International Journal of
Molecular Sciences

MDPI

Article

Protein Subcellular Localization with Gaussian Kernel Discriminant Analysis and Its Kernel Parameter Selection

Shunfang Wang [1,*,†], Bing Nie [1,†], Kun Yue [1,*], Yu Fei [2,*], Wenjia Li [1] and Dongshu Xu [1]

[1] Department of Computer Science and Engineering, School of Information Science and Engineering, Yunnan University, Kunming 650504, China; bingn2017@gmail.com (B.N.); woshiwenzi666@gmail.com (W.L.); qq78316519@gmail.com (D.X.)

[2] School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming 650221, China

* Correspondence: sfwang_66@ynu.edu.cn (S.W.); kyue@ynu.edu.cn (K.Y.); feiyukm@aliyun.com (Y.F.); Tel.: +86-137-6919-6517 (S.W.); +86-871-6593-3093 (K.Y.); +86-139-8765-9718 (Y.F.)

† These authors contributed equally to this work.

Received: 16 October 2017; Accepted: 5 December 2017; Published: 15 December 2017

Abstract: Kernel discriminant analysis (KDA) is a dimension reduction and classification algorithm based on nonlinear kernel trick, which can be novelly used to treat high-dimensional and complex biological data before undergoing classification processes such as protein subcellular localization. Kernel parameters make a great impact on the performance of the KDA model. Specifically, for KDA with the popular Gaussian kernel, to select the scale parameter is still a challenging problem. Thus, this paper introduces the KDA method and proposes a new method for Gaussian kernel parameter selection depending on the fact that the differences between reconstruction errors of edge normal samples and those of interior normal samples should be maximized for certain suitable kernel parameters. Experiments with various standard data sets of protein subcellular localization show that the overall accuracy of protein classification prediction with KDA is much higher than that without KDA. Meanwhile, the kernel parameter of KDA has a great impact on the efficiency, and the proposed method can produce an optimum parameter, which makes the new algorithm not only perform as effectively as the traditional ones, but also reduce the computational time and thus improve efficiency.

Keywords: protein subcellular localization; kernel parameter selection; kernel discriminant analysis (KDA); Gaussian kernel function; dimension reduction

1. Introduction

Some proteins can only play the role in one specific place in the cell while others can play the role in several places in the cell [1]. Generally, a protein can function correctly only when it is localized to a correct subcellular location [2]. Therefore, protein subcellular localization prediction is an important research area of proteomics. It is helpful to predict protein function as well as to understand the interaction and regulation mechanism of proteins [3]. Now, many methods have been used to predict protein subcellular location, such as green fluorescent protein labeling [4], mass spectrometry [5], and so on. However, these traditional experimental methods usually have many technical limitations, resulting in high cost of time and money. Thus, prediction of protein subcellular location based on machine learning has become a focus research in bioinformatics [6–8].

When we use the methods of machine learning to predict protein subcellular location, we must extract features of protein sequences. We can get some vectors after feature extraction, and then we use the classifier to process these vectors. However, these vectors are usually complex due to their high dimensionality and nonlinear property. In order to improve the prediction accuracy of protein subcellular location, an appropriate nonlinear method for reducing data dimension should be used before classification. Kernel discriminant analysis (KDA) [9] is a nonlinear reductive dimension algorithm based on kernel trick that has been used in many fields such as facial recognition and fingerprint identification. The KDA method not only reduces data dimensionality but also makes use of the classification information. This paper newly introduces the KDA method to predict protein subcellular location. The algorithm of KDA first maps sample data to a high-dimensional feature space by a kernel function, and then executes linear discriminant analysis (LDA) in the high-dimensional feature space [10], which indicates that kernel parameter selection will significantly affect the algorithm performance.

There are some classical algorithms used to select the parameter of kernel function, such as genetic algorithm, grid searching algorithm, and so on. These methods have high calculation precision but large amounts of calculation. In an effort to reduce computational complexity, recently, Xiao et al. proposed a method based on reconstruction errors of samples and used it to select the parameters of Gaussian kernel principal component analysis (KPCA) for novelty detection [11]. Their methods are applied into the toy data sets and UCI (University of CaliforniaIrvine) benchmark data sets to demonstrate the correctness of the algorithm. However, their innovation in the KPCA method aims at dimensional reduction rather than discriminant analysis, which leads to unsatisfied classification prediction accuracy. Thus, it is necessary to improve the efficiency of the method in [11] especially for some complex data such as biological data.

In this paper, an improved algorithm of selecting parameters of Gaussian kernel in KDA is proposed to analyze complex protein data and predict subcellular location. By maximizing the differences of reconstruction errors between edge normal samples and interior normal samples, the proposed method not only shows the same effect as the traditional grid-searching method, but also reduces the computational time and improves efficiency.

2. Results and Discussion

In this section, the proposed method (in Section 3.4) and the grid-searching algorithm (in Section 4.4) are both applied to predict protein subcellular localization. We use two standard data sets as the experimental data. The two used feature expressions are generated from PSSM (position specific scoring matrix) [12], which are the PsePSSM (pseudo-position specific scoring matrix) [12] and the PSSM-S (AAO + PSSM-AAO + PSSM-SAC + PSSM-SD = PSSM-S) [13]. Here AAO means consensus sequence-based occurrence, PSSM-AAO means evolutionary-based occurrence or semi-occurrence of PSSM, PSSM-SD is segmented distribution of PSSM and PSSM-SAC is segmented auto covariance of PSSM. The k-nearest neighbors (KNN) is used as the classifier in which Euclidean distance is adopted for the distance between samples. The flow of experiments is as follows.

- First, for each standard data set, we use the PsePSSM algorithm and the PSSM-S algorithm to extract features, respectively. Then totally we obtain four sample sets, which are GN-1000 (Gram-negative with PsePSSM which contains 1000 features), GN-220 (Gram-negative with PSSM-S which contains 220 features), GP-1000 (Gram-positive with PsePSSM which contains 1000 features) and GP-220 (Gram-positive with PsePSSM which contains 220 features).
- Second, we use the proposed method to select the optimum kernel parameter for the Gaussian KDA model and then use KDA to reduce the dimension of sample sets. The same procedure is also carried out for the traditional grid-searching method to form a comparison with the proposed method.
- Finally, we use the KNN algorithm to classify the reduced dimensional sample sets and use some criterions to evaluate the results and give the comparison results.

