molecules

MDPI

Article

Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics

Xiaohui Lin *, Chao Li, Yanhui Zhang, Benzhe Su, Meng Fan and Hai Wei

School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China; lizimu@mail.dlut.edu.cn (C.L.); yanhui.zhang.123@foxmail.com (Y.Z.); benzhe.su.123@foxmail.com (B.S.); mengfan2351@163.com (M.F.); weihai_dlut@163.com (H.W.)

* Correspondence: datas@dlut.edu.cn

Received: 10 November 2017; Accepted: 16 December 2017; Published: 26 December 2017

Abstract: Feature selection is an important topic in bioinformatics. Defining informative features from complex high dimensional biological data is critical in disease study, drug development, etc. Support vector machine-recursive feature elimination (SVM-RFE) is an efficient feature selection technique that has shown its power in many applications. It ranks the features according to the recursive feature deletion sequence based on SVM. In this study, we propose a method, SVM-RFE-OA, which combines the classification accuracy rate and the average overlapping ratio of the samples to determine the number of features to be selected from the feature rank of SVM-RFE. Meanwhile, to measure the feature weights more accurately, we propose a modified SVM-RFE-OA (M-SVM-RFE-OA) algorithm that temporally screens out the samples lying in a heavy overlapping area in each iteration. The experiments on the eight public biological datasets show that the discriminative ability of the feature subset could be measured more accurately by combining the classification accuracy rate with the average overlapping degree of the samples compared with using the classification accuracy rate alone, and shielding the samples in the overlapping area made the calculation of the feature weights more stable and accurate. The methods proposed in this study can also be used with other RFE techniques to define potential biomarkers from big biological data.

Keywords: SVM-RFE; overlapping degree; feature selection

1. Introduction

Feature selection is one of the main data analysis techniques in data mining, which has shown its power in many applications, such as insulator detection [1], medicine study [2], and environmental science [3]. Especially for big data analysis, how to define meaningful information is a key issue.

Along with the quick development of the high throughput techniques, genomics, metabolomics, and proteomics have been widely applied in disease study, drug research, etc. One characteristic of omics data is summarized in the expression "high dimensions, small samples", since omics data usually have a large number of features but few samples. As genomics data, metabolomics data and proteomics data usually contain many features, it has been critical to accurately measure the feature importance and select the most discriminative feature subset. Puthiyedth et al. [4] presented a combinatorial optimization approach for integrated feature selection and applied it to analyzing the data about prostate cancer. They have identified potential novel prostate cancer associated pathways and genes. Christin et al. [5] studied six feature selection methods, analyzed their performance for liquid chromatography-mass spectrometry based proteomics and metabolomics biomarker discovery. Zou et al. [6,7] presented the sequence based feature selection technique and dimensionality reduction strategy to realize the prediction of protein. Lin et al. [8] studied the feature selection method based on the overlapping area and defined the discriminative features of liver disease from the metabolomics dataset.

Support vector machine (SVM) [9] is a popular and efficient classification technique and has been widely applied in many fields such as biological data processing [10]. SVM-recursive feature elimination (SVM-RFE) [11] is a feature selection algorithm based on SVM. While the SVM learning model is built, the weights of the features are also computed. SVM-RFE iteratively removes the features with the lowest weights. The removing sequence of the features represents the feature importance ranking [11,12]. SVM-RFE has been adopted in many applications, such as signal processing [13], genomics [11,12], proteomics [14] and metabolomics [15,16], due to its superiority. Also, many studies have been done on it to get a more powerful performance. Tang et al. [17] proposed a two-stage SVM-RFE. In the first stage, multiple SVM-RFEs with different parameters were applied to remove the noise and non-informative data; in the second stage, the final feature subset was selected by a fine SVM-RFE. Li et al. [18] combined SVM-RFE with the T-statistic to define the genes associated with CRC development or metastasis. mRMR-SVM [19] tries to select an important and non-redundant feature subset by means of SVM-RFE and mRMR. R-SVM [20] is also a recursive feature selection method based on SVM, which combines SVM weights and class means to evaluate feature discriminative abilities. There are also some studies on determining how many features with the low weights are removed in each iteration of SVM-RFE [21,22].

