genes

MDPI

Article

The Cross-Entropy Based Multi-Filter Ensemble Method for Gene Selection

Yingqiang Sun [1], Chengbo Lu [2] and Xiaobo Li [2,*]

[1] School of Information Science and Engineering, Ningbo University, Ningbo 315000, China; 18363623303@163.com
[2] College of Engineering, Lishui University, Lishui 323000, China; lu.chengbo@aliyun.com
* Correspondence: oboaixil@126.com; Tel.: +86-578-227-1231

Received: 12 March 2018; Accepted: 2 May 2018; Published: 10 May 2018

Abstract: The gene expression profile has the characteristics of a high dimension, low sample, and continuous type, and it is a great challenge to use gene expression profile data for the classification of tumor samples. This paper proposes a cross-entropy based multi-filter ensemble (CEMFE) method for microarray data classification. Firstly, multiple filters are used to select the microarray data in order to obtain a plurality of the pre-selected feature subsets with a different classification ability. The top N genes with the highest rank of each subset are integrated so as to form a new data set. Secondly, the cross-entropy algorithm is used to remove the redundant data in the data set. Finally, the wrapper method, which is based on forward feature selection, is used to select the best feature subset. The experimental results show that the proposed method is more efficient than other gene selection methods and that it can achieve a higher classification accuracy under fewer characteristic genes.

Keywords: cross-entropy; multi-filter; gene expression profile; ensemble method; gene selection

1. Introduction

With the promotion of large-scale gene expression profiles, DNA chips can be used to obtain the expression level of thousands of genes in tissue samples, at the same time, in one experiment. The accurate classification of a tumor subtype at the molecular level is of great significance to the diagnosis and treatment of the tumor. The tumor gene expression profile data usually has a small sample size and high-dimensional feature space [1–4]. There is plenty of redundancy and noise data in the original data set, and thus the use of the feature selection method for classification can not only reduce the computational time, but it can also improve the classification accuracy [5,6]. Each sample in the data set records the level of expression of all of the measurable genes in the tissue sample, whereas only a few genes are actually related to the sample classification. Knowing how to select a group of genes that are critical to the classification of the sample is a key factor in establishing an effective classification model [7].

The gene selection consists of selecting a subset of genes from all of the attributes of the gene expression profile data [8], and the obtained genes have a strong ability to recognize the disease [9]. There are, in general, two approaches to gene selection, namely, filter [10] and wrappers [11]. The filter approach is based on the characteristics of the data itself for feature selection, and it does not depend on the classification algorithm to predict the selected subset [12,13]. The filter methods can be divided into two groups [14–17], namely, univariate and multivariate. The univariate methods measure the relationship of a single feature, with respect to a single evaluation criterion. In these methods, the dependencies between features play no role in the feature selection process. Methods such as the signal-to-noise ratio (SNR) [18], t-statistics (TS) [19], F-test (FT) [20], and Pearson correlation coefficient (PC) [21] have been shown to be effective for measuring the discriminative power of genes.

Unlike the univariate filter methods, the multivariate methods also consider the relationship between the features. This difference makes the multivariate methods relatively slower than their univariate counterparts. Well-known multivariate filter methods include correlation based feature selection (CFS) [22], minimum redundancy maximum relevance (mRMR) [23], and fast correlation based filter (FCBF) [24]. The filter approach is entirely based on the individual vector data for the feature selection, and the final subset evaluation criteria are independent of the classifier. The wrapper method selects a feature subset by some learning algorithms [25–27], and then evaluates it through the classifiers [28]. The evaluated data set can be selected and evaluated again, until the optimal feature subset is selected. Different learning methods have different feature subset evaluation criteria, some are based on an intelligent learning algorithm, some are based on a biological significance, and the others are based on its search space. However, the wrapper method is computationally slow and expensive [29]. In summary, the wrapper feature selection methods mainly select the final feature subset with the evaluation criteria of each learning algorithm.

The cross-entropy method is a new type of stochastic optimization algorithm that has emerged in recent years. It was first used to simulate low probability events and, later, it was extended to solve optimization problems [30–33]. The method has simple control parameters and a strong robustness. The cross-entropy is gradually applied to the field of bioinformatics, with its unique advantages, and has achieved some good results in the selection of the tumor feature genes. Su et al. [34] utilizes cross-entropy methods to deal with genes and gene pairs selection questions. Lin et al. [35] uses a cross-entropy Monte Carlo method for messenger RNA (mRNA) and microRNA studies problem. Bala et al. [36] uses mutual information in order to sort gene sets and uses the cross-entropy method to select the feature genes in descending order of the ranking results. In addition, this method is also applied to some practical problems in continuous multi-objective optimization and machine learning.

It is well known that many feature selection methods are very sensitive to data perturbations and lead to an instability of the selected classification models [37–39]. In order to select accurate classification subsets, many researchers have proposed the ensemble feature selection algorithm [40]. It is carried out by means of weighting several weak or base classifiers, and by combining them in order to obtain a classifier that outperforms each of the other classifiers. The ensemble methods can not only improve the classification accuracy and they can but also achieve the purpose of reducing dimension.