Some detailed information in experiments is as follows. For every sample set, we choose the class that contains the most samples to form the training set [8]. Let $S = [0.1, 0.2, 0.3, 0.4, 1, 2, 3, 4]$ be a candidate set of the Gaussian kernel parameter, which is proposed at random. When we use the KDA algorithm to reduce dimension, the number of retained eigenvectors must be less than or equal to C − 1 (C is the number of classes). Therefore, for sample sets GN-1000 and GN-220, the number of retained eigenvectors, which is denoted as d, can be from 1 to 7. For the sample sets GP-1000 and GP-220, d can be 1, 2, and 3. As far as the parameter u is concerned, when it is 5–8% of the average number of samples, good classification can be achieved [14]. Besides, we demonstrate the robustness of the proposed method with the variation of u in Section 2.2. So here we simply pick a general value for u, say 8. To sum up, in the following experiments, when certain parameters need to be fixed, their default values are as follows. The value of d is 7 for sample sets GN-1000 and GN-220, and 3 for GP-1000 and GP-220; the value of u is 8 and the k value in KNN classifier is 20.

2.1. The Comparison Results of the Overall Accuracy

2.1.1. The Accuracy Comparison between the Proposed Method and the Grid-Searching Method

In this section, first, the proposed method and the grid-searching method are respectively used in the prediction of protein subcellular localization with different d values. The experimental results are presented in Figure 1.

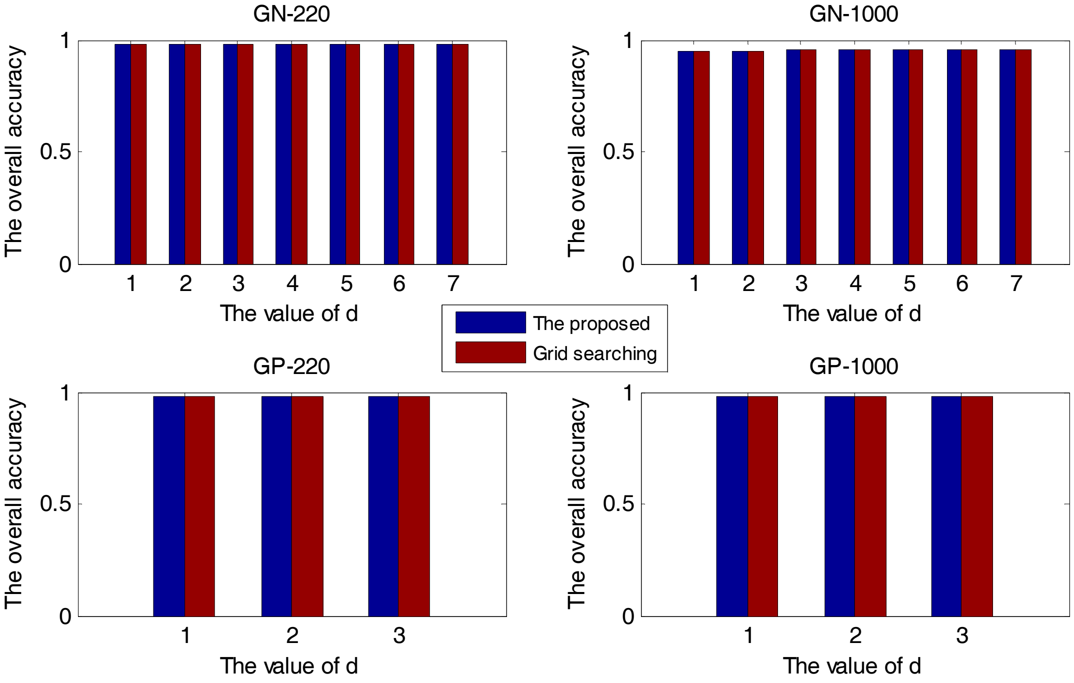

Figure 1. The overall accuracy versus d for four sample sets.

In Figure 1, all four sample sets suggest that when we use the KDA algorithm to reduce dimension, the larger the number of retained eigenvectors, the higher the accuracy. The overall accuracy of the proposed method is always the same as that of the grid-searching method, no matter which value of d. The proposed method is effective for selecting the optimal Gaussian kernel parameter.

Then, in the analyses and experiments, we find that superiority of the proposed method is the low runtime, which is demonstrated in Table 1 and Figure 2.

Table 1. The overall accuracy and the ratio of runtime for two methods.

Sample Sets	Overall Accuracy		Ratio (t_1/t_2)
GP-220 (PSSM-S)	The proposed method Grid searching method	0.9924 0.9924	0.7087
GP-1000 (PsePSSM)	The proposed method Grid searching method	0.9924 0.9924	0.7362
GN-220 (PSSM-S)	The proposed method Grid searching method	0.9801 0.9801	0.7416
GN-1000 (PsePSSM)	The proposed method Grid searching method	0.9574 0.9574	0.7687

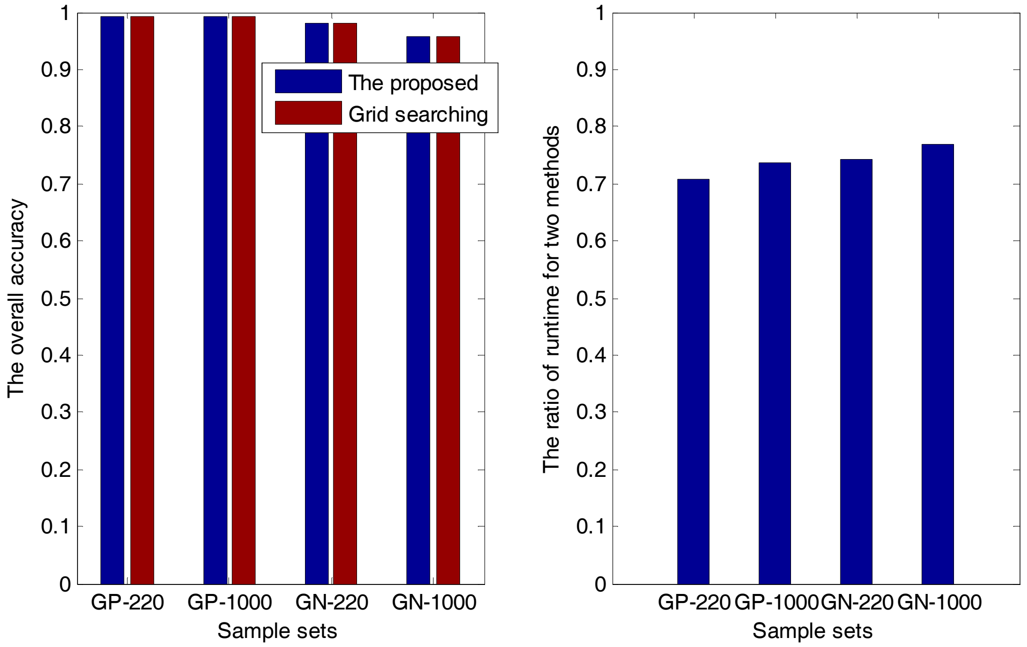

Figure 2. The overall accuracy and the ratio of runtime for two methods.