Basically, SVM-RFE ranks the features according to the feature deletion order during the iterations. The top ranked features which are removed in the last iteration of SVM-RFE are the most important, while the bottom ranked ones are the least informative and removed in the first iteration. For a specific application, it is not enough to obtain a feature importance ranking; it needs to determine how many top ranked features (such as genes and metabolites) should be selected. Thus, based on the selected features, we can study the disease phenotype and disease mechanism. In some studies, the top ranked features were selected according to a predetermined number [23,24]. In other studies, the top features that can induce a classifier with a "best" classification accuracy rate were selected [11,16].

It is not practical to specify the number of features to be selected in advance in some applications. However, it is well known that the feature subset selected should have a powerful discriminative ability. If a feature subset has a powerful discriminative ability, then the classifier based on it usually has a high prediction accuracy rate, and the different sample groups on the selected subspace should show different distributions with little overlapping areas. Hence, this study proposes a method, SVM-RFE-OA, which determines the number of features to be selected from the feature rank of SVM-RFE by combining the classification accuracy rate and the average overlapping ratio of the samples together. In addition, to weigh the features more accurately, this study also proposes a modified SVM-RFE-OA (M-SVM-RFE-OA) algorithm, which temporally screens out the samples lying in a heavily overlapping area in each iteration. The experiments on the eight public biological datasets show the validation of the two techniques proposed.

2. Methods

2.1. Overlapping Degree

Let $X = \{x_1, x_2, \ldots x_n\}$ be the dataset containing n samples, C be the class label set, $Label(x_i) \in C$ be the class label of sample $x_i \in X$. For a sample $x_i \in X$, the number of its neighbor samples that do not belong to the same class as x_i reflects whether it lies in an overlapping area [25,26]. If most of its neighbors do not belong to the same class as x_i, then x_i heavily mixes with the heterogeneous samples and locates in an overlapping area. Here we define $r(x_i)$ to represent the overlapping degree of sample x_i based on the ratio of the heterogeneous samples in its neighborhood as follows:

$$r(x_i) = \frac{Difflabel(x_i)}{k} - OR(x_i) \quad (1)$$

where $Difflabel(x_i) = \{x \mid x \in kNN(x_i)$ && $Label(x) \neq Label(x_i)\}$, $kNN(x_i)$ is the set of the k nearest samples of x_i [25,26], $OR(x_i) = \{x \mid x \in X, Label(x) \neq Label(x_i)\}/n$. $Difflabel(x_i)/k$ is the heterogeneous sample ratio in the neighborhoods of x_i, and $OR(x_i)$ is the heterogeneous sample ratio in the training data.

$r(x_i) > 0$ means that, in the neighbor area of sample x_i, the ratio of the samples belonging to the different class as x_i is larger than the ratio of the samples belonging to the different class as x_i in the whole training data, there are too many heterogeneous samples in the neighbor area of sample x_i.

To measure the overlapping degrees of the samples without bias, $r(x_i)$ is normalized as follows:

$$Nr(x_i) = \frac{r(x_i)}{OR(x_i)} \tag{2}$$

Therefore, $Nr(x)$ represents the degree that sample x mixes with heterogeneous samples, and the average $Nr(x)$ of all samples in the dataset reflects the mixing degree of the different class samples on the current subspace. If different classes show different distributions on the current feature subspace, then there is a clear separation among different classes, and the average $Nr(x)$ of all samples is small. If different classes show almost the same distribution, they mix together on the subspace, and the average $Nr(x)$ of all samples is large. Hence, the average $Nr(x)$ can express how much discriminative information the current feature subset contains.

2.2. Feature Selection Based on SVM-RFE, the Overlapping Degree, and the Accuracy Rate

SVM-RFE [11,12] is a backward feature deletion method. At first, the current feature subset F contains all the input features. In each loop, an SVM learning model is built based on the current feature subset F, the weight ($|w|$) of each feature in F is calculated according to the support vectors on the hyper-plane of the SVM classifier. The features are then ranked based on $|w|$, and the bottom ranked features are removed from F. This procedure is repeated until F is empty. The feature removing sequence represents the feature importance rank [11,12]. The later the features are removed from F, the more important the features are. The top ranked features are those that are removed from F in the last iteration of SVM-RFE.