In view of the characteristics of the tumor gene expression data, and in order to obtain the highest possible sample classification rate and to reduce the time complexity of the algorithm, using as few as possible information genes, a cross-entropy based multi-filter ensemble (CEMFE) gene selection algorithm is proposed in this paper, which can select the feature subset with the best classification performance, using the wrapper method, which is based on the forward feature selection, with the accuracy as the criterion.

2. Materials and Methods

2.1. Datasets

To validate the performance of the proposed algorithm, five public biological datasets were used in this paper, including the Colon, Prostate, Leukemia, Lymphoma, and Lung datasets, which were downloaded [41]. Table 1 gives the detailed information of the five data sets.

2.2. Filtering Process

2.2.1. Signal-to-Noise Ratio

$$SNR = \frac{\left|\mu_{g^+} - \mu_{g^-}\right|}{\delta_{g^+} + \delta_{g^-}} \tag{1}$$

where *SNR* is the signal-to-noise ratio of gene g, μ_{g^+}, μ_{g^-}, is the mean value of expression level in different sample classes, and δ_{g^+}, δ_{g^-} is the standard deviation of the expression level. The signal-to-noise ratio of each gene is calculated, and the genes are sorted from high to low [42].

Table 1. Experimental datasets.

Dataset	Samples	Number of Samples		Classes
		Class 1	Class 2	
Colon	2000	40 (T)	22 (N)	2
Prostate	12,600	52 (T)	50 (N)	2
Leukemia	7129	25 (AML)	47 (ALL)	2
Lymphoma	7129	58 (DLBCL)	19 (FL)	2
Lung	12600	31 (MPM)	150 (ADCA)	2

T: tumor; DLBCL: diffuse large B-cell lymphoma; N: normal; FL: follicular lymphoma; AML: acute myeloid leukemia; ALL: acute lymphoblastic leukemia; MPM: malignant pleural mesothelioma; ADCA: adenocarcinoma.

2.2.2. *t*-Statistic

$$t_j = \frac{\overline{x_{2j}} - \overline{x_{1j}}}{\sqrt{{s_{1j}}^2/n_1 + {s_{2j}}^2/n_2}} \tag{2}$$

$\overline{x_{1j}}$ and $\overline{x_{2j}}$ is the mean value of the feature j in the two different sample classes. ${s_{1j}}^2$ and ${s_{2j}}^2$ represent the variance of feature j in the two different categories of samples. The larger the value that was calculated by the Equation (2), the greater the difference in the expression of the feature j was in the two categories [43].

2.2.3. Pearson Correlation Coefficient

$$PC = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}} \tag{3}$$

where PC is the Pearson correlation coefficient, and represents the values corresponding to the class and denotes the average of the features and categories. The greater the value of the PC, the greater was the relevance of the feature to the category [44].

2.2.4. Combination of Filtered Genes

In this work, the three most commonly filters (SNR, TS, and PC) were used in order to select the genes. The three filters could obtain three orderly data sets, and each data set was sorted by value, from high to low. The top N genes of each data set were selected so as to integrate a new gene set.

Suppose the number of data set was S, and the number of filter was L. We obtained the gene set F, which contained the top N genes in L data sets, as follows (where we supposed that S was larger than N):

(1) Suppose $G = \{g_1, g_2, \ldots, g_S\}$, and $F = \phi$;
(2) Use the filter FT_i to calculate the statistical scores and rank them, where $i \in \{1, 2, \ldots, L\}$;
(3) Select the N genes with the top ranking score in each list, add N into F, and delete the N genes from G;
(4) Take the union of the L filtered lists, which consolidates the overlapping genes and reduces the size of the combined list F of the filtered genes;
(5) Repeat steps (2)–(4) until all of the top N genes are added to F and there are no duplication genes.

2.3. Cross-Entropy Method

The cross-entropy method [45] was a new optimization method that was proposed by Professor Reuven Y. Rubinstein in 1998 [46]. It was first used to simulate the low probability events and it was later extended in order to solve optimization problems. In recent years, the cross-entropy method was widely applied to the solution of many combinatorial optimization problems, and a solution algorithm was designed for the different application areas [47,48]. In general, the filtered gene subset F would contain a large number of redundant genes, which would affect the classification accuracy and the robustness of the classification model. Therefore, deleting the redundant genes in F played a key role in the selection of the best feature subset.

Several methods were proposed to measure the dependency of the variables [49], such as mutual information, entropy, and cross-entropy. The notion of cross-entropy was used in this work to compute the redundancy of a feature set [36,50].

The cross-entropy of $f(X)$ and $g(X)$, denoted by $D(f(X), g(X))$, is as follows:

$$D(f(X), g(X)) = \sum f(X) \log \frac{f(X)}{g(X)} \tag{4}$$

If $f(X) = p(x_1, x_2, \ldots, x_n)$, and $g(X) = p(x_1)p(x_2)\ldots p(x_n)$, the cross-entropy $D(f(X), g(X))$ can be written as D_n for short.

$$D_n = \int \ldots \int p(x_1, \ldots, x_n) \log[p(x_1, \ldots, x_n)/p(x_1)\ldots p(x_n)] \tag{5}$$

If $x_1 \ldots x_n$ are independent, then $D_n = 0$. Otherwise, it is as follows:

$$D_n = \int \ldots \int p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n) - \sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{6}$$