In Table 1, t_1 and t_2 are the runtimes of the proposed method and the grid-searching method, respectively. The overall accuracy and the ratio of t_1 and t_2 are presented in both Table 1 and Figure 2, from which we can see that for each sample set, the accuracy of the proposed method is always the same as that of the grid-searching method; meanwhile, the runtime of the former is about 70–80% of that of the latter, indicating that the proposed method has a higher efficiency than the grid-searching method.

2.1.2. The Comparison between Methods with and without KDA

In this experiment, we compare the overall accuracies between the cases of using KDA algorithm or not, with k values of the KNN classifier varying from 1 to 30. The experimental results are shown in Figure 3.

For each sample set, Figure 3 shows that the accuracy with KDA algorithm to reduce dimension is higher than that of without it. However, the kernel parameter has a great impact on the efficiency of the KDA algorithm, and the proposed method can be used to select the optimum parameter that makes the KDA perform perfect. Therefore, accuracy can be improved by using the proposed method to predict the protein subcellular localization.

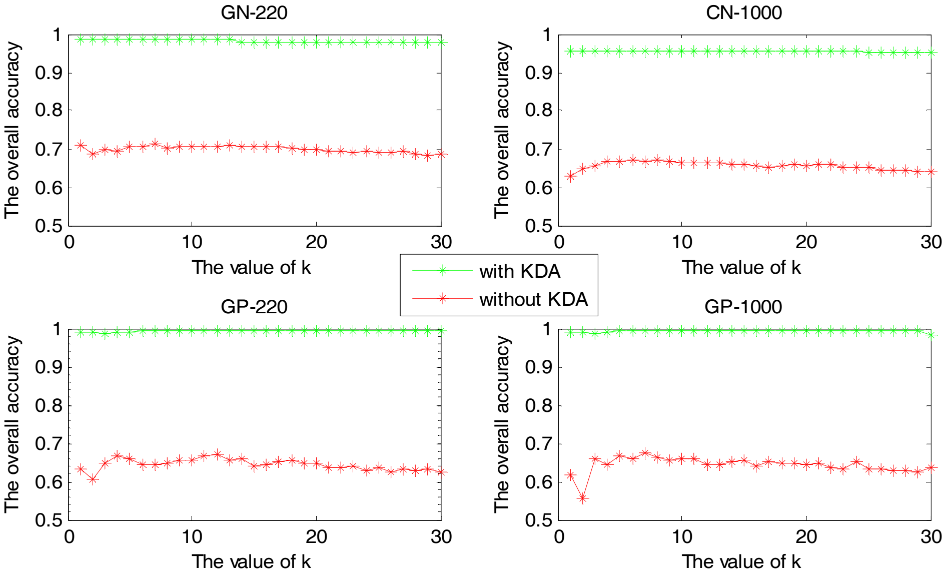

Figure 3. The overall accuracy versus k value with or without KDA algorithm.

2.2. The Robustness of the Proposed Method

In the proposed method, the value of u will have an impact on the radius value of neighborhood so that it can affect the number of the selected internal and edge samples. Figure 4 shows the experimental results when the value of u ranges from 6 to 10, in which the overall accuracies of the proposed method and the grid-searching method are given.

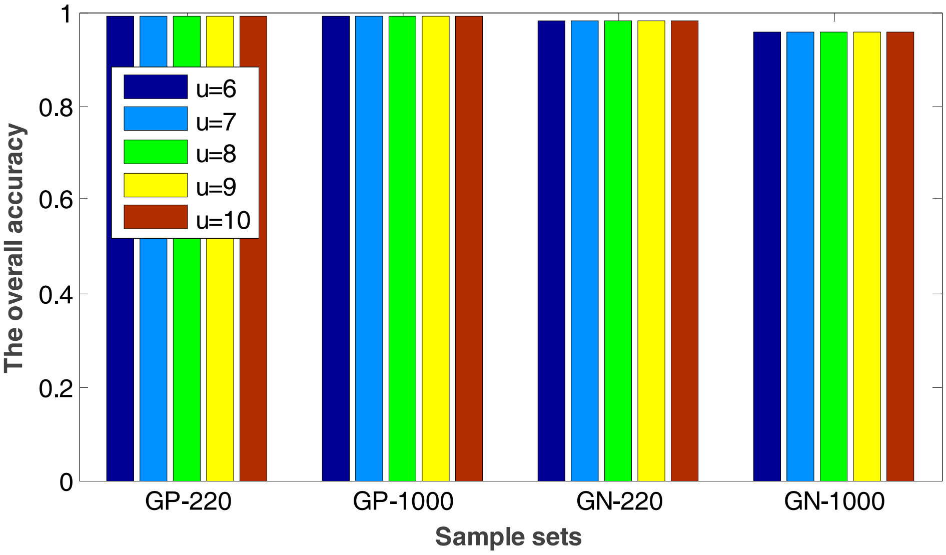

Figure 4. The overall accuracy for four sample sets with different u values.

It is easily seen from Figure 4 that the accuracy keeps invariable with different u values. The number of the selected internal and edge samples has little effect on the performance of the proposed method. Therefore, the method proposed in this paper has a good robustness.

2.3. Evaluating the Proposed Method with Some Regular Evaluation Criterions

In this subsection, we compute the values of some regular evaluation criterions with the proposed method for two standard data sets, which is show in Tables 2 and 3, respectively. In Table 3, "-" means an infinity value, corresponding to the cases when the denominator is 0 in MCC.

Table 2. The values of evaluation criterion with the proposed method for the Gram-positive.

Sample Set	Protein Subcellular Locations			
	Cell Membrane	Cell Wall	Cytoplasm	Extracell
	Sensitivity			
GP-220	1	0.9444	0.9904	0.9919
GP-1000	0.9943	0.9444	1	0.9837
	Specificity			
GP-220	0.9943	1	1	09950
GP-1000	0.9971	1	0.9937	0.9925
	Matthews coefficient correlation (MCC)			
GP-220	0.9914	0.9709	0.9920	0.9841
GP-1000	0.9914	0.9709	0.9921	0.9840
	Overall accuracy (Q)			
GP-220	0.9924			
GP-1000	0.9924			

Table 3. The values of evaluation criterion with the proposed method for the Gram-negative.