Thus, we can obtain a feature rank via SVM-RFE. However, for a certain data analysis, how many top ranked features should be selected from the feature rank of SVM-RFE is still to be considered. In some cases, the number of features to be selected is decided according to prior knowledge or is simply decided subjectively [23]. In other cases, the "optimal" feature subset is kept during the iteration as the final selected feature subset [11,16]. That is, in each iteration, the accuracy rate of the SVM learning model and the feature weights are calculated, and the features having the smallest weights are removed from F. When the procedure terminates, the feature subset corresponding to the maximal accuracy rate is kept as the final selected feature subset [11,16].

In biological data analysis, defining the most informative features (such as genes and metabolites) from the large complex data is of great importance to disease diagnosis and drug study. SVM-RFE is very efficient in analysis of large complex data. However, it is quite difficult to predetermine how many top ranked features should be selected from the feature rank of SVM-RFE. The classification accuracy rate of d-fold cross validation on the training dataset can be applied to determine which feature subset is selected during the backward feature deletion, i.e., the number of the selected features is determined by the classification accuracy rate on the training dataset. However, classification accuracy reflects the discriminative ability of the feature subset based on the classifier, and the distribution of the samples can also reflect the discriminative ability of the feature subset. If different class samples mix together on the current subspace, the overlapping degree of the samples is large, and the subspace has little discriminative information. Both the classification accuracy and the overlapping degree of the samples can tell us how much discriminative information the feature subset has. They evaluate the feature subset from two different aspects, respectively. The discriminative ability of the feature subset could be evaluated more comprehensively by combining these two terms. Hence, we propose SVM-RFE-OA (see Algorithm 1 SVM-RFE-OA), which measures the feature subset during the iterations of SVM-RFE by integrating the average overlapping degree of samples and the classification accuracy rate, and selects a feature subset that has a large accuracy rate and a small overlapping degree. In each iteration, SVM-RFE-OA calculates the average accuracy rate (T_c_acc) of d-fold cross validation and

the average $Nr(x)$ (T_c_oa) of all the samples in the training data, and the feature subset having the largest "$T\_c\_acc - T\_c\_oa$" is kept as the final selected feature subset.

Algorithm 1 SVM-RFE-OA

Input: training dataset X, t.
Output: selected feature subset FS.
Begin
  $c\_acc = 0$;
  $c\_oa = \infty$;
  F = {all input features};
  While ($|F| > 0$) Do
    Construct an SVM based on X and F;
    T_c_acc = d-fold cross validation accuracy rate of SVM;
    T_c_oa = average $Nr(x)$ of the samples in X based on F;
    Rank the features in F by $|w|$ in descending order;
    If $T\_c\_acc - T\_c\_oa > c\_acc - c\_oa$ Then
      $c\_acc = T\_c\_acc$;
      $c\_oa = T\_c\_oa$;
      $FS = F$;
    Endif;
    $F = F -$ {$t \times |F|$ bottom ranked features in F};
  Endwhile;
  Return FS;
End.

t ($0 < t < 100\%$) is the filter factor. In each iteration of SVM-RFE, $t \times |F|$ bottom ranked features are removed from the current feature subset F.

2.3. Modified-SVM-RFE-OA

In the calculation of feature weights, only the samples on the hyper-plane of the SVM learning model are considered [11,12]. The hyper-plane is constructed based on the training samples and the current subspace. The quality of the training data can affect the hyper-plane construction and the computation of feature weights. If different group samples mix heavily on the subspace, overfitting may occur, which can induce the bias of the calculation of feature weights. Therefore, to get a more accurate calculation of the feature weights, we propose a modified algorithm based on SVM-RFE-OA (M-SVM-RFE-OA), which temporally screens out the samples lying in a heavy overlapping area in each iteration (see Algorithm 2 M-SVM-RFE-OA). That is, (1) in each iteration, $Nr(x)$ of each sample in the training data is calculated based on the current subspace F; (2) the samples with $Nr(x) > 0$ are temporarily set aside and are not used in SVM training in this iteration. At most, one-third of the samples in each class in the training data are screened out to make sure that there are enough samples kept for the training. Since the samples in the heavy overlapping area are shielded in the training procedure, there is little chance that overfitting occurs, and the bias becomes small.

