancer fatalities continuously declined from 215.1 deaths per 100,000 in 1991-171.8 in 2010, this was mostly because of widespread screening and early diagnosis rather than effective cancer management strategies and therapies. Diagnosis tools based on gene expression profiles have made a significant contribution to progress in cancer studies. DNA microarrays can be used to monitor the expression levels of thousands of genes simultaneously Harrington *et al*., 2000) for cancer diagnosis, prognosis (Sungheetha and Suganthi, 2013) and classification (Perez-Diez *et al*., 2007). These techniques are used to extract patterns and build classification models for gene expression data and have significantly aided in cancer prediction (Rocha *et al*., 2007) and prognosis (Vanneschi *et al*., 2011).

A common difficulty for all these techniques is the large number of genes (features) compared to the small sample size, which has a negative impact on their speed and accuracy. To overcome this limitation, feature selection techniques are applied to recognize differentially expressed genes from redundant genes and remove irrelevant genes. Feature selection can improve the accuracy and speed of classification systems by reducing dimensionality (Shang and Shen, 2005). Some of the improvements are (Wang *et al*., 2015; Zhou and Dickerson, 2014). According to Nguyen and Rocke (2002), feature selection involves selecting a subset of novel features to investigate the relationships between specific diseases and genes and to identify a compact set of

Corresponding Author: Naeimeh Elkhani Center for Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

discriminative genes to develop a pattern classifier with good generalizability and limited complexity. A metric typically used to assess the performance of feature selection methods is the classification accuracy defined as the number of correct predictions made divided by the total number of predictions made multiplied by 100:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

Where:
FP = False positives
TN = True negatives
FN = False negatives
TP = True positives

Hira and Gillies (2015) reviewed different feature selection and extraction methods and their classification accuracy in terms of the number of genes evaluated. The researchers concluded that incorporation of prior knowledge from various biological sources increases the accuracy and reduces the computational complexity. However, their conclusions regarding robust feature selection methods were made without investigating experimental evidence for different methods or different microarray data sets and without reviewing variations in classification accuracy for different combinations of feature selection methods and classifiers. This deficiency motivated us to review feature selection methods with a specific focus on the classification accuracy for different cancer data sets. We therefore compared the robustness of feature selection methods for classification in terms of accuracy for the most popular cancer microarray data sets.

Feature selection methods for cancer microarray data: In general, three main methods are used for feature selection in microarray data sets: wrapper, filter and embedded methods. Wrappers which are general purpose algorithms, search the feature space and then test the performance of subsets according to a learning algorithm. A classifier is required to evaluate the performance quality. Filter methods mostly use a feature ranking function that assigns a relevance score to each feature. A higher rank is allocated to features that are more relevant. Filter methods are independent of classification algorithms. Embedded methods select features during an implicit process for learning of optimal parameters. As in wrapper methods, feature selection depends on a classification algorithm.

Randomized wrappers: According to Hira and Gillies (2015), wrappers can be divided into two categories: randomized and deterministic strategies. Although, they have a higher risk of overfitting in comparison to deterministic methods, randomized wrappers are less prone to local optima and are computationally intensive. Randomized wrapper methods which mostly use Genetic Algorithms (GAs) are more prevalent in feature selection for microarray cancer data.

The GA was initially developed by Holland (1975). According to Li *et al*. (2001), the GA imitates natural selection. In addition, this algorithm helps in identifying optimal solutions by mimicking evolution in biological systems (Liu *et al*., 2004). We determined the performance of wrapper methods such as the GA for feature selection in combination with different classifiers.

GA/K Nearest Neighbors (KNN): The GA/KNN approach combines a GA feature selection method and a KNN classifier to identify genes that can discriminate between different sample classes such as tumor and normal tissue (Li *et al*., 2001). The GA feature selection method can identify small subsets of genes for training and then applies an evolutionary tool. Srinivas and Patnaik (1994) demonstrated that the GA/KNN Method can identify the existence of different subtypes within classes. The method can also be used in computationally intensive searches for many chromosomes (near-optimal solutions) for which approximately 10,000 near-optimal solutions are usually needed for a typical run. The parameters are first loaded and then the initial generation is produced using approximately, 200 chromosomes. Each chromosome has 30 genes that are picked randomly from the gene pool. After a preselection step to enhance the probability of identifying the most optimal chromosomes, the top 100 chromosomes are chosen to create the initial generation. In the next stage, the program goes through mutation and crossover operations. If optimal chromosomes are found, they are recorded and placed into a panel of discriminated genes; mutation and crossover operations are run for five more loops before beginning a new iteration with a new initial generation. The next generation is subjected to mutation and crossover operations up to the maximum number of iterations. This entire process is continued until the recognized number of optimal chromosomes is produced.