Sample Set	Protein Subcellular Locations							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Sensitivity							
GN-220	1	0.9699	1	0	0.9982	0	0.9677	1
GN-1000	1	0.9323	1	0	0.9659	0	0.9516	0.9556
	Specificity							
GN-220	0.9924	0.9902	1	1	0.9978	1	1	0.9953
GN-1000	0.9608	0.9872	1	1	0.9967	1	1	0.9992
	Matthews coefficient correlation (MCC)							
GN-220	0.9866	0.9324	1	-	0.9956	-	0.9823	0.9814
GN-1000	0.9346	0.8957	1	-	0.9681	-	0.9733	0.9712
	Overall accuracy (Q)							
GN-220	0.9801							
GN-1000	0.9574							

(1) Cytoplasm, (2) Extracell, (3) Fimbrium, (4) Flagellum, (5) Inner membrane, (6) Nucleoid, (7) Outer membrane, (8) Periplasm.

Tables 2 and 3 show that the values of the evaluation criterion are close to 1 for the proposed method. Then the selection of the kernel parameter using the proposed method will benefit the protein subcellular localization.

3. Methods

3.1. Protein Subcellular Localization Prediction Based on KDA

To improve the localization prediction accuracy, it is necessary to reduce dimension of high-dimensional protein data before subcellular classification. The flow of protein subcellular localization prediction is presented in Figure 5.

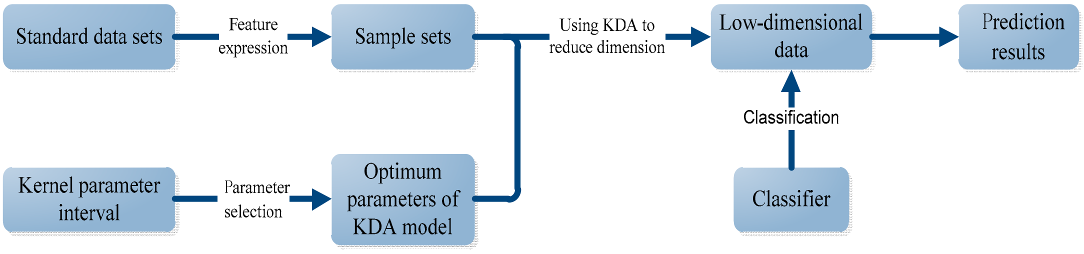

Figure 5. The flow of protein subcellular localization.

As shown in Figure 5, first, for a standard data set, some features of protein sequences such as PSSM-based expressions are extracted to form the sample sets. The specific feature expressions used in this paper are discussed in Section 4.2. Second, the kernel parameter is selected in an interval based on the sample sets to reach its optimal value in KDA model. Third, with this optimal value, we used the KDA to realize the dimension reduction of the sample sets. Lastly, the low dimensional data is treated by certain classifier to realize the classification and the final prediction.

In the whole process of Figure 5, dimension reduction with KDA is very important, in which the kernel selection is a key step and constructs the research focus of this paper. Kernel selection includes the choice of the type of kernel function and the choice of the kernel parameters. In this paper, Gaussian kernel function is adopted for KDA because of its good nature, learning performance, and catholicity. So, the emphasis of this study is to decide the scale parameter of the Gaussian kernel, which plays an important role in the process of dimensionality reduction and has a great influence on prediction results. We put forward a method for selecting the optimum Gaussian kernel parameter with the starting point of reconstruction error idea in [15].

3.2. Algorithm Principle

Kernel method constructs a subspace in the feature space by the kernel trick, which makes normal samples locate in or nearby this subspace, while novel samples are far from it. The reconstruction error is the distance of a sample from the feature space to the subspace [11], so the reconstruction errors of normal samples should be different from those of the novel samples. In this paper, we use the Gaussian KDA as the descending algorithms. Since the values of the reconstruction errors are influenced by the Gaussian kernel parameters, the reconstruction errors of normal samples should be differentiated from those of the novel samples by suitable parameters [11].

In the input space, we usually call the samples on the boundary as edge samples, and call those within the boundary as internal samples [16,17]. The edge samples are much closer to novel samples than the internal samples, while the internal samples are much closer to normal states than the edge samples [11]. We usually use the internal samples as the normal samples and use the edge samples as the novel samples, since there are no novel samples in data sets. Therefore, the principle is that the optimal kernel parameter makes the reconstruction errors have a reasonable difference between the internal samples and the edge samples.

3.3. Kernel Discriminant Analysis (KDA) and Its Reconstruction Error

KDA is an algorithm by applying kernel trick into linear discriminant analysis (LDA). LDA is an algorithm of linear dimensionality reduction together with classifying discrimination, which aims to find a direction that maximizes the between-class scatter while minimizing the within-class scatter [18]. In order to extend the LDA theory to the nonlinear data, Mika et al. proposed the KDA algorithm, which makes the nonlinear data linearly separable in a much higher dimensional feature space than before [9]. The principle of the KDA algorithm is shown as follows.

Suppose the N samples in X can be divided into C classes and the ith class contains N_i samples satisfying $N = \sum_{i=1}^{C} N_i$. The between-class scatter matrix S_b^{ϕ} and the within-class scatter matrix n_n^{ϕ} of X are defined in the following equations, respectively:

$$S_b^{\phi} = \sum_{i=1}^{C} N_i \left(m_i^{\phi} - m^{\phi}\right) \left(m_i^{\phi} - m^{\phi}\right)^{T} \tag{1}$$

$$S_w^{\phi} = \sum_{i=1}^{C} \sum_{j=1}^{N_i} \left[\phi\left(x_j^i\right) - m_i^{\phi}\right] \left[\phi\left(x_j^i\right) - m_i^{\phi}\right]^{T} \tag{2}$$

where $m_i^{\phi} = \frac{1}{N_i} \sum_{j=1}^{N_i} \phi\left(x_j^i\right)$ is the mean vector of the ith class, and $m^{\phi} = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i)$ is the total mean of X. To find the optimal linear discriminant, we need to maximize $J(W)$ as follows:

$$\max J(W) = \frac{W^T S_b^{\phi} W}{W^T S_w^{\phi} W} \tag{3}$$

where $W = [w_1, w_2, \cdots, w_d]^T (1 \leq d \leq C - 1)$ is a projection matrix, and $w_k (k = 1, 2, \cdots, d)$ is a column vector with N elements. Through certain algebra, it can be deduced that W is made up of the eigenvectors corresponding to the top d eigenvalues of $S_w^{\phi\,-1} S_b^{\phi}$. Also, the projection vector w_k can be represented by a linear combination of the samples in the feature space:

$$w_k = \sum_{j=1}^{N} a_j^k j(x_j) \tag{4}$$

where a_j^k is a real coefficient. The projection of the sample X onto w_k is given by:

$$w_k^T \times \phi(x) = \sum_{i=1}^{N} a_i^k K(x, x_j) \tag{5}$$

Let $a = \left[a^1, a^2, \cdots, a^d\right]^T$ be the coefficient matrix where $a^k = \left[a_1^k, a_2^k, \cdots, a_N^k\right]^T$ is the coefficient vector. Combining Equations (1)–(5), we can obtain the linear discriminant by maximizing the function $J(a)$:

