Investigating the SVM method with four different kernels including: MLP, RBF, Adaptive wavelet and Gaussian wavelet

Abstraction

A support vector machine (SVM) is a supervised learning machine method. This method is categorized between pattern recognition algorithms. It can be used to control unmanned aircraft, route simulation, welding quality analysis, computer quality analysis, machine maintenance analysis, loan risk estimation and many other applications. SVM algorithms use a set of mathematical functions that are defined as kernels. The task of the kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. In this project, the SVM method with four different kernels is investigated. These kernels are MLP, RBF, Adaptive wavelet and Gaussian wavelet kernel also a method which uses a combination structure is evaluated. This method uses four mentioned kernels.

In this work, to evaluate the performances of these methods confusion and confidence matrices have been used, which show the level of accuracy and confidence in decisions.

Keywords: SVM, MLP, RBF, Adaptive Wavelet, Gaussian wavelet.

1. Introductions

SVM is a supervised machine learning algorithm that separates data samples represented as points in space using a line or hyper-plane. This separation is such that the data points are on the same side of the line are similar to each other and are placed in the same group. New data samples will be placed in one of the existing categories after being added to the same space.

In this section four SVM kernels methods are discussed briefly. These kernels are MLP ،RBF ،Adaptive wavelet and Gaussian wavelet.

1-1 MLP

MLP function is as below. The determining parameters are p1 and p2.

$$K(U,V) = \tanh(p_1 U V^T + p_2)$$

1-2 RBF

RBF kernel function parameter is σ and RBF equation is:

$$K(U,V) = \exp\left(-\frac{||U - V||^2}{2\sigma^2}\right)$$

1-3 Adaptive Wavelet

In adaptive wavelet, the parameters train according to figure 1.

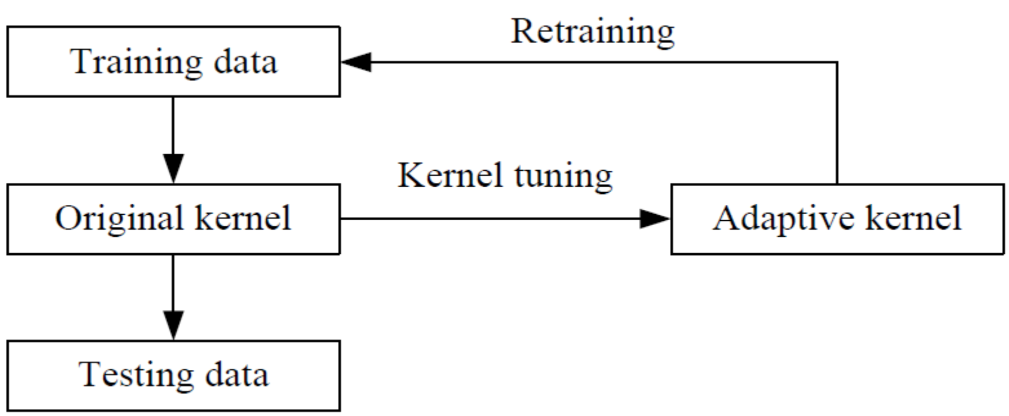

Figure 1. Adaptive kernel training steps

The adaptive kernel function is generally obtained from the following relation:

$$\mathrm{k}^*(\mathrm{x},\mathrm{x'}) = \sum_{i\in sv} a_i e^{\frac{-\|\mathrm{x}-\mathrm{x_i}\|^2}{\mathrm{d_i^2}}} \sum_{i\in sv} a_i e^{\frac{-\|\mathrm{x'}-\mathrm{x_i}\|^2}{\mathrm{d_i^2}}} \prod_{\mathrm{i}=1}^{\mathrm{n}} \left(1 - \frac{\|\mathrm{x}-\mathrm{x'}\|^2}{\mathrm{a_i}}\right) \exp\left(-\frac{\|\mathrm{x}-\mathrm{x'}\|^2}{2\mathrm{a_i}}\right)$$

Which a_i is the Lagrange multiplier of the i-th support vector and the data density around the i-th support vector is defined as follows:

$$d_i = \frac{\sum_a \|x_a - x_i\|}{k}$$

Based on this formula, k is the closest data number and a is the location of the closest data.

1-4 Gaussian Wavelet

For all x parameters, bellow relations are established:

$$(x) = \prod_{i=1}^{d} \Psi(\frac{x_i}{a}) = \prod_{i=1}^{d} (-1)^{p/2} [\exp(-\frac{\|x_i\|^2}{a^2})]^p$$

$$\int \exp\left(-j(\omega x)\right) \prod_{i=1}^{d} (-1)^{\frac{p}{2}} \left[\exp\left(-\frac{\|x_i\|^2}{a^2}\right)\right]^p dx$$

$$= \prod_{i=1}^{d} \int_{-\infty}^{\infty} \exp\left(-j(\omega x_i)\right) (-1)^{\frac{p}{2}} \left[\exp\left(-\frac{\|x_i\|^2}{a^2}\right)\right]^p dx$$

$$= \prod_{i=1}^{d} (-1)^{\frac{p}{2}} (j\omega)^p \int_{-\infty}^{\infty} \exp\left(-j(\omega x_i)\right) \left[\exp\left(-\frac{\|x_i\|^2}{a^2}\right)\right]^p dx$$

$$= \prod_{i=1}^{d} (-1)^{\frac{p}{2}} (j\omega)^p |a| \sqrt{\pi} \exp\left(-\frac{a^2\omega^2}{4}\right) >$$

2. Performance evaluation

In this section, the performances of these four methods are investigated. The kernel functions that are examined here are RBF, MLP and adaptive wavelet and Gaussian wavelet. Each of the kernel functions is highlighted. Firstly, the optimum solution for choices by using the grid search method is calculated and then their performances are compared.