Let $H = -\int \ldots \int p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n)$, $H_i = \sum_{i=1}^{n} p(x_i) \log p(x_i)$, the above formula can be simplified as follows:

$$D_n = -H + \sum_{i=1}^{n} H_i \tag{7}$$

Since $H_i \leq H(i = 1, \ldots, n)$, we have the following:

$$H_1 + H_2 + \ldots + H_n \leq nH \tag{8}$$

For Equations (7) and (8), we have the following: $D_n = H_1 + H_2 + \ldots + H_n - H \leq (n-1)H$. Therefore, D_n can be normalized as follows:

$$\overline{D_n} = (H_1 + H_2 + \ldots + H_n - H)/((n-1)H) \tag{9}$$

where $0 \leq \overline{D_n} \leq 1$.

In general, $\overline{D_n}$ measures the dependency of the n variables. The larger the value of $\overline{D_n}$ was, the more dependent the variables were. In order to select the independent features, the threshold of independence should have been set. If the threshold of independence was T, and a feature set had $\overline{D_n} \leq T$, then the features in this feature set were considered to be independent.

2.4. Calculation of Redundancy

The gene set F, which was integrated after the filter process, was then calculated by cross-entropy, and the new gene set was obtained by removing the redundancy genes. Let $F = \{g_1, g_2, \ldots, g_m\}$ be a feature set, which contained the m genes and their values of closeness with a class, we eliminated the redundant features and obtained the non-redundant feature set G as follows:

(1) Set the threshold of independence be T, and $G = \phi$;

(2) Use $\overline{D_n}(G, g_j)$ to calculate the cross-entropy between two genes, where $g_j \in F$;

(3) If $\overline{D_n}(G, g_j) < T$, then $G = G \cup \{g_j\}$, $F = F - \{g_j\}$, and go to step (4);

(4) If $\overline{D_n}(G, g_j) \geq T$, then $F = F - \{g_j\}$, and go to step (4).

(5) Repeat (2)–(3), until $F = \phi$.

2.5. Selection of Optimal Subset

To obtain the best gene set R, the wrapper method was used, based on the forward feature selection to select the optimal subset, with the largest classification accuracy as the criterion.

(1) Initialization $R = \Phi$;

(2) For each $x_i \in G$, calculate the classification accuracy for classifier M;

(3) Select a subset of the genes x_k with the highest accuracy h, $R = R \cup x_k$ $G = G - \{x_k\}$;

(4) For each $x_i \in G$, calculate the classification of $R \cup \{x_i\}$, which is referred to as h';

(5) If $h' > h$, then $R = R \cup \{x_k\}$, $G = G - \{x_k\}$;

(6) Repeat (4)–(5), until the accuracy is 100 or G is null.

2.6. Flowchart of CEMFE Method

In this paper, we used three of the most commonly used filters (SNR, TS, and PC) in order to select the genes from the microarray datasets and to integrate the top N genes in the subset of each filter, in order to form a new gene set F. The cross-entropy algorithm was used to get the gene set G from F. Finally, in order to get the best subset, the wrapper method was used, which was based on the forward feature selection with the accuracy as the criterion. The model that was proposed in this paper was based on the CEMFE gene selection method, shown in Figure 1:

Our proposed algorithm can be described as follows:

```
Input: data set $S$, number of filter $L$, number of union filtered gene $P$, number of genes subset ($G$) $Q$,
classifier $M$
Output: optimal feature subset $R$
   For $i = 1$ to $L$ do
    $S_i$ = use the filter $FT_i$ calculate the statistical scores and rank it
    $m_i$ = select $m$ genes with top ranking score in each list
    End of For
        F /*the union of the list of genes*/
    Initialization: $G = \phi$
    For $j = 1$ to $P$ do
    Calculate $\overline{D_n}\left(G, x_j\right)$ /*For all $x_j \in F$*/
    If $\overline{D_n}\left(G, x_j\right) < T, G = G \cup \left\{x_j\right\}, F = F - \left\{x_j\right\}$
    End of For
    Return G
    Initialization: $R = \phi$/*optimal feature subset*/
    $x_k = \max Class\_Acc, R = R \cup \{x_k\}$
    For $k = 2$ to $Q$ do
    $newClass\_Acc$ = calculate classification accuracy of $R \cup \{x_k\}$
    If $newClass\_Acc > \max Class\_Acc$
    $\max Class\_Acc = newClass\_Acc, R = R \cup \{x_k\}$
    End of For
Return $R$, $\max Class\_Acc$
```