$$\max J(a) = \frac{a^T \widetilde{M} a}{a^T \widetilde{L} a} \tag{6}$$

where $\widetilde{M} = \sum_{i=1}^{C} N_i (M_i - M)(M_i - M)^T$, $\widetilde{L} = \sum_{i=1}^{C} K_i \left(E - \frac{1}{N_i} I\right) K_i^T$, the kth component of the vector M_i is $(M_i)_k = \frac{1}{N_i} \sum_{j=1}^{N_i} K\left(x_k, x_j^i\right) (k = 1, 2, \cdots, N)$, the kth component of the vector M is $(M)_k = \frac{1}{N} \sum_{j=1}^{N} K\left(x_k, x_j^i\right) (k = 1, 2, \cdots, N)$, K_i is a $N \times N_i$ matrix with $(K_i)_{mn} = K(x_m, x_n^i)$, E is the $N_i \times N_i$ identity matrix, and $\frac{1}{N_i} I$ is the $N_i \times N_i$ matrix that all elements are $\frac{1}{N_i}$ [9]. Then, the projection matrix a is made up of the eigenvectors corresponding to the top d eigenvalues of $\widetilde{L}^{-1} \widetilde{M}$.

According to the KDA algorithm principle in (3) or (6), besides the Gaussian kernel parameter s, the number of retained eigenvectors d also affects the algorithm performance. Generally, in this paper, the proposed method is mainly used to screen an optimum S under a predetermined d value.

The Gaussian kernel function is defined as follows:

$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2}) \tag{7}$$

where σ is the scale parameter which is generally estimated by s. Note that $\|\phi(x)\|^2 = K(x, x) = 1$.

The kernel-based reconstruction error is defined in the following equation:

$$\begin{aligned} RE(x) &= \|\phi(x) - W\,t(x)\,\|^2 = \|\phi(x)\,\|^2 - \|\,t(x)\,\|^2 \\ &= K(x, x) - \|\,t(x)\,\|^2 \end{aligned} \tag{8}$$

where t(x) is the vector obtained by projecting $\phi(x)$ onto a projection matrix a.

3.4. The Proposed Method for Selecting the Optimum Gaussian Kernel Parameter

The method of kernel parameter selection relies on the reconstruction errors of the internal samples and the edge samples. Therefore, first we find a method to select the edge samples and the interior samples, then we propose the method for selection of the Gaussian kernel parameter.

3.4.1. The Method for Selecting Internal and Edge Samples

Li and Maguire present a border-edge pattern selection method (BEPS) to select the edge samples based on the local geometric information [16]. Xiao et al. [11] modified the BEPS algorithm so that it can select both the edge samples and internal samples. However, their algorithm has the risk of making all samples in the training set become the edge samples. For example, when all samples are distributed on a spherical surface in a three-dimensional space, every sample in the data set will be selected as the edge samples since its neighbors are all located on one side of its tangent plane. In order to solve this problem, this paper innovatively combines the ideas in [19,20] to select the internal and edge samples, respectively, which is not dependent on the local geometric information. The main principle is that the edge sample is usually surrounded by the samples belonging to other classes while the internal sample is usually surrounded by the samples belonging to its same class. Further, the edge samples are usually far from the centroid of this class, while the internal samples are usually close to the centroid. So, a sample will be selected as the edge sample if it is far from the centroid of this class and there are samples around it that belongs to other classes, otherwise it will be selected as the internal sample.

Specifically, suppose the ith class $X_i = \{x_1, x_2, \cdots, x_{N_i}\}$ in the sample set X is picked out as the training set. Denote c_i be the centroid of this class:

$$c_i = \frac{1}{N_i}\sum_{i=1}^{N_i} x_i \tag{9}$$

We use the median value m of the distances from all samples in a class to its centroid to measure the distance from a sample to the centroid of this class. A sample is conserved to be far from the centroid of this class if the distance from this sample to the centroid is greater than the median value. Otherwise, the sample is considered to be close to the centroid.

Denote $\mathrm{dist}(x_i, x_j)$ as the distance between any two samples x_i and x_j, and $N_\varepsilon(x)$ as the ε-neighborhood of X:

$$N_\varepsilon(x) = \{y | \mathrm{dist}(x, y) \le \varepsilon, y \in X\} \tag{10}$$

The value of neighborhood ε is given as follows. Let u be a given number which satisfies $0 < u < N_i$. $\mathrm{Density}_u(X_i)$ is the mean radius of neighborhood of X_i for the given number u:

$$\mathrm{Density}_u(X_i) = \frac{1}{N_i}\sum_{i=1}^{N_i} \mathrm{dist}_u(x_i) \tag{11}$$

where $\mathrm{dist}_u(x_i)$ is the distance from x_i to its uth nearest neighbor. So, $\mathrm{Density}_u(X_i)$ is used as the value of ε for the training set X_i. The flow for the selection of the internal and edge samples is shown in Table 4.

Table 4. The Selection of Internal and Edge Samples.

Input: $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_C}\}$**, the training set** $\mathbf{X_i} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_{N_i}}\}$ $(\mathbf{1 \le i \le C})$**.**
1. Calculate the radius of neighborhood ε using Equation (11). 2. Calculate the centroid c_i of the i^{th} class according to Equation (9). 3. Calculate the distances $\mathrm{dist}_j(j = 1, 2, \cdots, N_i)$ from all samples in training set to c_i, respectively, and the median value m of them. 4. For each training sample x_j of the set X_i • Calculate the $N_\varepsilon(x_j)$ according to Equation (10). • If $\mathrm{dist}_j > m$ and there are samples in $N_\varepsilon(x_j)$ belonging to other classes, x_j is selected as an edge sample. • If $\mathrm{dist}_j < m$ and no sample in $N_\varepsilon(x_j)$ belongs to other classes, x_j is selected as an internal sample.
Output: the selected internal sample set $\mathbf{\Omega_{in}}$**, the selected edge sample set** $\mathbf{\Omega_{ed}}$**.**

In Table 4, a sample X is considered to be the edge one when the distance from X to the centroid is larger than the median m and there are samples in $N_\varepsilon(x)$ belonging to other classes in this case. A sample X is considered to be the internal one when the distance from X to the centroid is less than m and in this case all samples of $N_\varepsilon(x)$ belong to this class.