The principle underlying KNN classifiers is supervised learning. A classifier initially searches for the nearest k samples when a new sample appears in the pre-existing training data. These new samples are classified according to the most similar class.

Adaptive GA/KNN (AGA/KNN): A GA solely finds the nearest optimal solution while in each run of high-dimensional data, nearest optimal string is not similar. To solve this problem, AGA was developed by adding three techniques which were comprised of immigration and extinction strategy, adaptive possibilities of mutation and crossover and elitist strategy. Afterwards, it was combined with KNN. The reason for the combination of AGA and KNN is: some assumptions are needed for most feature selection methods while to use in high-dimensional scopes they are not appropriate. AGA is an appropriate search tool for analyzing high-dimensional and noisy data, since it follows biological principles in searching for near-optimal strings. KNN is a simple and effective classifier among trialed and implemented classification techniques.

The examining AGA/KNN in different areas indicates satisfactory result (Li *et al*., 2001; Srinivas and Patnaik, 1994). Lee *et al*. (2011) were the first to use a combination of KNN and AGA in five different stages: termination, genetic operators (this element incorporates the adaptive probabilities and selection of mutation and crossover), fitness function, initial population and encoding.

Ga/support Vector Machine (GA/SVM): GA/SVM is found to be the genetic algorithm which is based on the SVM classification utilized to identify or recognize the optimal parameters for a conventional SVM classifier (Chen and Yang, 2012).

Studies show that SVM is one of the most powerful and integrated learning classifiers which offers great opportunity to have an effective pattern recognition approach. Initially, the paradigm of SVM classifier was identified by various researchers and practitioners (Huang and Wang, 2006). It has been observed that SVM utilizes a linear separating plane which is referred as the hyper-plane. One of the major objectives of this plane is to increase or maximize the distance amid two classes (Pierna *et al*., 2004).

Huang and Wang (2006) reported that four common kernel functions are used as SVM classifiers: sigmoid, radial basic function, polynomial and linear functions. The kernel parameters for these functions need to be set properly to maximize the SVM classification accuracy (Pierna *et al*., 2004; Huang and Wang, 2006).

Binary Coded GA (BCGA) and Real Coded GA (RCGA): A BCGA is a probabilistic search algorithm that transforms a population set (mathematical objects with a uniform length) to a new offspring according to the Darwinian principle of natural selection. In particular, a set of chromosomes that individually represent one probable solution is modified and transformed via genetic processes to create a new population. The entire process continues up to a predetermined number of iterations or until further enhancement and improvements are achieved. According to a mutation operator, a given chromosome or a gene is selected in a random manner and its value is exchanged (e.g., 1 for 0 and vice versa) (Holland, 1975).

RCGA operates on a population set that represents a variable of the problem and the chromosome size is kept the same as the length of the problem solution. The starting point is the initial population and then the main loop algorithm is run. The main loop comprises preprocessing, three genetic operations and post-processing and it continues until the termination condition is satisfied. The process includes a fitness function test to avoid premature junctions in the initial stages of the evolution process and to stimulate convergence in the more advanced stages of the process. In addition, genetic operators incorporate mutation, crossover and selection.

MATERIALS AND METHODS

Information Gain (IG) is a popular feature selection method used to rank genes in a data set according to their significance. According to Hira and Gillies (2015), IG is a univariate filter method. Univariate IG ranking approximates the conditional distribution (C/F), where C is the class label and F is the feature vector. IG is used as a surrogate for the conditional distribution.

Another popular method is based on t-score feature selection. Markov blanket filtering based on t-scores is categorized as a multivariate filter method (Hira and Gillies, 2015). Multivariate Markov blanket filtering finds features that are independent of the class label so that their removal will not affect the accuracy. In multivariate methods, paired t-scores are used to evaluate gene pairs depending on how well they can separate two classes; the aim is to identify genes that work together to provide a better classification (Bo and Jonassen, 2002).

Multivariate methods are able to find relationships among the features while univariate methods consider each feature separately. We now describe examples of these filter methods.