Algorithm 2 M-SVM-RFE-OA

Input: training dataset X, t.
Output: selected feature subset FS.
Begin
  $c\_acc = 0$;
  $c\_oa = \infty$;
  F = {all input features};
  While ($|F| > 0$) Do
    Calculate $Nr(x)$ for each $x \in X$ based on F;
    $X_t = X$;
    For each $c \in C$ Do
      $X_c = \{x \mid x \in X, Label(x) = c \text{ and } Nr(x) > 0\}$;
      $\theta = |\{x \mid x \in X, Label(x) = c\}|$;
      If $|X_c| > \theta/3$ Then
        Rank the samples in X_c based on $Nr(x)$ in descending order;
        $X_t = X_t - \{\theta/3$ top ranked samples in $X_c\}$;
      Else
        $X_t = X_t - X_c$;
      End if;
    End for;
    Construct an SVM based on X_t and F;
    T_c_acc = accuracy rate of SVM;
    T_c_oa = average $Nr(x)$ of the samples in X_t based on F;
    If $T\_c\_acc - T\_c\_oa > c\_acc - c\_oa$ Then
      $c\_acc = T\_c\_acc$;
      $c\_oa = T\_c\_oa$;
      $FS = F$;
    End if;
    Rank the features in F by $|w|$ in descending order;
    $F = F - \{t \times |F|$ bottom ranked features in $F\}$;
  End while;
  Return FS;
End

3. Results and Discussion

To show the performance of the two techniques proposed, SVM-RFE-OA and M-SVM-RFE-OA were compared with SVM-RFE where the selected feature subset was determined by the classification accuracy rate. Eight public biological datasets were used in the comparison of the three algorithms, where Breast2, Colon, Lymphoma, Prostate, Brain_data, Srbct are from http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html, and the last two datasets are from www.gems-system.org. Breast2 [27,28] contains 77 samples, including 33 samples that developed distant metastases within 5 years and 44 samples that remained disease-free for over 5 years. The Colon [27,29] dataset includes 40 tumors samples and 22 normal colon tissues samples with 2000 genes by Affymetrix technology. The DLBCL_GEMS [30] dataset includes 58 diffuse large B-cell lymphomas (DLBCL) samples and 19 follicular lymphomas samples. The Lymphoma [27,31] dataset contains the most prevalent adult lymphoid malignancies. The total sample size is 62, including 42 samples of diffuse large B-cell lymphoma, 9 follicular lymphoma samples, and 11 chronic lymphocytic leukemia samples. Prostate [27,32] contains 52 prostate tumors samples and 50 non-tumor prostate samples. Brain_data [27,33] contains 42 samples, which include 5 different tumors of the central nervous system, 10 medulloblastomas samples, 10 malignant gliomas samples, 10 atypical teratoid/rhabdoid tumors (AT/RTs) samples, 8 primitive neuro-ectodermal tumors (PNETs) samples, and 4 human

cerebella samples. The Leukemia2_GEMS [30] dataset contains 24 acute lymphoblastic leukemia (ALL) samples, 28 acute myeloid leukemia (AML) samples, and 20 mixed-lineage leukemia (MLL) samples. The Srbct [27,34] dataset, named the small, round blue cell tumors of childhood, includes 23 neuroblastoma (NB) samples, 20 rhabdomyosarcoma (RMS) samples, 12 non-Hodgkin lymphoma (NHL) samples, and 8 the Ewing family of tumors (EWS) samples. Table 1 provides detailed information of the eight datasets. Four of them are binary problems.

SVM-RFE, SVM-RFE-OA, and M-SVM-RFE-OA were implemented in C++. SVM was obtained from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Linear kernel was adopted in the SVM, and t was set to 5%. In SVM-RFE-OA and M-SVM-RFE-OA, k was set to 9. Five-fold cross validation was run 50 times for each method. The average classification accuracy rates and the average standard deviations are given in Table 2. For the four binary datasets, sensitivities and specificities are given in Tables 3 and 4, respectively. In the tables, the bold numbers represent the largest value in a dataset among the three methods.