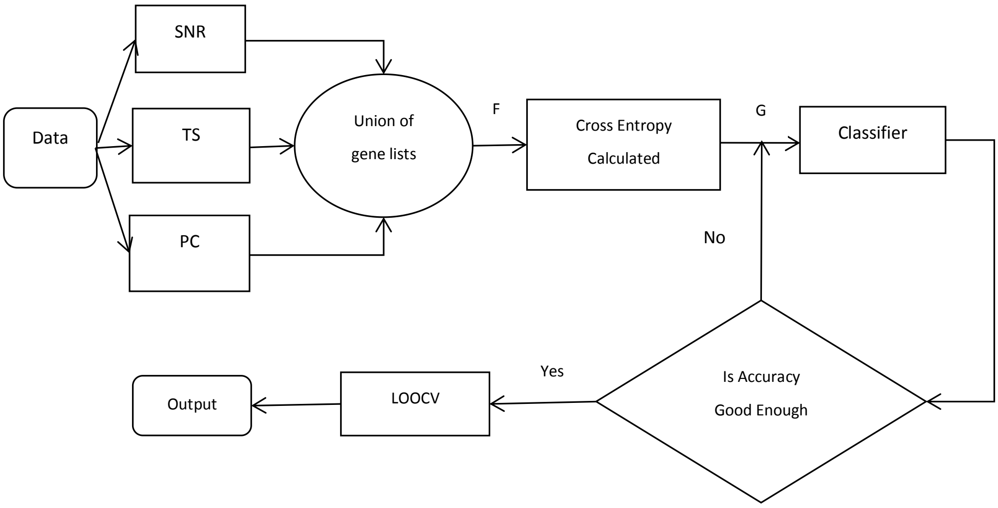

Figure 1. Flowchart of the cross-entropy multi-filter ensemble (CEMFE) algorithm. TS: *t*-statistic; SNR: signal-to-noise ratio; PC: Pearson correlation coefficient; LOOCV: leave one out cross validation.

3. Results and Discussion

The experiments were performed on a Windows 7, 2.2 GHz 8 G personal computer All of the experiments were implemented in matlab R2015b and weka 3.8.0 [51], and three kinds of classification models, namely, the Naive Bayes, Support Vector Machine (SVM), and k-nearest neighbor, were constructed, in which the value of k was 3, and the kernel function of the SVM was set as the linear kernel function. All of the experiments were used in the K-fold cross validation method, where K was taken as 10.

3.1. Results on Microarray Data

The experimental datasets were first normalized using the Z-score. Then, the gene expression profile datasets were filtered, using three filters, namely, SNR, TS, and PC. Each filter produced an ordered list of genes, and the top N (N = 50) genes, with the highest rankings in each list (with a significant score on the classification), were merged to form a new list of genes. Taking the union of the three lists consolidated the overlapping genes and reduced the size of the combined list of the filtered genes. The cross-entropy algorithm was used for the redundant computing, the value of T was varied from 0.1 to 0.9, with a step size of 0.1. It was observed that the non-redundant gene sets were the same for the values between 0.4 and 0.8, therefore, we took T = 0.5 during the process of eliminating the redundant feature genes, (i.e., all the genes with $\overline{D_n}$ greater than 0.5 were rejected as dependent genes). Finally, the best feature subset was obtained using the wrapper-based forward feature selection method, and the accuracy was used as the criterion in the selection process. In learning the classification algorithm, the support vector machine (SVM) could avoid a dimensionality disaster and had better robustness [52], the training speed of Naive Bayesian (NB) was faster, and the k-nearest neighbor (KNN) was easy to implement, with no need to estimate the parameters and no need for training. In order to validate the classification model of our proposed algorithm, the three kinds of learning algorithms, including NB, SVM, and KNN, were used in order to verify their classification performance. Table 2 shows the number of feature genes and the best classification accuracy that was obtained by the different algorithms, so as to achieve the best classification accuracy. The CEMFE represents the method that was proposed in this paper; the signal–noise ration and cross-entropy (SNRCE) meant that only the signal-to-noise ratio and the cross-entropy method were used; the *t*-statistic and cross-entropy method (TSCE) meant that only the *t*-statistic and cross-entropy method were used; the Pearson correlation coefficient and cross-entropy method (PCCE) represented that only the Pearson correlation coefficient and the cross-entropy method were used.

Table 2. Experimental contrast of all of the kinds of algorithms on different data sets, feature gene number, and the best classification accuracy.

Dataset	Classifier	CEMFE	SNRCE	TSCE	PCCE
Colon	SVM	93.55 (23)	90.32 (7)	90.32 (18)	93.55 (33)
	KNN	96.77 (14)	88.71 (17)	83.87 (13)	90.32 (7)
	NB	96.77 (9)	88.71 (17)	85.48 (6)	91.91 (21)
Prostate	SVM	97.10 (17)	88.24 (11)	97.10 (22)	91.18 (7)
	KNN	98.04 (9)	94.12 (26)	90.20 (10)	98.04 (15)
	NB	96.10 (9)	89.22 (13)	93.14 (22)	92.16 (27)
Leukemia	SVM	97.22 (6)	95.83 (17)	90.28 (17)	95.83 (35)
	KNN	98.61 (7)	89.06 (28)	88.89 (23)	93.06 (8)
	NB	100 (12)	83.33 (26)	96.88 (19)	94.44 (33)
Lymphoma	SVM	100 (26)	88.31 (34)	96.10 (66)	84.42 (36)
	KNN	98.70 (16)	93.51 (7)	79.22 (14)	97.40 (41)
	NB	98.70 (18)	94.81 (9)	96.10 (14)	80.52 (22)
Lung	SVM	100 (4)	98.34 (24)	100 (36)	99.45 (33)
	KNN	100 (3)	100 (21)	100 (40)	100 (18)
	NB	98.90 (9)	96.13 (17)	98.90 (23)	98.90 (41)

SVM: support vector machine; KNN: k-nearest neighbor; NB: Naive Bayesian; CEMFE: cross-entropy based multi-filter ensemble; SNRCE: signal–noise ration and cross-entropy; TSCE: *t*-statistic and cross-entropy method; PCCE: Pearson correlation coefficient and cross-entropy method.