IG-GA/KNN: IG is a feature ranking technique based on decision trees and has good classification performance (Valdivia *et al*., 2008). The principle behind IG is to choose features that present information

about classes. According to Mukras, Wiratunga, Lothian, Chakraborti and Harper, these features are discriminative in nature and occur within a single class. In the initial stage, a subset of the original feature set is usually acquired by implementing IG in the form of filtering criteria. This is usually performed by categorizing the genes. Genes with an information value that exceeds the threshold are eligible to enter the next stage. In the second stage the GA is applied to the set of filtered genes. The IG-GA/KNN method uses a KNN classifier to check the cross-validation accuracy.

Hybrid Particle Swarm Optimization Tabu Search (HPSOTS): HPSOTS is a hybrid classification model comprising Particle Swarm Optimization (PSO) and Tabu Search (TS) (Shen *et al*., 2008). Shen *et al*. (2008) proposed a modified discrete PSO with the information-sharing mechanism of PSO. Before the heuristic search procedure, genes with lower absolute t-test values among normal and tumor samples are removed. A heuristic search method (HPSOTS, pure TS and PSO) is then applied to the data set.

TS can provide solutions for various difficult optimization problems (Glover, 1986). TS is an iterative process in which the fitness function of a random solution is first evaluated. Then, by tracing the current suggested solution, all neighbors of this solution are identified and evaluated. The tracing is based on a primitive transformation. A new current solution is selected if TS does not identify best neighbors for the current solution. The best neighbor identified is constantly compared with the current solution; if it is worse than the new one, TS tracing is continued upwards. Using this approach, local minima can be easily overcome (Shen *et al*., 2008).

Minimum Redundancy Maximum Relevance (mRMR): mRMR is a two-stage feature selection algorithm that is based on a specific formula (Peng *et al*., 2005). In the first step, a contender feature set is allocated with the feature selection method, specifically mRMR feature selection method. Afterwards, these schemes are used in order to select compact subsets from nominated sets.

Similarity-Preserving Feature Selection (SPFS): SPFS is used for samples with redundant features. The process starts with a conventional combinatorial optimization formula, K-$X_A X_A^T$ where = (f_{i1}, …, f_{ik}), $i_p \sigma A$, p = 1, …, k. Feature selection in the SPFS framework can be established as a multiple-output regression problem (Zhao *et al*., 2011).

Trace Ratio (TR): TR is an iterative algorithm developed to identify the optimal subset of features for which the subset level score is maximized.

Embedded models: In embedded models, similar to wrapper methods, feature selection is linked to the classification stage but this link is much stronger. Embedded methods offer the same advantages as wrapper methods concerning the interaction between feature selection and classification. Moreover, they have better computational complexity since feature selection is directly included in the classifier construction during training. The Genetic Swarm Algorithm (GSA), Large-Margin Subspace Learning (LMSL) and Local Linear Feature Selection (LLFS) are examples of embedded methods.

GSA: GSA combines the strengths of the GA and PSO (Kumar *et al*., 2012). The GSA design is based on a fuzzy expert system for classification of microarray data. PSO is a population-based algorithm that uses swarm intelligence to solve optimization problems. Each individual in a population is referred to as a particle and designates a solution. Moreover, each particle flies with an adaptable speed in the search space according to its own experience or that of other particles. Thus, each particle tries to progress by imitating the traits of other particles traits. This is possible because each particle has memory to hold positions in the search space that are met. The best position is denoted pbest and the best particle in the population is denoted gbest (Tse and Tso, 1993).

LMSL: LMSL is a subspace learning algorithm that is based on a large-margin framework (Liu *et al*., 2013). Initially, it uses the nearest neighbor along with the same label and different labels for a given sample.

LLFS: Sun *et al*. (2010) proposed LLFS which is based on well-structured numerical and machine learning analysis techniques, without making any assumptions regarding the underlying allocation of data. LLFS can process a wide range of features within minutes on a personal computer while achieving quite high reliability and integrity that is approximately insensitive to an increasing number of irrelevant features. Table 1 shows the methodology used by various models for analysis of microarray data.