Table 1. Data description.

Datasets	No. of Samples	No. of Features	No. of Classes
Breast2 [27,28]	77	4869	2
Colon [27,29]	62	2000	2
DLBCL_GEMS [30]	77	5469	2
Lymphoma [27,31]	62	4026	3
Prostate [27,32]	102	6033	2
Brain_data [27,33]	42	5597	5
Leukemia2_GEMS [30]	72	11225	3
Srbct [27,34]	63	2308	4

Table 2. Comparison in accuracy (%).

Datasets	SVM-RFE	SVM-RFE-OA	M-SVM-RFE-OA
Breast2	61.96 ± 4.57	61.13 ± 4.17	**65.19 ± 3.77**
Colon	80.39 ± 3.99	83.92 ± 2.97	**88.61 ± 1.44**
DLBCL_GEMS	89.09 ± 4.35	**94.29 ± 2.25**	93.69 ± 2.97
Lymphoma	94.14 ± 2.63	95.07 ± 2.25	**95.63 ± 2.38**
Prostate	89.46 ± 2.14	91.84 ± 1.82	**92.24 ± 1.56**
Brain_data	71.78 ± 5.09	80.63 ± 4.58	**81.98 ± 3.21**
Leukemia2_GEMS	89.83 ± 2.80	94.39 ± 2.28	**94.69 ± 2.03**
Srbct	95.21 ± 2.79	98.43 ± 1.45	**98.59 ± 1.44**

Bold: the largest value in a dataset among the three methods.

Table 3. Comparison in sensitivity (%).

Datasets	SVM-RFE	SVM-RFE-OA	M-SVM-RFE-OA
Breast2	66.95 ± 4.99	66.45 ± 4.94	**68.27 ± 5.07**
Colon	86.35 ± 3.40	88.65 ± 2.27	**90.15 ± 1.63**
DLBCL_GEMS	81.16 ± 8.97	**88.32 ± 5.55**	88.32 ± 6.49
Prostate	89.38 ± 2.67	**91.19 ± 2.17**	90.23 ± 2.01

Bold: the largest value in a dataset among the three methods.

Table 4. Comparison in specificity (%).

Datasets	SVM-RFE	SVM-RFE-OA	M-SVM-RFE-OA
Breast2	55.27 ± 7.23	53.94 ± 7.55	**61.03 ± 6.27**
Colon	69.45 ± 9.72	75.27 ± 6.76	**85.82 ± 2.70**
DLBCL_GEMS	91.72 ± 4.14	**96.24 ± 2.43**	95.45 ± 2.78
Prostate	89.52 ± 3.65	92.48 ± 2.84	**94.32 ± 2.19**

Bold: the largest value in a dataset among the three methods.

First, we compared SVM-RFE-OA with SVM-RFE to examine the performance of SVM-RFE-OA. Table 2 shows that SVM-RFE-OA outperforms SVM-RFE for seven of the eight biological datasets in classification accuracy rate. The accuracy rate of SVM-RFE-OA is higher than that of SVM-RFE by 8.95% for Brain_data. Only for Breast2 is the average accuracy rate of SVM-RFE-OA lower than that of SVM-RFE (by 0.83%). The sensitivities and specificities (see Tables 3 and 4) for the four binary problems also show the superiority of SVM-RFE-OA over SVM-RFE. The sensitivities of SVM-RFE-OA are higher than those of SVM-RFE for three of the four binary datasets, and its specificities are higher than those of SVM-RFE for three of the four datasets, too. Hence, the discriminative ability of the feature subset could be measured more accurately by combining the classification accuracy rate with the average overlapping degree of samples than by using the classification accuracy rate alone. The classification accuracy reflects feature distinguishing ability via the classification model, while the average overlapping degree of the samples represents the discriminative information that the feature subset contains by means of the sample distribution. Combining the two criteria induces a more comprehensive measurement of the feature subset. This technique can be used in other RFE analyses to determine the final selected feature subset.