3.4.2. The Proposed Method

In order to select the optimum kernel parameter, it is necessary to propose a criterion aiming to distinguish reconstruction errors of the edge samples from those of the internal samples. A suitable parameter not only maximizes the difference between reconstruction errors of the internal samples and those of the edge samples, but also minimizes the variance (or standard deviation) of reconstruction errors of the internal samples [11]. According to the rule, an improved objective function is proposed in this paper. The optimal Gaussian kernel parameter S is selected by maximizing this objective function.

$$s = \underset{s}{\mathrm{argmax}} f(s) = \arg\max_{s} \frac{\|\mathrm{RE}(\Omega_{ed})\|_\infty - \|\mathrm{RE}(\Omega_{in})\|_\infty}{\mathrm{std}\{\mathrm{RE}(\Omega_{in})\}} \tag{12}$$

where $\|\cdot\|_\infty$ is the infinite norm which computes the maximum absolute component of a vector and $\mathrm{std}(\cdot)$ is a function of the standard deviation. Note that in the objective function $f(s)$, our key improvement is to use the infinite norm to compute the size of reconstruction error vector since it can lead to a higher accuracy than many other measurements, which has been verified by a series of our experiments. The reason is probably that the maximum component is more reasonable to evaluate the size of a reconstruction error vector than others such as the 1-norm, p-norm $(1 < p < +\infty)$ and the minimum component of a reconstruction error vector in [11].

According to (8), when the number of retained eigenvectors is determined, we can select the optimum parameter s from a candidate set using the proposed method. The optimum parameter ensures that the Gaussian KDA algorithm performs well in dimensionality reduction, which improves the accuracy of protein subcellular location prediction. The proposed method for selecting the Gaussian kernel parameter can be presented in Table 5.

Table 5. The Method for Selecting the Gaussian KDA Parameter.

```
Input: A reasonable candidate set $S = \{s_1, s_2, \cdots, s_m\}$ for Gaussian kernel parameter, $X = \{X_1, X_2, \cdots, X_C\}$,
the training set $X_i = \{x_1, x_2, \cdots, x_{N_i}\}$ $(1 \leq i \leq C)$, the number of retained eigenvectors d.

1. Get the internal sample set $\Omega_{in}$ and the edge sample set $\Omega_{ed}$ from the training set $X_i$ using Algorithm 1.
2. For each parameter $s_i \in S, \quad i = 1, 2, \cdots, m$
    • Calculate the kernel matrix K using Equation (7).
    • Reduce dimension of the K using the Gaussian KDA algorithm.
    • Calculate $RE(\Omega_{ed})$ and $RE(\Omega_{in})$ using Equation (8).
    • Calculate the value of objective function $f(s_i)$ using Equation (12).
3. Select the optimum parameter $s = \underset{s_i \in S}{\mathrm{argmax}} f(s_i)$

Output: the optimum Gaussian kernel parameter S.
```

As the end of this section, we want to summarize the position of the proposed method in protein subcellular localization once more. First, two kinds of regularization forms of PSSM are used to extract the features in protein amino acid sequences. Then, the KDA method is performed on the extracted features for dimension reduction and discriminant analysis according to the KDA algorithm principle in Section 3.3 with formulas (1)–(6). During the procedure of KDA, the novelty of our work is to give a new method for selecting the Gaussian kernel parameter, which is summarized in Table 5. Finally, we choose the k-nearest neighbors (KNN) as the classifier to cluster the dimension-reduced data after KDA.

4. Materials

In this section, we introduce the other processes in Figure 5 except KDA model and its parameter selection, which are necessary materials for the whole experiment.

4.1. Standard Data Sets

In this paper, we use two standard datasets that have been widely used in the literature for Gram-positive and Gram-negative subcellular localizations [13], whose protein sequences all come from the Swiss-Prot database.

For the Gram-positive bacteria, the standard data set we found in the literature [13,14,21] is publicly available on http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi/Data.htm. There are 523 locative protein sequences in the data set that are distributed in four different subcellular locations. The number of proteins in each location is given in Table 6.

Table 6. The name and the size of each location for the Gram-positive data set.

No.	Subcellular Localization	Number of Proteins
1	cell membrane	174
2	cell wall	18
3	cytoplasm	208
4	extracell	123

For the Gram-negative bacteria, the standard data set of subcellular localizations is presented in the literature [13,22], which can be downloaded freely from http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi/Data.htm. The data set contains 1456 locative protein sequences located in eight different subcellular locations. The number of proteins in each location is shown in Table 7.

Table 7. The name and the size of each location for the Gram-negative data set.

No.	Subcellular Localization	Number of Proteins
1	cytoplasm	410
2	extracell	133
3	fimbrium	32
4	flagellum	12
5	inner membrane	557
6	nucleoid	8
7	outer membrane	124
8	periplasm	180

4.2. Feature Expressions and Sample Sets

In the prediction of protein subcellular localizations with machine learning methods, feature expressions are important information extracted from protein sequences, which have certain proper mathematical algorithms. There are many efficient algorithms used to extract features of protein sequences, in which two of them, PsePSSM [12] and PSSM-S [13], are used in this paper. The two methods rely on the position-specific scoring matrix (PSSM) for benchmarks which is obtained by using the PSI-BLAST algorithm to search the Swiss-Prot database with the parameter E-value of 0.01. The PSSM is defined as follows [12]:

$$\mathrm{P}_{\mathrm{PSSM}} = \begin{bmatrix} \mathrm{M}_{1\to1} & \mathrm{M}_{1\to2} & \cdots & \mathrm{M}_{1\to20} \\ \mathrm{M}_{2\to1} & \mathrm{M}_{2\to2} & \cdots & \mathrm{M}_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{M}_{\mathrm{i}\to1} & \mathrm{M}_{\mathrm{i}\to2} & \cdots & \mathrm{M}_{\mathrm{i}\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{M}_{\mathrm{L}\to1} & \mathrm{M}_{\mathrm{L}\to2} & \cdots & \mathrm{M}_{\mathrm{L}\to20} \end{bmatrix} \tag{13}$$

where $\mathrm{M}_{\mathrm{i}\to\mathrm{j}}$ represents the score created in the case when the ith amino acid residue of the protein sequence is transformed to the amino acid type j during the evolutionary process [12].

Note that, usually, multiple alignment methods are used to calculate PSSM, whose chief drawback is being time-consuming. The reason why we select PSSM instead of simple multiple alignment in this paper to form the total normalized information content is as follows. First, since our focus is to demonstrate the effectiveness of dimensional reduction algorithm, we need to construct high-dimensional feature expressions such as PsePSSM and PSSM-S, whose dimensions are as high as 1000 and 220, respectively. Second, PSSM has many advantages, such as those described in [23]. As far as the information features are concerned, PSSM has produced the strongest discriminator feature between fold members of protein sequences. Multiple alignment methods are used to calculate PSSM, whose chief drawback is being time-consuming. However, in spite of the time-consuming nature of constructing a PSSM for the new sequence, the extracted feature vectors from PSSM are so informative that are worth the cost of their preparation [23]. Besides, for a new protein sequence, we only need to construct a PSSM for the first time, which could be used repeatedly in the future for producing new normalization forms such as PsePSSM and PSSM-S.