Algorithm accuracy for cancer data sets: Yang *et al*., 2010) presented an IG-GA/KNN framework combining a wrapper method (GA) and filter method (IG)

Table 1: Methodology used by models for analysis of microarray data

Model	Type	Reference	Methodology
LMSL/SVM	Embedded	(Liu *et al*., 2013)	In three algorithms including (LMSL, LLFS, RFS) fivefold cross validation used to determine regularization parameter. Samples randomly selected 70%-30% as training and test data, respectively. Finally, the highlighted features are classified by linear SVM to determine accuracy of classification
RFS/SVM	Embedded		
LLFS/SVM	Embedded		
SPFS/SVM	Filter		
mRMR/SVM	Filter	(Shen *et al*., 2008)	Before a heuristic search, genes of top-ranked are selected by t-test filtering algorithm. In terms of datasets, 50 colon samples are randomly selected as training datasets from 62 samples and 12 samples are used as test dataset. The result of classification shows 83.87% accuracy by fivefold cross validation. In the case of leukemia 38 samples are selected as training dataset (27 ALL, 11 AML) and 24 samples (20 ALL, 14 AML) are selected as test data. The result of classification shows 90.28% accuracy by fivefold cross validation. In terms of breast cancer, from a total of 49 breast tumor samples, 40 samples are selected as training set and the rest nine samples are selected as test dataset. The result of classification shows 85.71% accuracy by fivefold cross validation
TR/SVM	Filter		
t-test/HPSOTS	Filter		
GA/SVM	Wrapper	(Chen and Yang, 2012)	In this study Four different methods including: all genes (All), 70 correlation-selected genes (C70), 15 medical literature-selected genes (R15), and 50 t-test-selected genes (T50) are used for gene selection. The results of classification accuracy indicate 95% for T50 and 90% for C70 or R15
AGA/KNNGA/KNN	WrapperWrapper	(Lee *et al*., 2011)	The study used three groups of genes including; group 1, 50 genes which are selected by AGA/KNN, group 2, 50 genes which have the smallest max-T adjusted p value and group 3 with 50 genes are selected randomly. First dataset is colon data with 62 samples which 40 are tumor and 22 are normal genes. From 62 samples, 40 samples selected as training set and the remaining are test datasets. Second data set is Small, Round Blue Cell Tumors (SRBCTs) with 2308 genes are divided to 63 training samples from 23 tumors
GSABC	Embedded	(Kumar *et al*., 2012)	For gene selection from the original gene profile, mutual information technique is used followed by fuzzy expert system for classification. Fuzzy system is including if-then rules using GA and membership functions evolving by PSO. Standard Leave-One-Out Cross-Validation (LOOCV) determine generalizability of the proposed system. Data sets including colon cancer, leukemia and lymphoma are considered in simulations
GARCGA	WrapperWrapper		
IG-GA/KNNIG/ KNNGA/KNN	FilterFilter Wrapper	(Yang *et al*., 2010)	First, information gain used for feature selection, second GA as a random wrapper method followed by KNN classifier. Standard Leave-one-Out Cross-Validation (LOOCV) determine generalizability and accuracy of the proposed system
NRGAIG/KNNIG-GA/ KNNIG-NRGA/KNN	FilterFilter FilterFilter	(Sungheetha and Suganthi, 2013)	First, information gain and genetic algorithm used for pre-feature selection of microarray data, then, non-dominated ranked GA (NRGA) is used as actual feature selection and KNN used to evaluate the NRGA algorithm. The details of datasets are as follow: Brain tumor: five human brain tumor types, 90 samples, 5920 genesLung cancer: five lung cancer types and normal tissues, 203 samples, 1260 genes Prostate tumor: two prostate tumors and normal tissues, 102 samples, 10509 genes

for feature selection among microarray data sets (Table 2). They used IG to choose significant gene subsets from all elements in the gene expression data, and applied a GA for selection of actual features. The KNN method with LOOCV was applied to evaluate IG-GA. Using a KNN classifier, the accuracy for a brain cancer data set was 93.33% for IG-GA feature selection, compared to 88.89% for IG and 92.22% for GA. For a leukemia data set the accuracy was 100% for IG-GA, 93.06% for IG and 97.22% for GA. For a lung cancer data set the accuracy was 95.57% for IG-GA, compared to 90.15% for IG and 94.09% for GA. For a prostate cancer data set, IG-GA had 96.08% accuracy, compared to 89.22% for IG and 91.18% for GA.

Lee *et al*. (2011) found that a KNN classifier with AGA feature selection can reduce the dimensionality of a data set. For pediatric Small, Round, Blue-Cell Tumors (SRBCTs), all test samples were categorized correctly after 70 runs while with GA feature selection accuracy reached just 80% after 1000 runs.