From Table 2, we can see that for the different data sets, the different feature gene selection methods showed a different classification performance on five different classifiers. Compared with the other methods, it was shown that the accuracy of the CEMFE method was relatively high, and the number of the selected feature genes were fewer. In the colon data set, the accuracy of the CEMFE in the KNN and the NB were 96.77%, which were significantly higher than the other methods. The minimum number of genes, nine, were obtained on the NB classifier. In prostate dataset, a maximum classification accuracy of 98.04% was achieved with nine genes in the KNN classifier, using CEMFE. In the leukemia dataset, the maximum classification accuracy of 100% was achieved with 12 genes in the NB classifier, using CEMFE. The classification accuracies obtained using the CEMFE algorithm in the SVM and the KNN classifiers were also much higher than the other algorithms. In the lymphoma dataset, the maximum classification accuracy of 100% was achieved in the SVM classifier, using the CEMFE. The same classification accuracy of 98.70% was achieved in the KNN and NB classifiers. For the lung dataset, the maximum classification accuracy of 100% was achieved in the SVM and KNN classifier, using the gene that was selected by our method, CEMFE. The number of genes that were selected by CEMFE were significantly less, in comparison with the other three kinds of classification algorithms. The best result was obtained for the KNN, using only three genes. Therefore, the CEMFE algorithm that was proposed in this paper could obtain a subset of the best feature genes with a high correlation and low redundancy as a whole and could effectively improve the accuracy of the feature gene classification algorithm.

Figure 2 shows the results of the classification accuracy that was obtained by the different methods, as the genes were added one by one in the prostate dataset. It was observed that the classification accuracy that was obtained by our algorithm was much better than the SNRCE, TSCE, and PCCE, with the same number of genes for all of the classifiers. Similar results were also observed in the other datasets.

In order to further verify the efficiency and stability of the CEMFE classification algorithm that was proposed in this paper, we compared our methods (CEMFE) with two model-free gene selection methods, namely, mRMR [23] and FCBF [24]. The reason for choosing them was that they were typical and popular gene selection algorithms. The FCBF [24] measured the relevance between the genes using symmetric uncertainty and eliminated the irrelevant genes by virtue of an approximate Markov blanket. In mRMR [23], only those genes that might have brought more relevance to the class and less redundancy to the selected genes, at the same time, would be selected. The Naive Bayes (NB) and k-nearest-neighbor (KNN), were used in order to build the classifiers on the selected gene subsets. For KNN, K = 3 and its distance was calculated by the Euclidean formula, in our experiments. The k-fold cross-validation was used to evaluate the performance of the experiment, and K = 10.

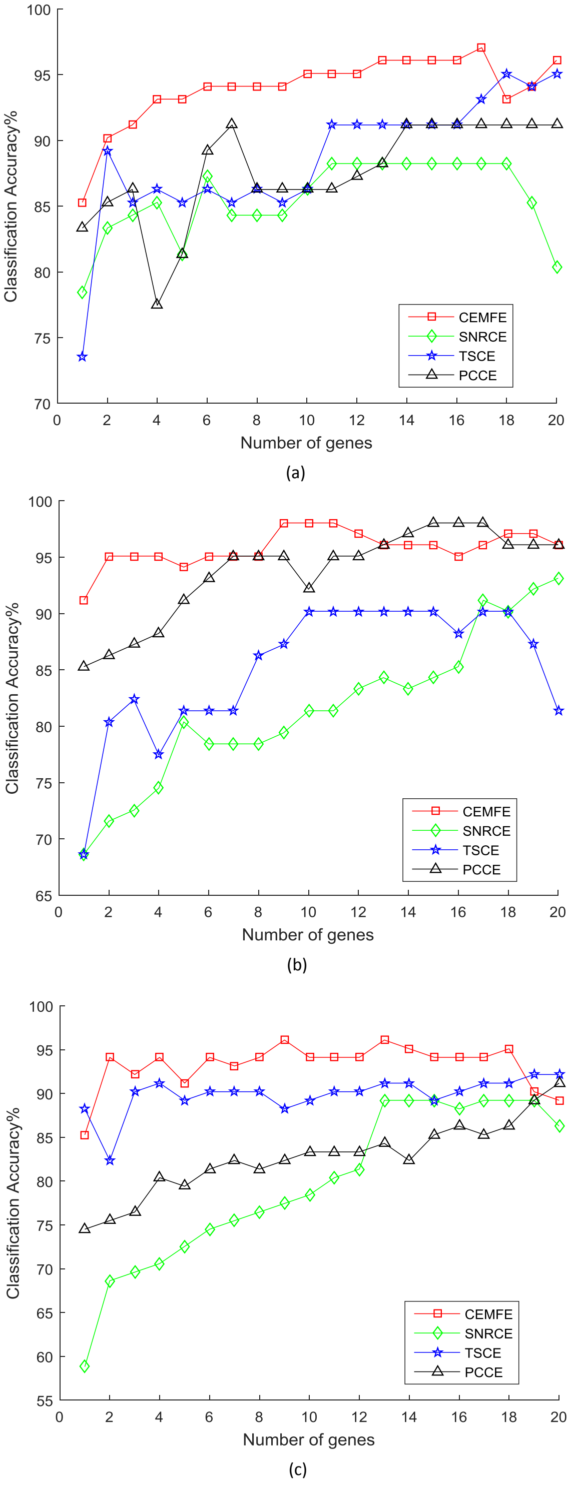

Figure 2. Classification accuracy vs. number of genes for the prostate dataset, using the (**a**) Support Vector Machine, (**b**) k-nearest neighbor, and (**c**) Naive Bayesian.