Secondly, we compared M-SVM-RFE-OA with SVM-RFE-OA, thereby examining the performance of temporally screening out the poor samples lying in an overlapping area. Both M-SVM-RFE-OA and SVM-RFE-OA combine the classification accuracy rate and the average overlapping degree to calculate the discriminative ability of the feature subset and determine the number of top ranked features to be selected. To measure the feature importance more accurately, M-SVM-RFE-OA temporarily shields the samples in the overlapping area in each iteration. The comparison between M-SVM-RFE-OA and SVM-RFE-OA shows that temporarily screening out the samples mixed with heterogeneous samples in each iteration benefits the calculation of feature weights. Table 2 clearly shows that M-SVM-RFE-OA outperforms SVM-RFE-OA for seven of the eight datasets in terms of the accuracy rate. Tables 3 and 4 also represent the superiority of M-SVM-RFE-OA over SVM-RFE-OA in sensitivity and specificity. Therefore, we have that the quality of the training data influences the construction of the SVM model and the calculation of feature weights. M-SVM-RFE-OA produces a more accurate calculation of the feature weights by temporally screening out the samples with high overlapping ratios in each iteration, finally obtaining a more powerful feature subset.

The comparisons between SVM-RFE and SVM-RFE-OA and between SVM-RFE-OA and M-SVM-RFE-OA validate the two techniques proposed in this study. Finally, it can be seen that M-SVM-RFE-OA outperforms SVM-RFE for all eight datasets in terms of accuracy rate and outperforms SVM-RFE for all the four binary datasets in terms of sensitivity and specificity. Especially for Brain_data, the accuracy rate of M-SVM-RFE-OA is higher than that of SVM-RFE by 10.2%.

Meanwhile, SVM-RFE-OA and M-SVM-RFE-OA are more stable than SVM-RFE. The standard deviations of M-SVM-RFE-OA on accuracy rate, sensitivity, and specificity are lower than those of SVM-RFE in most cases. Hence, from two different aspects, the classification accuracy rate and the average overlapping degree of samples, which reflects the sample distribution on the feature subspace (top ranked feature subset), we can obtain a more comprehensive measurement of the feature subset. Further, temporally shielding the samples with high overlapping ratios in each iteration could make the computation of feature importance more accurate.

Table 5 gives the average number of features selected in five-fold cross validation run 50 times for each method. It can be seen that the average number of features selected by SVM-RFE is less than those selected by SVM-RFE-OA and M-SVM-RFE-OA. However, the classification accuracy rates of SVM-RFE-OA and M-SVM-RFE-OA are higher than those of SVM-RFE, and the standard deviations of SVM-RFE-OA and M-SVM-RFE-OA are lower than those of SVM-RFE (see Table 2). For the Lymphoma dataset, the average number of selected features by SVM-RFE is 3.48, while SVM-RFE-OA and M-SVM-RFE-OA increase the classification accuracy rate 0.93% and 1.49% by 1.59 and 1.62 more features, respectively. Although the average numbers of features selected by SVM-RFE-OA and

M-SVM-RFE-OA are larger than those by SVM-RFE, the two new methods are much more efficient and stable than SVM-RFE.

Table 5. The average number of features selected.

Datasets	SVM-RFE	SVM-RFE-OA	M-SVM-RFE-OA
Breast2	17.54	52.34	44.89
Colon	12.41	29.46	52.36
DLBCL_GEMS	7.62	39.54	34.50
Lymphoma	3.48	5.07	5.10
Prostate	12.73	60.16	57.50
Brain_data	12.94	48.15	121.68
Leukemia2_GEMS	9.61	78.49	73.99
Srbct	7.22	31.04	30.18

4. Conclusions

In systems biology, it is very significant to select the most meaningful features from large complex genomics, metabolomics, and proteomics data, which could help in classifying different disease samples, studying disease mechanisms, and developing new drugs. This paper proposes two techniques of selecting discriminative feature subsets based on SVM-RFE. One is measuring the feature subset by combining the classification accuracy rate with the average overlapping degree of samples, and the other is temporally screening out the samples in a heavily overlapping area in each loop of the SVM-RFE. Experiments on eight public biological datasets show the validation of these techniques and prove that filtering out the samples that lie in the heavily overlapping area could make the measurement of feature weights more accurate.