Chen and Yang (2012) applied the GASVM model to data for 97 patients with breast cancer. They used four different gene selection strategies: all genes, 70 correlation-selected genes, 15 medical literature-selected genes and 50 t-test-selected genes. GA feature selection improved the SVM classification accuracy (90%) in comparison to SVM (60%), correlation (83%), decision tree (89.47%), nearest-centroid with multiple random validation (69%), Bayesian network (74%), ANN (78.65%), and nearest neighbor (76.34%) methods.

Kumar *et al*. (2012) compared the performance of GSA, PSO, RCGA and BCGA. For a colon cancer data set, GSA (including GA feature selection and PSO) achieved 58.7% accuracy, compared to 56.5% for BCGA, 52.8% for RCGA and 51.2% for PSO. For leukemia classification, GSA yielded 81.2% accuracy, compared to 79.1% for BCGA, 75.5% for RCGA and 76.3% for PSO. For a lymphoma data set, GSA achieved 69.5% accuracy, compared to 65.2% for BCGA, 66.7% for RCGA and 68.9%

Table 2: Accuracy of algorithms for cancer data sets

Author/s	Method	Proposed algorithm	Comparator algorithm/s	Classification accuracy (%)		Data set
				Proposed algorithm	Comparator algorithm/s	
Yang *et al.* (2010)	Statistics, machine learning	IG-GA/KNN	IG/KNN	93.33	88.89	Brain tumor
			GA/KNN		92.22	
				100.00	93.06	Leukemia
					97.22	
				95.57	90.15	Lung cancer
					94.09	
				96.08	89.22	Prostate tumor
					91.18	
Lee *et al.* (2011)	Machine learning	AGA/KNN	GA/KNN	~90% after 40 runs, increasing to ~100% after 70 runs	~80% when >1000 runs were executed	Pediatric SRBCTs
Chen and Yang (2012)	Machine learning	GASVM	+SVM+Correlation-based method+ Decision tree+ Nearest-centroid with multiple random validation+Bayesian network+ANN+Nearest neighbors	90	60 83 89.47 69 74 78.65 76.34	Breast cancer
(Kumar *et al.* (2012)	Fuzzy expert system	GSA	BCGA RCGA PSO	58.7	56.5 52.8 51.2	Colon cancer
				81.2	79.1 75.5 76.3	Leukemia
				69.5	65.2 66.7 68.9	Lymphoma
Sungheetha and Suganthi (2013)	Machine learning	NRGA/KNN	IG/KNN	89.1	7073.4	Brain tumor
			IG-GA/KNN	77.4	70.15 74.8	Lung cancer
				86.3	82.22 77.6	Prostate tumor
Shen *et al.* (2008)	Machine learning	HPSOTS	Pure TS Pure PSO	93.55	90.32 90.33	Colon
Liu *et al.* (2013)	Machine learning	LMSL	RFS LLFSS PFS mRMRTR	95.61	95.10 93.10 94.73 94.52 95.03	Lung
				92.31	91.31 91.46 79.79 81.51	Prostate

for PSO. The simulation results show that GSA generates a compact and integrated fuzzy system with higher levels of accuracy for all the data sets compared to the other approaches.

A hybrid NRGA/KNN proposed by Sungheetha and Suganthi (2013) incorporates IG GA for feature selection in microarray data sets. IG is used to choose significant gene subsets from all elements in the gene expression data, whereas NRGA is applied for selection of actual features. The KNN method is utilized to examine the NRGA algorithm. Using the KNN classifier for a brain cancer data set, NRGA feature selection achieved 89.1% accuracy, compared to 70% for IG and 73.4% for IG-GA. For lung cancer classification, NRGA yielded 77.4% accuracy, compared to 70.15% for IG and 74.8% for IG-GA. For a prostate cancer data set, NRGA achieved 86.3% accuracy, compared to 82.22% for IG and 77.6% for IG-GA. The experimental results indicate that NRGA/KNN effectively simplifies the number of gene expression levels and provides more accurate and reliable classification.

Shen *et al.* (2008) compared the performance of HPSOTS to that of pure PSO and TS algorithms. For a colon cancer data set, t-test feature selection with the HPSOTS classifier achieved 93.55% accuracy compared to 90.32% for a t-test with pure TS and 90.33% for a *t*-test with pure PSO.