Table 3 summarizes the best classification accuracy of NB and KNN, using three gene selectors. In the colon dataset, the accuracy of the CEMFE algorithm in both classifiers was 96.77%, which was much higher than that of the FCBF and mRMR algorithms. In the prostate dataset, the accuracy of the CEMFE algorithm in the NB classifier was slightly lower than the other two algorithms, but the classification accuracy was better for the KNN classifier, and similarly, for the leukemia dataset. In the lymphoma and lung datasets, the classification accuracy of the CEMFE algorithm on two classifiers was obviously higher than that of the FCBF and mRMR algorithms. Overall, the average classification accuracy of the CEMFE algorithm in all the five datasets was higher than the accuracy of the other two classification algorithms. In other words, the CEMFE algorithm that was proposed in this paper could obtain a highly relevant subset of the feature genes and improve the classification performance.

Table 3. The best classification accuracy of the CEMFE, FCBF, and mRMR algorithms in different data sets.

Dataset	NB			KNN		
	FCBF	mRMR	CEMFE	FCBF	mRMR	CEMFE
Colon	91.94	88.79	96.77	88.71	77.42	96.77
Prostate	97.06	98.04	96.08	97.06	97.06	98.04
Leukemia	100	100	100	100	100	98.61
Lymphoma	93.51	94.81	98.70	93.51	97.40	98.70
Lung	86.67	99.13	100	83.33	96.13	98.90
Average	93.84	96.15	98.31	92.52	93.60	98.20

mRMR: minimum redundancy maximum relevance; FCBF: fast correlation based filter.

3.2. Discussion

To study the effect of varying the number of genes that were selected by each filter, we repeated all the experiments of Section 3 for each dataset, with N = 100 and 200, in order to ascertain whether N = 50 was a reasonable choice. The results for the different values of N are summarized in Table 4. Where the average number of genes and the average of performance, to the average of the total number of genes and the best classification accuracy that were obtained by the k-fold cross-validation on each SVM, KNN, and NB classifier are shown. For all the datasets, the best accuracy could be achieved with N = 50, except for lymphoma, where the accuracy had decreased from the case of N = 100 to that of N = 200, as a result of the increase of one misclassification. Theoretically, the use of a larger gene subset should have always increased the accuracy. However, the CEMFE algorithm was not guaranteed to converge to the global optimal solution, and the results of Table 4 suggested that the use of an unnecessarily large gene set might have caused the algorithm to be trapped at a local minimum, as was the case when N was increased from 100 to 200 for the leukemia, lymphoma, and lung data sets. Hence, for the data sets that were under consideration, N = 50 would be the best choice for the number of genes that were to be retained by each filter.

Table 4. Classification accuracies variation on different number of genes selected by each filter.

N/Data Set		Colon	Prostate	Leukemia	Lymphoma	Lung
50	Avg. no. of genes	15.3	12.3	8.3	20	5.3
	Avg. performance	95.70	97.08	98.61	99.13	99.63
100	Avg. no. of genes	10.6	12.3	11.6	20	5.3
	Avg. performance	93.55	95.93	93.06	100	99.63
200	Avg. no. of genes	10.6	12.3	13.3	22.3	3.6
	Avg. performance	91.94	95.93	90.28	98.70	98.90

N: number of genes selected; Avg: average; no: number.

4. Conclusions

This study aimed to select an optimal subset of the features from the high dimensional and small sample gene expression datasets for the classification of cancer genes. For this purpose, we proposed a cross-entropy based multi-filter ensemble method (CEMFE), where multi-filter ensemble algorithm is used to classify the microarray data, and the smallest feature subset that is associated with cancer classification is selected. Firstly, the multi-filter is employed to select a set of relevant genes, and cross entropy is used to determine the independent genes, which provides a set of independent and relevant genes, and reduces the size of the gene set significantly. Secondly, the gene subset is selected using the wrapper method, based on the forward feature selection. Finally, the final gene subset, with the highest classification accuracy, is obtained. In the above process, the unrelated and redundant genes in the original gene set were screened, so that the obtained genes were closely related to the class labels and were independent of each other, and the best feature subset was selected, based on the criterion of the accuracy rate. The experimental results show that the method that has been proposed in this paper can not only obtain a high accuracy, but also, the number of genes that are obtained is less than that of the other methods.

Author Contributions: All authors reviewed and approved the final version of the manuscript. The literature search was done by Y.S. and C.L.; Y.S. and X.L. conceived and designed the experiments; Y.S. and C.L. performed the experiments; Y.S. analyzed the data; and Y.S. and X.L. wrote the paper.