4.2.1. Pseudo Position-Specific Scoring Matrix (PsePSSM)

Let P be a protein sample, whose definition of PsePSSM is given as follows [12]:

$$\mathrm{P}^{\xi}_{\mathrm{Pse-PSSM}} = \left[\overline{\mathrm{M}}_1\overline{\mathrm{M}}_2\cdots\overline{\mathrm{M}}_{20}\mathrm{G}^{\xi}_1\mathrm{G}^{\xi}_2\cdots\mathrm{G}^{\xi}_{20}\right]^{\mathrm{T}}(\xi = 0,1,2,\cdots,49) \tag{14}$$

$$\overline{\mathrm{M}}_{\mathrm{j}} = \frac{1}{\mathrm{L}}\sum_{\mathrm{i}=1}^{\mathrm{L}} \mathrm{M}_{\mathrm{i}\to\mathrm{j}} (\mathrm{j} = 1, 2, \cdots, 20) \tag{15}$$

$$\mathrm{G}_{\mathrm{j}}^{\xi} = \frac{1}{\mathrm{L}-\xi}\sum_{\mathrm{i}=1}^{\mathrm{L}-\xi} \left[\mathrm{M}_{\mathrm{i}\to\mathrm{j}} - \mathrm{M}_{(\mathrm{i}+\xi)\to\mathrm{j}}\right]^2 (\mathrm{j} = 1, 2, \cdots, 20; \xi < \mathrm{L}) \tag{16}$$

where L is the length of P, $\mathrm{G}_{\mathrm{j}}^{\xi}$ is the correlation factor by coupling the ξ-most contiguous scores [22]. According to the definition of PsePSSM, a protein sequence can be represented by a 1000-dimensional vector.

4.2.2. PSSM-S

Dehzangi et al. [13] put forward a new feature extraction method, PSSM-S, which combines four components: AAO, PSSM-AAO, PSSM-SD, and PSSM-SAC. According to the definition of the PSSM-S, it can be represented a feature vector with 220 (20 + 20 + 80 + 100) elements.

4.2.3. Sample Sets

For the two benchmark data, PsePSSM and PSSM-S are used to extract features, respectively. Finally we get four experimental sample sets GN-1000, GN-220, GP-1000 and GP-220, shown in Table 8.

Table 8. Sample sets.

Sample Sets	Benchmarks for Subcellular Locations	Extraction Feature Method	The Number of Classes	The Dimension of Feature Vector	The Number of Samples
GN-1000	Gram-negative	PsePSSM	8	1000	1456
GN-220	Gram-negative	PSSM-S	8	220	1456
GP-1000	Gram-positive	PsePSSM	4	1000	523
GP-220	Gram-positive	PSSM-S	4	220	523

4.3. Evaluation Criterion

To evaluate the performance of the proposed method, we use Jackknife cross-validation, which has been widely used to predict protein subcellular localization [13]. The Jackknife test is the most objective and rigorous cross-validation procedure in examining the accuracy of a predictor, which has been used increasingly by investigators to test the power of various predictors [24,25]. In the Jackknife test (also known as leave-one-out cross-validation), every protein is removed one-by-one from the training dataset, and the predictor is trained by the remaining proteins. The isolated protein is then tested by the trained predictor [26]. Let x be a sample set with N samples. For each sample, it will be used as the test data, and the remaining $\mathrm{N}-1$ samples will be used to construct the training set [27]. In addition, we use some criterion to assess the experimental results, defined as follows [12]:

$$\mathrm{MCC(k)} = \frac{\mathrm{TP_k} \times \mathrm{TN_k} - \mathrm{FN_k} \times \mathrm{FP_k}}{\sqrt{(\mathrm{TP_k}+\mathrm{FN_k})(\mathrm{TP_k}+\mathrm{FP_k})(\mathrm{TN_k}+\mathrm{FP_k})(\mathrm{TN_k}+\mathrm{FN_k})}} \times 100\% \tag{17}$$

$$\mathrm{Sen(k)} = \frac{\mathrm{TP_k}}{\mathrm{TP_k}+\mathrm{FN_k}} \times 100\% \tag{18}$$

$$\mathrm{Spe(k)} = \frac{\mathrm{TN_k}}{\mathrm{TP_k}+\mathrm{FP_k}} \times 100\% \tag{19}$$

$$\mathrm{Q} = \frac{\sum_{\mathrm{k}=1}^{\mathrm{C}} \mathrm{TP_k}}{\mathrm{N}} \times 100\% \tag{20}$$

where TP is the number of true positive, TN is the number of true negative, FP is the number of false positive, and FN is the number of false negative [12]. The value of MCC (Matthews coefficient correlation) varies between -1 and 1, indicating when the classification effect goes from a bad to a good one. The values of Specificity (Spe), sensitivity (Sen), and the overall accuracy (Q) all vary between 0 and 1, and the classification effect is better when their values are closer to 1, while the classification effect is worse when their values are closer to 0 [13].

4.4. The Grid Searching Method Used as Contrast

In this section, we introduce a normal algorithm for searching S, the grid-searching algorithm, which is used as a contrast with the proposed algorithm in Section 3.4.

The grid-searching method is usually used to select the optimum parameter, whose steps are as follows for the candidate parameter set S [28].

- Compute the kernel matrix k for each parameter $s_i \in S$, $i = 1, 2, \cdots, m$.
- Use the Gaussian KDA to reduce the dimension of K.
- Use the KNN algorithm to classify the reduced dimensional samples.
- Calculate the classification accuracy.
- Repeat the above four steps until all parameters in S have been traversed. The parameter corresponding to the highest classification accuracy is selected as the optimum parameter.

5. Conclusions

Biological data is usually high-dimensional. As a result, it is necessary to reduce dimension to improve the accuracy of the protein subcellular localization prediction. The kernel discriminant analysis (KDA) based on Gaussian kernel function is a suitable algorithm for dimensional reduction in such applications. As is known to all, the selection of a kernel parameter affects the performance of KDA, and thus it is important to choose the proper parameter that makes this algorithm perform well. To handle this problem, we propose a method of the optimum kernel parameter selection, which relies on reconstruction error [15]. Firstly, we use a method to select the edge and internal samples of the training set. Secondly, we compute the reconstruction errors of the selected samples. Finally, we select the optimum kernel parameter that makes the objective function maximum.