Liu *et al.* (2013) performed a wide range of experiments to evaluate the LMSL efficiency in comparison to five characteristic feature selection algorithms. Using SVM as the classifier for a lung cancer data set, LMSL for feature selection yielded better accuracy (95.61%) than RFS (95.10%), LLFS (93.10%), SPFS (94.73%), mRMR (94.52%) and TR (95.03%). LMSL also achieved better accuracy (92.31%) than RFS (91.31%), LLFS (91.46%), SPFS (77.93%), mRMR (79.79%), and TR (81.51%) for a prostate cancer data set.

The RFS and LLFS algorithms are closely associated with LMSL: the principle of large margins underlies both LLFS and LMSL and LMSL benefits from RFS for effective resolution of objective functions. SPFS, mRMR and TR are state-of-the-art feature selection algorithms with different effective characteristics. mRMR removes redundant elements by considering them in a pairwise manner. TR characterizes data set structures via a Laplacian graph and has considerably better performance than similar algorithms such as Laplacian Score. The SVM classifier showed accuracy of 30% which is much lower than the accuracy of TR, RFS and LMSL. After RFS LMSL is the next fastest method. Moreover, LMSL takes more time than RFS in PROS (0.06 sec). LLFS is considerably slower than LMSL for all three data sets. LMSL is slower than RFS because it requires calculation of the sample margins for improved and integrated feature selection and better performance.

RESULTS AND DISCUSSION

The most important issue for models built for classification problems is the model accuracy, defined as the number of correct predictions among all predictions made. Accuracy can be misleading, as it is sometimes desirable to select a model with lower accuracy but greater predictive power. When tp< fp, then accuracy will always increase as the classification rule changes to always produce a "negative" output. Conversely, when tN<FN, the same holds true as the rule changes to always generate a "positive" output. Thus, for the accuracy results reported from the studies reviewed here, it is supposed that the possibility of misleading classification accuracy was considered.

Yang *et al.* (2010) found similar classification accuracy for IG-GA/KNN and GA-KNN for brain cancer, leukemia and lung cancer data sets (Table 3). For a prostate cancer data set, the GA-KNN accuracy was 4.9% better than that of IG-GA/KNN. Thus, the IG filter did not greatly improve the GA-KNN accuracy. However, significant gap was reported when IG-GA/KNN was compared with IG-KNN as a pure filter method. The accuracy result of IG-GA/KNN shows +4.44% improvement in brain cancer classification accuracy, +6.94% in leukemia dataset, +5.42% in lung and +6.86% in prostate. Thus, filter feature selection (IG) combined with wrapper feature selection (GA) yields better classification accuracy than a pure filter method (IG-KNN) for all cancer data sets investigated, although IG did not add much value to GA-KNN as a wrapper method. Sungheetha and Suganthi (2013) used IG to select important subsets among features in a gene expression data set and an NRGA for actual feature selection. The accuracy of NRGA was compared to that of IG-GA and IG feature selection methods. The results indicate that NRGA/KNN outperforms IG-GA/KNN for brain, lung and prostate cancer data sets. For brain cancer classification, NRGA/KNN improved the accuracy by +15.7% in comparison to IG-GA/KNN. For lung and prostate cancers, the improvements in accuracy were +2.6% and +8.7%, respectively. NRGA feature selection achieved an absolute improvement in accuracy of +19.1% for brain, +7.25% for lung and +4.08% for prostate cancer classification in comparison tothe pure IG filter method.

Comparison of AGA/KNN and GA/KNN shows that both algorithms are good at reducing dimensionality. However, the efficiency and classification rate of AGA/KNN are better than those of GA/KNN for pediatric SRBCTs (Lee *et al.*, 2011). Thus, researchers can use AGA/KNN to perform dimension reduction when analyzing microarray data. Then biologists can efficiently identify relevant genes from gene subsets and correctly classify test samples. The results indicate that the GASVM Model has the potential to better assist physicians in assessing breast cancer prognosis using both clinical and microarray data (Chen and Yang, 2012).

In simulations by Kumar *et al.* (2012), the GSA approach generated a compact fuzzy system with better classification accuracy for colon cancer, leukemia, and lymphoma data sets when compared with BCGA, RCGA, and PSO. GSA led to a significant improvement for colon and leukemia data sets in comparison to RCGA (+5.9, +5.7%) and PSO (+7.5, 4.9%). Although, BCGA can yield comparable accuracy, the results obtained are hard to interpret. The main strength of GSA is that it provides a deeper understanding of biological and clinical issues because of its fuzzy system in contrast to so-called black box methods that focus only on classification performance.