Acknowledgments: The study has been supported by the National Natural Science Foundation of China (21375011).

Author Contributions: X.L. and M.F. conceived and designed the experiments; C.L., M.F. and B.S. performed the experiments; M.F., Y.Z., and H.W. searched the datasets; C.L. and M.F. analyzed the results.

Conflicts of Interest: We declare no conflict of interest.

References

1. Jabid, T.; Uddin, M.Z. Rotation invariant power line insulator detection using local directional pattern and support vector machine. In Proceedings of the IEEE Conference on Innovations in Science, Engineering and Technology (ICISET), Dhaka, Bangladesh, 28–29 October 2016; pp. 1–4. [CrossRef]
2. Jothi, G. Hybrid Tolerance Rough Set–Firefly based supervised feature selection for MRI brain tumor image classification. *Appl. Soft Comput.* **2016**, *46*, 639–651. [CrossRef]
3. Lou, I.; Xie, Z.; Ung, W.K.; Mok, K.M. Integrating support vector regression with particle swarm optimization for numerical modeling for algal blooms of freshwater. In *Advances in Monitoring and Modelling Algal Blooms in Freshwater Reservoirs*; Springer: Dordrecht, The Netherlands, 2017; pp. 125–141, ISBN 978-94-024-0933-8. [CrossRef]
4. Puthiyedth, N.; Riveros, C.; Berretta, R.; Moscato, P. A new combinatorial optimization approach for integrated feature selection using different datasets: A prostate cancer transcriptomic study. *PLoS ONE* **2015**, *10*, 1–26. [CrossRef] [PubMed]
5. Christin, C.; Hoefsloot, H.C.J.; Smilde, A.K.; Hoekman, B.; Suits, F.; Bischoff, R.; Horvatovich, P. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol. Cell. Proteom.* **2013**, *12*, 263–276. [CrossRef] [PubMed]
6. Wei, L.; Xing, P.; Shi, G.; Ji, Z. L.; Zou, Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE ACM T. Comput. Biol. Bioinform.* **2017**, in press. [CrossRef] [PubMed]
7. Zou, Q.; Wan, S.; Ju, Y.; Tang, J.; Zeng, X. Pretata: Predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* **2016**, *10*, 401–412. [CrossRef] [PubMed]