Acknowledgments: The work was supported by a grant from the National Natural Science Foundation of China (61373057), a grant from the Zhejiang Provincial Natural Science Foundation of China (LY17F020003, LY18F030003), and a grant from the Zhejiang Provincial Key Laboratory of Digital Design and Intelligent Manufacturing of Characteristic Cultural and Creative Products (2016E10007).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rakkeitwinai, S.; Lursinsap, C.; Aporntewan, C.; Mutirangura, A. New feature selection for gene expression classification based on degree of class overlap in principle dimensions. *Comput. Biol. Med.* **2015**, *64*, 292–298. [CrossRef] [PubMed]
2. Zhou, W.; Dickerson, J.A. A novel class dependent feature selection method for cancer biomarker discovery. *Comput. Biol. Med.* **2014**, *47*, 66–75. [CrossRef] [PubMed]
3. Zhang, X.; Song, Q.; Wang, G.; Zhang, K.; He, L.; Jia, X. A dissimilarity-based imbalance data classification algorithm. *Appl. Intell.* **2015**, *42*, 544–565. [CrossRef]
4. Xiong, H.; Zhang, Y.; Chen, X.W.; Yu, J. Cross-platform microarray data integration using the normalized linear transform. *Int. J. Data Min. Bioinform.* **2010**, *4*, 142–157. [CrossRef] [PubMed]
5. Kabir, M.M.; Shahjahan, M.; Murase, K. A new local search based hybrid genetic algorithm for feature selection. *Neurocomputing* **2011**, *74*, 2914–2928. [CrossRef]
6. Pugalendhi, G.; Vijayakumar, A.; Kim, K.J. A new data-driven method for microarray data classification. *Inter. J. Data Min. Bioinform.* **2016**, *15*, 101–124. [CrossRef]
7. Marafino, B.J.; John Boscardin, W.; Dudley, R.A. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J. Biomed. Inform.* **2015**, *54*, 114–120. [CrossRef] [PubMed]
8. You, W.; Yang, Z.; Yuan, M.; Ji, G. TotalPLS: local dimension reduction for multicategory microarray data. *IEEE Trans. Hum. Mach. Syst.* **2014**, *44*, 125–138.
9. Magendiran, N.; Selvarajan, S. Substantial Gene Selection in Disease Prediction based on Cluster Centre Initialization Algorithm. *Indian J.* **2016**, *6*, 258. [CrossRef]
10. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef] [PubMed]
11. Kohavi, R.; John, G.H. *Wrappers for Feature Subset Selection*; Elsevier Science Publishers Ltd.: New York, NY, USA, 1997.
12. Kamkar, I.; Gupta, S.K.; Phung, D.; Venkatesh, S. Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso. *J. Biomed. Inform.* **2015**, *53*, 277. [CrossRef] [PubMed]

13. Liu, M.; Zhang, D. Feature selection with effective distance. *Neurocomputing* **2016**, *215*, 100–109. [CrossRef]
14. Xu, J.; Li, T.; Sun, L.; Li, Y. Feature selection method based on signal-to-noise ratio and neighborhood rough set. *Data Acquis. Process.* **2015**, *30*, 973–981.
15. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. Data classification using an ensemble of filters. *Neurocomputing* **2014**, *135*, 13–20. [CrossRef]
16. Leung, Y.; Hung, Y. A Multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2010**, *7*, 108–117. [CrossRef] [PubMed]
17. Meyer, P.E.; Schretter, C.; Bontempi, G. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE J. Sel. Top. Signal Proc.* **2008**, *2*, 261–274. [CrossRef]
18. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 205–214. [CrossRef]
19. Speed, T. *Statistical Analysis of Gene Expression Microarray Data*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2003.
20. Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. In Proceedings of the IEEE Bioinformatics Conference, Stanford, CA, USA, 11–14 August 2003; pp. 523–528.
21. Leung, Y.Y.; Chang, C.Q.; Hung, Y.S.; Fung, P.C.W. Gene selection for brain cancer classification. In Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society Embs '06, New York, NY, USA, 30 August–3 September 2006; pp. 5846–5849.
22. Hall, M.A. *Correlation-Based Feature Selection for Machine Learning*; University of Waikato: Hamilton, New Zealand, 1999; p. 19.
23. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Comput. Soc.* **2005**, *8*, 1226–1238.
24. Yu, L.; Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the Twentieth International Conference on International Conference on Machine Learning, Washington, DC, USA, 21–24 August 2003; pp. 856–863.
25. Liu, J.; Zhou, H.B. Tumor classification based on gene microarray data and hybrid learning method. In Proceedings of the International Conference on Machine Learning and Cybernetics, Xi'an, China, 5 November 2003; Volume 4, pp. 2275–2280.
26. Shreem, S.S.; Abdullah, S.; Nazri, M.Z.A.; Alzaqebah, M. Hybridizing relief, mRMR filters and GA wrapper approaches for gene selection. *J. Theor. Appl. Inform. Technol.* **2013**, *46*, 1034–1039.
27. Brahim, A.B.; Limam, M. Robust ensemble feature selection for high dimensional data sets. In Proceedings of the International Conference on High Performance Computing and Simulation, Helsinki, Finland, 1–5 July 2013; pp. 151–157.
28. Inza, I.; Larrañaga, P.; Blanco, R.; Cerrolaza, A.J. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.* **2004**, *31*, 91–103. [CrossRef] [PubMed]
29. Tabakhi, S.; Moradi, P.; Akhlaghian, F. An unsupervised feature selection algorithm based on ant colony optimization. *Eng. Appl. Artif. Intell.* **2014**, *32*, 112–123. [CrossRef]
30. Choe, Y. Information criterion for minimum cross-entropy model selection. *arXiv*, **2017**, arXiv:1704.04315.
31. Rubinstein, R.Y.; Kroese, D.P. *The Cross-Entropy Method: A unified Approach to Combinatiorial Optimization, Monte-Carlo Simulation and Machine Learning*; Springer: New York, NY, USA, 2004; pp. 92–94.
32. Botev, Z.I.; Kroese, D.P.; Rubinstein, R.Y.; L'Ecuyer, P. The cross-entropy method for optimization. *Handb. Stat.* **2013**, *31*, 35–59.
33. Benham, T.; Duan, Q.; Kroese, D.P.; Liquet, B. CEoptim: cross-entropy R package for optimization. *arXiv*, **2015**, arXiv:1503.01842.
34. Su, Y.; Li, Y.; Zhang, Z.; Pan, L. Feature identification for phenotypic classification based on genes and gene pairs. *Curr. Bioinform.* **2017**, *12*. [CrossRef]
35. Lin, S.; Ding, J. Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA Studies. *Biometrics* **2009**, *65*, 9. [CrossRef] [PubMed]
36. Bala, R.; Agrawal, R.K. Mutual information and cross entropy framework to determine relevant gene subset for cancer classification. *Informatica* **2011**, *35*, 375–382.
37. Li, X.; Lu, H.; Wang, M. A hybrid gene selection method for multi-category tumor classification using microarray data. *Int. J. Bioautomation* **2013**, *17*, 249–258.