Liu *et al.* (2013) found that the LMSL method significantly improved the accuracy of prostate cancer classification compared to SPFS (+14.38%), mRMR (12.52%) and TR (10.8%) but did not differ much in accuracy compared to RFS and LLFS. In the case of a lung cancer data set, there was generally not much difference in accuracy among the LMSL, RFS, LLFS, SPFS, mRMR and TR methods.

According to Table 3, the time complexity is slightly lower for IG/KNN and GA/KNN than for IG-GA/KNN. However, many studies used a filter (e.g., IG) to select relevant feature information for IG/KNN and GA/KNN, thereby reducing the number of features compared to a GA. Table 3 shows that the computation time is shorter for a KNN classifier than for an SVM. KNN is a readily available algorithm that needs few parameters, so it is often used for classification problems. According to,

Table 3: Time complexity of the microarray analysis methods

Methods	Time cost	References
IG-GA/KNN	O (n log n+nm pg)[a]	
GA/KNN	O (nmpg)[a]	
IG/KNN	O (n log n+nm)[a]	
SVM	O ($C^4_n \log^n/\varepsilon_n$)	
KNN	O(nm)	Yang *et al.* (2010)
AGA/KNN (120 runs)	10 min	
GA/KNN (1000 runs)	26 min	
BCGA	318.8 s	Lee *et al.* (2011)
RCGA	287.8 s	
PSO	216.5 s	
GSA	248.2 s	
	Brain, lung, prostate data sets	Kumar *et al.* (2012)
IG KNN	0.354 ms, 0.762 ms, 0.482 ms	
IG GA KNN	0.258 ms, 0.673 ms, 0.363 ms	
NRGA KNN	0.248 ms, 0.543 ms, 0.313 ms	

[a]n is the number of samples, m is the dimension of the data sets, p is the population size and g is the number of generations; [b]C is the number of classes, n is the size of the training set and en is obtained via an appropriately normaliz

Lee *et al*., 2011), the AGA/KNN time cost was significantly better (by 16 min) than that of GA/KNN. The GSA method proposed by Kumar *et al.* (2012) has a lower time cost compared to RCGA and BCGA but not pure PSO. Further, investigations with different methodologies revealed that NRGA/KNN has better accuracy than IG/KNN and IG-GA/KNN and a lower time cost (Sungheetha and Suganthi, 2013).

CONCLUSION

We reviewed feature selection methods for reducing the dimensionality of microarrays to improve the classification accuracy for cancer data. We described and categorized filter, wrapper and embedded approaches according to their methodology. We compared the methods in terms of their classification accuracy for various cancer microarray data sets. We discussed improvements in accuracy and the time complexity of selection methods for specific cancer data sets, and made some suggestions regarding suitable methods for cancer data sets.

ACKNOWLEDGEMENTS

Researchers acknowledge grant support from the Development of Membrane Computing Software (project number 01-01-02-SF1104) for this research.

REFERENCES

Bo, T.H. and I. Jonassen, 2002. New feature subset selection procedures for classification of expression profiles. Genome Boil., 3: 1-11.

Chen, A.H. and C. Yang, 2012. The improvement of breast cancer prognosis accuracy from integrated gene expression and clinical data. Expert Syst. Appl., 39: 4785-4795.

Glover, F., 1986. Future paths for integer programming and links to artificial intelligence. Comput. Operat. Res., 13: 533-549.

Harrington, C.A., C. Rosenow and J. Retief, 2000. Monitoring gene expression using DNA microarrays. Current Opin. Microbiol., 3: 285-291.

Hira, Z.M. and D.F. Gillies, 2015. A review of feature selection and feature extraction methods applied on microarray data. Adv. Bioinf., 2015: 1-13.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology Control and Artificial Intelligence. U Michigan Press, Michigan, Pages: 183.

Huang, C.L. and C.J. Wang, 2006. A GA-based feature selection and parameters optimizationfor support vector machines. Expert Syst. Applic., 31: 231-240.

Kumar, P.G., T.A.A. Victoire, P. Renukadevi and D. Devaraj, 2012. Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm. Expert Syst. Appl., 39: 1811-1821.

Lee, C.P., W.S. Lin, Y.M. Chen and B.J. Kuo, 2011. Gene selection and sample classification on microarray data based on adaptive genetic algorithm-k-nearest neighbor method. Expert Syst. Appl., 38: 4661-4667.

Li, L., C.R. Weinberg, T.A. Darden and L.G. Pedersen, 2001. Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA-KNN method. Bioinf., 17: 1131-1142.