8. Lin, X.; Song, H.; Fan, M.; Ren, W.; Li, L.; Yao, W. The feature selection algorithm based on feature overlapping and group overlapping. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, Shenzhen, China, 15–18 December 2017; pp. 619–624. [CrossRef]
9. Suthaharan, S. Support vector machine. In *Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems*; Springer: Boston, MA, USA, 2016; pp. 207–235, ISBN 978-1-4899-7640-6. [CrossRef]
10. Butkiewicz, M.; Lowe, E.; Mueller, R.; Mendenhall, J.; Teixeira, P.; Weaver, C.; Meiler, J. Benchmarking ligand-based virtual high-throughput screening with the pubchem database. *Molecules* **2013**, *18*, 735–756. [CrossRef] [PubMed]
11. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
12. 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]
13. Hidalgo-Muñoz, A.R.; López, M.M.; Pereira, A.T.; Tomé, A. Spectral turbulence measuring as feature extraction method from EEG on affective computing. *Biomed. Signal Process. Control* **2013**, *8*, 945–950. [CrossRef]
14. Dao, F.Y.; Yang, H.; Su, Z.D.; Yang, W.R.T.; Wu, Y.; Ding, H.; Chen, W.; Tang, H.; Lin, H. Recent advances in conotoxin classification by using machine learning methods. *Molecules* **2017**, *22*, 1057. [CrossRef] [PubMed]
15. Mahadevan, S.; Shah, S.L.; Marrie, T.J.; Slupsky, C.M. Analysis of metabolomic data using support vector machines. *Anal. Chem.* **2008**, *80*, 7562–7570. [CrossRef] [PubMed]
16. Lin, X.H.; Yang, F.F.; Zhou, L.N.; Yin, P.Y.; Kong, H.W.; Xing, L.; Jia, L.W.; Wang, Q.C.; Xu, G.W. A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *J. Chromatogr. B* **2012**, *910*, 149–155. [CrossRef] [PubMed]
17. Tang, Y.; Zhang, Y.Q.; Huang, Z. Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2007**, *4*, 365–381. [CrossRef] [PubMed]
18. Li, X.B.; Peng, S.H.; Chen, J.; Lü, B.J.; Zhang, H.H.; Lai, M.D. 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]
19. Mundra, P.A.; Rajapakse, J.C. SVM-RFE with MRMR filter for gene selection. *IEEE Trans. Nanobiosci.* **2010**, *9*, 31–37. [CrossRef] [PubMed]
20. Zhang, X.G.; Lu, X.; Shi, Q.; Xu, X.Q.; Hon-Chiu, E.L.; Harris, L.N.; Iglehart, J.D.; Miron, A.; Liu, J.S.; Wong, W.H. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinform.* **2006**, *7*, 1–13. [CrossRef]
21. Bolón-Canedo, V.; Sánchez-Marono, N.; Alonso-Betanzos, A.; Benítez, J.M.; Herrera, F. A review of microarray datasets and applied feature selection methods. *Inf. Sci.* **2014**, *282*, 111–135. [CrossRef]
22. Ding, Y.; Wilkins, D. Improving the performance of SVM-RFE to select genes in microarray data. *BMC Bioinform.* **2006**, *7*, S12. [CrossRef] [PubMed]
23. Zhou, X.; Tuck, D.P. MSVM-RFE: Extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics* **2007**, *23*, 1106–1114. [CrossRef] [PubMed]
24. Zou, Q.; Zeng, J.; Cao, L.; Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* **2016**, *173*, 346–354. [CrossRef]
25. Lee, J.; Batnyam, N.; Oh, S. RFS: Efficient feature selection method based on R-value. *Comput. Biol. Med.* **2013**, *43*, 91–99. [CrossRef] [PubMed]
26. Statnikov, A.; Henaff, M.; Narendra, V.; Konganti, K.; Li, Z.G.; Yang, L.Y.; Pei, Z.H.; Blaser, M.J.; Aliferis, C.F.; Alekseyenko, A.V. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* **2013**, *1*, 1–11. [CrossRef] [PubMed]
27. Díaz-Uriarte, R.; Andrés, S.A.D. Gene selection and classification of microarray data using random forest. *BMC Bioinform.* **2006**, *7*, 3. [CrossRef] [PubMed]
28. Van't Veer, L.J.; Dai, H.; Van De Vijver, M.J.; He, Y.D.; Hart, A.A.; Mao, M.; Peterse, H.L.; Van Der Kooy, K.; Marton, M.J.; Witteveen, A.T.; et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **2002**, *415*, 530–536. [CrossRef] [PubMed]

29. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 6745–6750. [CrossRef] [PubMed]
30. Statnikov, A.; Tsamardinos, I.; Dosbayev, Y.; Aliferis, C.F. Gems: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *Int. J. Med. Inform.* **2005**, *74*, 491–503. [CrossRef] [PubMed]
31. Alizadeh, A.A.; Eisen, M.B.; Davis, R.E.; Ma, C.; Losses, I.S.; Rosenwald, A.; Boldrick, J.C.; Sabet, H.; Tran, T.; Yu, X.; et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **2000**, *403*, 503–511. [CrossRef] [PubMed]
32. Singh, D.; Febbo, P.G.; Ross, K.; Jackson, D.G.; Manola, J.; Ladd, C.; Tamayo, P.; Renshaw, A.A.; D'Amico, A.V.; Richie, J.P.; et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **2002**, *1*, 203–209. [CrossRef]
33. Pomeroy, S.L.; Tamayo, P.; Gaasenbeek, M.; Sturla, L.M.; Angelo, M.; McLaughlin, M.E.; Kim, J.Y.; Goumnerova, L.C.; Black, P.M.; Lau, C.; et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **2002**, *415*, 436–442. [CrossRef] [PubMed]
34. Khan, J.; Wei, J.S.; Ringner, M.; Saal, L.H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C.R.; Peterson, C.; et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **2001**, *7*, 673–679. [CrossRef] [PubMed]

Sample Availability: Not available.

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).