38. Utkin, L.V.; Zhuk, Y.A.; Chekh, A.I. An ensemble-based feature selection algorithm using combination of support vector machine and filter methods for solving classification problems. *Eur. J. Technol. Des.* **2013**, *1*, 70–76. [CrossRef]
39. Duan, K.B.; Rajapakse, J.C.; Wang, H.; Azuaje, F. Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Trans. Nanobiosci.* **2005**, *4*, 228–234. [CrossRef]
40. Abeel, T.; Helleputte, T.; van de Peer, Y.; Dupont, P.; Saeys, Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* **2010**, *26*, 392–398. [CrossRef] [PubMed]
41. Microarray Datasets. Available online: http://csse.szu.edu.cn/staff/zhuzx/Datasets.html (accessed on 11 July 2017).
42. Hengpraprohm, S. GA-Based Classifier with SNR weighted features for cancer microarray data classification. *Int. J. Signal Proc. Syst.* **2013**, *1*, 29–33. [CrossRef]
43. Li, X.; Peng, S.; Chen, J.; Lü, B.; Zhang, H.; Lai, M. SVM-T-RFE: A novel gene selection algorithm for identifying metastasis-related genes in colorectal cancer using gene expression profiles. *Biochem. Biophys. Res. Commun.* **2012**, *419*, 148–153. [CrossRef] [PubMed]
44. Benesty, J.; Chen, J.; Huang, Y. On the importance of the Pearson correlation coefficient in noise reduction. *IEEE Trans. Audio Speech Lang. Proc.* **2008**, *16*, 757–765. [CrossRef]
45. Hui, K.P.; Bean, N.; Kraetzl, M.; Kroese, D.P. The Cross-Entropy method for network reliability estimation. *Ann. Oper. Res.* **2005**, *134*, 101. [CrossRef]
46. Rubinstein, R. The Cross-Entropy Method for Combinatorial and Continuous Optimization. *Methodol. Comput. Appl. Probab.* **1999**, *1*, 127–190. [CrossRef]
47. Chan, J.C.; Eisenstat, E. Marginal likelihood estimation with the cross-entropy method. *Econom. Rev.* **2015**, *34*, 256–285. [CrossRef]
48. Qi, X.; Liang, C.; Zhang, J. Generalized cross-entropy based group decision making with unknown expert and attribute weights under interval-valued intuitionistic fuzzy environment. *Comput. Ind. Eng.* **2015**, *79*, 52–64. [CrossRef]
49. Li, X.; Peng, S. Identification of metastasis-associated genes in colorectal cancer through an integrated genomic and transcriptomic analysis. *Chin. J. Cancer Res.* **2013**, *25*, 623–636. [PubMed]
50. Kapur, J.N.; Kesavan, H.K. Entropy optimization principles and Their Applications. *Water Sci. Technol. Libr.* **1992**, *9*, 3–20.
51. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]
52. Li, X.; Gong, X.; Peng, X.; Peng, S. SSiCP: a new SVM based Recursive Feature Elimination Algorithm for Multiclass Cancer Classification. *Int. J. Multimed. Ubiquituos Eng.* **2014**, *9*, 347–360. [CrossRef]