Li, L., T.A. Darden, C.R. Weingberg, A.J. Levine and L.G. Pedersen, 2001. Gene assessment and sample classification for gene expression data using a genetic algorithm-k-nearest neighbor method. Comb. Chem.. High Throughput Screening, 4: 727-739.

Li, W., 2006. The-more-the-better and the-less-the-better. Bioinf., 22: 2187-2188.

Liu, B., B. Fang, X. Liu, J. Chen and Z. Huang *et al.*, 2013. Large margin subspace learning for feature selection. Pattern Recognit., 46: 2798-2806.

Liu, D., T. Shi, J.A. DiDonato, J.D. Carpten and J. Zhu *et al*., 2004. Application of genetic algorithm-k-nearest neighbor method to the classification of renal cell carcinoma. Proceedings of the 2004 IEEE Conference on Computational Systems Bioinformatics (CSB), August 19-19, 2004, IEEE, New York, USA., ISBN: 0-7695-2194-0, pp: 558-559.

Nguyen, D.V. and D.M. Rocke, 2002. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinf., 18: 1216-1226.

Peng, H., F. Long and C. Ding, 2005. Feature selection based on mutual information criteria of max-dependency max-relevance and min-redundancy. IEEE. Trans. Pattern Anal. Mach. Intell., 27: 1226-1238.

Perez-Diez, A., A. Morgun and N. Shulzhenko, 2007. Microarrays for cancer diagnosis and classification. Adv. Exp. Med. Biol., 593: 74-85.

Pierna, J.A., V. Baeten, A.M. Renier, R.P. Cogdill and P. Dardenne, 2004. Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds. J. Chemom., 18: 341-349.

Rocha, M., R. Mendes, P. Maia, D.G. Pena and F.F. Riverola, 2007. A platform for the selection of genes in DNA microarraydata using evolutionary algorithms. Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, July 7-11, 2007, ACM, London, England, ISBN: 978-1-59593-697-4, pp: 415-423.

Shang, C. and Q. Shen, 2005. Aiding classification of gene expression data with feature selection: A comparative study. Int. J. Comput. Intell. Res., 1: 68-76.

Shen, Q., W.M. Shi and W. Kong, 2008. Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. Comput. Biol. Chem., 32: 53-60.

Srinivas, M. and L.M. Patnaik, 1994. Adaptive probabilities of crossover and mutation in genetic algorithms. IEEE Trans. Syst. Man Cybernet., 24: 656-667.

Sun, Y., S. Todorovic and S. Goodison, 2010. Local-learning-based feature selection for high-dimensional data analysis. IEEE Trans. Pattern Anal. Machine Intell., 32: 1610-1626.

Sungheetha, A. and J. Suganthi, 2013. An efficient clustering-classification method in an information gain NRGA-KNN algorithm for feature election of micro array data. Life Sci. J., 10: 691-700.

Tan, Y., L. Shi, W. Tong, G.G. Hwang and C. Wang, 2004. Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. Comput. Boil. Chem., 28: 235-243.

Tse, G.T. and S.K. Tso, 1993. Refinement of conventional PSS design in multimachine system by modal analysis. IEEE Trans. Power Syst., 8: 598-605.

Valdivia, M.T.M., M.C.D. Galiano, A.M. Raez and L.L.A. Urena, 2008. Using information gain to improve multi-modal information retrieval systems. Inf. Process. Manage., 44: 1146-1158.

Vanneschi, L., A. Farinaccio, G. Mauri, M. Antoniotti and P. Provero *et al*., 2011. A comparison of machine learning techniques for survival prediction in breast cancer. Bio. Data Mining, 4: 1-12.

Wang, A., N. An, G. Chen, L. Li and G. Alterovitz, 2015. Improving PLS-RFE based gene selection for microarray data classification. Comput. Boil. Med., 62: 14-24.

Yang, C.H., L.Y. Chuang and C.H. Yang, 2010. IG-GA: A hybrid filter-wrapper method for feature selection of microarray data. J. Med. Biol. Eng., 30: 23-28.

Zhao, Z., L. Wang, H. Liu and J. Ye, 2011. On similarity preserving feature selection. IEEE Trans. Knowl. Data Eng., 25: 619-632.

Zhou, W. and J.A. Dickerson, 2014. A novel class dependent feature selection method for cancer biomarker discovery. Comput. Boil. Med., 47: 66-75.