The proposed method is applied to the prediction of protein subcellular locations for Gram-negative bacteria and Gram-positive bacteria. Compared with the grid-searching method, the proposed method gives higher efficiency and performance.

Since the performance of the proposed method largely depends on the selection of the internal and edge samples, in the future study, researchers may pay more attention to select more representative internal and edge samples from the biological data set to improve the prediction accuracy of protein subcellular localization. Besides this, it is also meaningful to research how to further improve the proposed method to make it suitable for selecting parameters of other kernels.

Acknowledgments: This research is supported by grants from National Natural Science Foundation of China (No. 11661081, No. 11561071 and No. 61472345) and Natural Science Foundation of Yunnan Province (2017FA032).

Author Contributions: Shunfang Wang and Bing Nie designed the research and the experiments. Wenjia Li, Dongshu Xu, and Bing Nie extracted the feature expressions from the standard data sets, and Bing Nie performed all the other numerical experiments. Shunfang Wang, Bing Nie, Kun Yue, and Yu Fei analyzed the experimental results. Shunfang Wang, Bing Nie, Kun Yue, and Yu Fei wrote this paper. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Chou, K.C. Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Mol. Biosyst.* **2013**, *9*, 1092–1100. [CrossRef] [PubMed]
2. Zhang, S.; Huang, B.; Xia, X.F.; Sun, Z.R. Bioinformatics Research in Subcellular Localization of Protein. *Prog. Biochem. Biophys.* **2007**, *34*, 573–579.

3. Zhang, S.B.; Lai, J.H. Machine Learning-based Prediction of Subcellular Localization for Protein. *Comput. Sci.* **2009**, *36*, 29–33.
4. Huh, W.K.; Falvo, J.V.; Gerke, L.C.; Carroll, A.S.; Howson, R.W.; Weissman, J.S.; O'Shea, E.K. Global analysis of protein localization in budding yeast. *Nature* **2003**, *425*, 686–691. [CrossRef] [PubMed]
5. Dunkley, T.P.J.; Watson, R.; Griffin, J.L.; Dupree, P.; Lilley, K.S. Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteom.* **2004**, *3*, 1128–1134. [CrossRef] [PubMed]
6. Hasan, M.A.; Ahmad, S.; Molla, M.K. Protein subcellular localization prediction using multiple kernel learning based support vector machine. *Mol. Biosyst.* **2017**, *13*, 785–795. [CrossRef] [PubMed]
7. Teso, S.; Passerini, A. Joint probabilistic-logical refinement of multiple protein feature predictors. *BMC Bioinform.* **2014**, *15*, 16. [CrossRef] [PubMed]
8. Wang, S.; Liu, S. Protein Sub-Nuclear Localization Based on Effective Fusion Representations and Dimension Reduction Algorithm LDA. *Int. J. Mol. Sci.* **2015**, *16*, 30343–30361. [CrossRef] [PubMed]
9. Baudat, G.; Anouar, F. Generalized Discriminant Analysis Using a Kernel Approach. *Neural Comput.* **2000**, *12*, 2385–2404. [CrossRef] [PubMed]
10. Zhang, G.N.; Wang, J.B.; Li, Y.; Miao, Z.; Zhang, Y.F.; Li, H. Person re-identification based on feature fusion and kernel local Fisher discriminant analysis. *J. Comput. Appl.* **2016**, *36*, 2597–2600.
11. Xiao, Y.C.; Wang, H.G.; Xu, W.L.; Miao, Z.; Zhang, Y.; Hang, L.I. Model selection of Gaussian kernel PCA for novelty detection. *Chemometr. Intell. Lab.* **2014**, *136*, 164–172. [CrossRef]
12. Chou, K.C.; Shen, H.B. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* **2007**, *360*, 339–345. [CrossRef] [PubMed]
13. Dehzangi, A.; Heffernan, R.; Sharma, A.; Lyons, J.; Paliwal, K.; Sattar, A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. *J. Theor. Biol.* **2015**, *364*, 284–294. [CrossRef] [PubMed]
14. Shen, H.B.; Chou, K.C. Gpos-PLoc: An ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Eng. Des. Sel.* **2007**, *20*, 39–46. [CrossRef] [PubMed]
15. Hoffmann, H. Kernel PCA for novelty detection. *Pattern Recogn.* **2007**, *40*, 863–874. [CrossRef]
16. Li, Y.; Maguire, L. Selecting Critical Patterns Based on Local Geometrical and Statistical Information. *IEEE Trans. Pattern Anal.* **2010**, *33*, 1189–1201.
17. Wilson, D.R.; Martinez, T.R. Reduction Techniques for Instance-Based Learning Algorithms. *Mach. Learn.* **2000**, *38*, 257–286. [CrossRef]
18. Saeidi, R.; Astudillo, R.; Kolossa, D. Uncertain LDA: Including observation uncertainties in discriminative transforms. *IEEE Trans. Pattern Anal.* **2016**, *38*, 1479–1488. [CrossRef] [PubMed]
19. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.* **2010**, *31*, 651–666. [CrossRef]
20. Li, R.L.; Hu, Y.F. A Density-Based Method for Reducing the Amount of Training Data in kNN Text Classification. *J. Comput. Res. Dev.* **2004**, *41*, 539–545.
21. Chou, K.C.; Shen, H.B. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Sci.* **2010**, *2*, 1090–1103. [CrossRef]
22. Chou, K.C.; Shen, H.B. Large-Scale Predictions of Gram-Negative Bacterial Protein Subcellular Locations. *J. Proteome Res.* **2007**, *5*, 3420–3428. [CrossRef] [PubMed]
23. Kavousi, K.; Moshiri, B.; Sadeghi, M.; Araabi, B.N.; Moosavi-Movahedi, A.A. A protein fold classifier formed by fusing different modes of pseudo amino acid composition via PSSM. *Comput. Biol. Chem.* **2011**, *35*, 1–9. [CrossRef] [PubMed]
24. Shen, H.B.; Chou, K.C. Nuc-PLoc: A new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Eng. Des. Sel.* **2007**, *20*, 561–567. [CrossRef] [PubMed]
25. Wang, T.; Yang, J. Using the nonlinear dimensionality reduction method for the prediction of subcellular localization of Gram-negative bacterial proteins. *Mol. Divers.* **2009**, *13*, 475.
26. Wei, L.Y.; Tang, J.J.; Zou, Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform. Sci.* **2017**, *384*, 135–144. [CrossRef]

27. Shen, H.B.; Chou, K.C. Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *J. Theor. Biol.* **2010**, *264*, 326–333. [CrossRef] [PubMed]
28. Bing, L.I.; Yao, Q.Z.; Luo, Z.M.; Tian, Y. Gird-pattern method for model selection of support vector machines. *Comput. Eng. Appl.* **2008**, *44*, 136–138.