The methods are examined in order, and for each of them the confusion matrix, confidence matrix, the required time for the test and training data, histograms and two- dimensional and three-dimensional scatters are obtained separately. At the end, the four proposed methods are compared and a combined method which uses these four kernels is presented and evaluated.

2-1 MLP

To check the performance of the classifier, the test data is used. The confusion matrix after normalization is obtained as follows:

Table 1. Normalized confusion matrix of MLP method

decision / real	ω1	ω2	ω3	ω4	ω5	ω6	ω7
ω1	0.9909	0	0	0	0.0091	0	0
ω2	0	1	0	0	0	0	0
ω3	0.0364	0	0.8	0.0455	0.1182	0.0364	0
ω4	0.0364	0	0.0182	0.5364	0.0455	0.3636	0
ω5	0.0364	0	0.2182	0.0818	0.6636	0	0
ω6	0	0	0	0.03636	0	0.9636	0
ω7	0	0	0	0	0	0	1

To calculate the CCR criterion, we calculate the ratio of the total diagonal elements to the total data:

$CCR = 0.8506$

$P(Error) = 1 - CCR = 0.1494$

It can be seen that the diagonal elements in these two matrices have more value than the others, but some non-diagonal elements such as (5,3) also have a large value and this shows that the data of classes 3 and 5 aren't correct. Also for element (4,6) is the same problem and it shows that the class 6 data is mostly mistaken for class 4 data.

The confidence matrix for this method is shown in the table below:

Table 2. The confidence matrix of the MLP method

decision / real	ω1	ω2	ω3	ω4	ω5	ω6	ω 7
ω1	0.6865	0	0	0.0714	0	0	0
ω2	0	0.7598	0	0	0	0	0
ω3	0	0	0.6846	0.0714	0	0	0
ω4	0	0	0.0857	0.7248	0.1270	0.1429	0
ω5	0	0	0	0	0.588	0	0
ω6	0	0	0.1071	0.01036	0	0.7040	0
ω٧	0	0	0	0	0	0	0.8096

The confidence matrix shows how confidently the classifier has classified the data, each element shows a confidence level that the data of class i is correctly classified in class j.

To calculate each element of the confidence matrix, we should add the confidence value of each data of class i which is classified in class j together and assign it to the desired level. The confidence level of each data is equal to the difference between the first maximum of the class in which the data is classified and the second maximum of the class in which it would be classified in the absence of the first class. The closer the unit is, the closer it is to the ideal state.

According to table 2, some non-diagonal elements such as (6,3) have a high value, which indicates that the wrong decision was made.

The required time for training and testing is shown in table 3.

Table 3. The required time for training and testing in MPL

	Training time	Testing time
MLP	2.0695 sec	14.3574sec

According to the forward selection algorithm, the priority of the selected features is as follows. In this program, the selected features are as follows:

Table 4. The order of selected features in each step in the MLP method

Step	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
# new feature selected in this step	11	2	13	19	6	10	3	4	16	12	14	15	5	8	17	1	18	9	7

So feature number 11 is used for the histogram diagram, features number 11 and 2 for a two-dimensional scatter chart, and features number 11, 2 and 13 for a three-dimensional scatter chart are used.

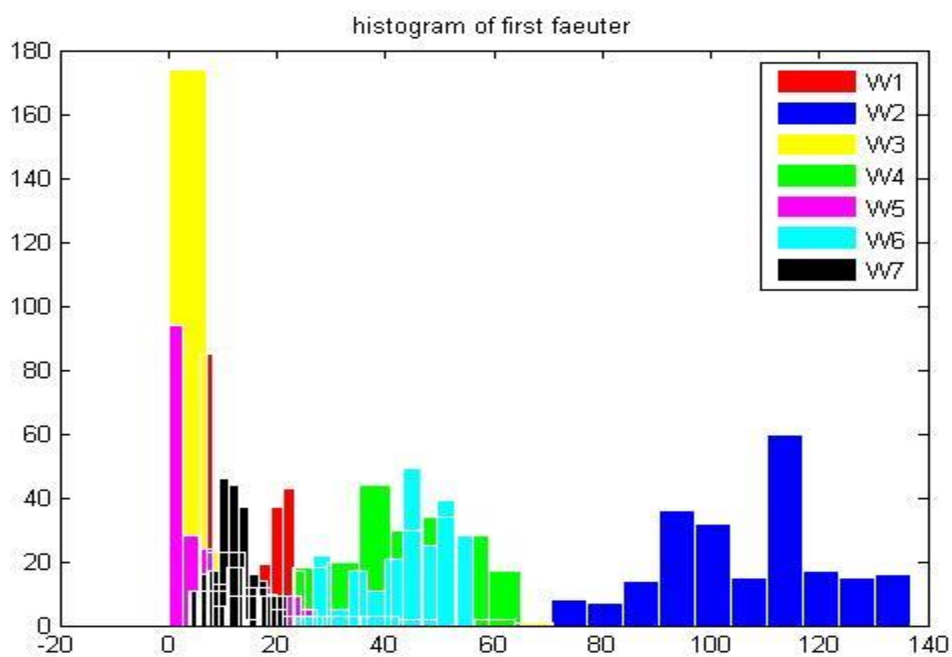

Figure 2. Histogram chart of MLP method

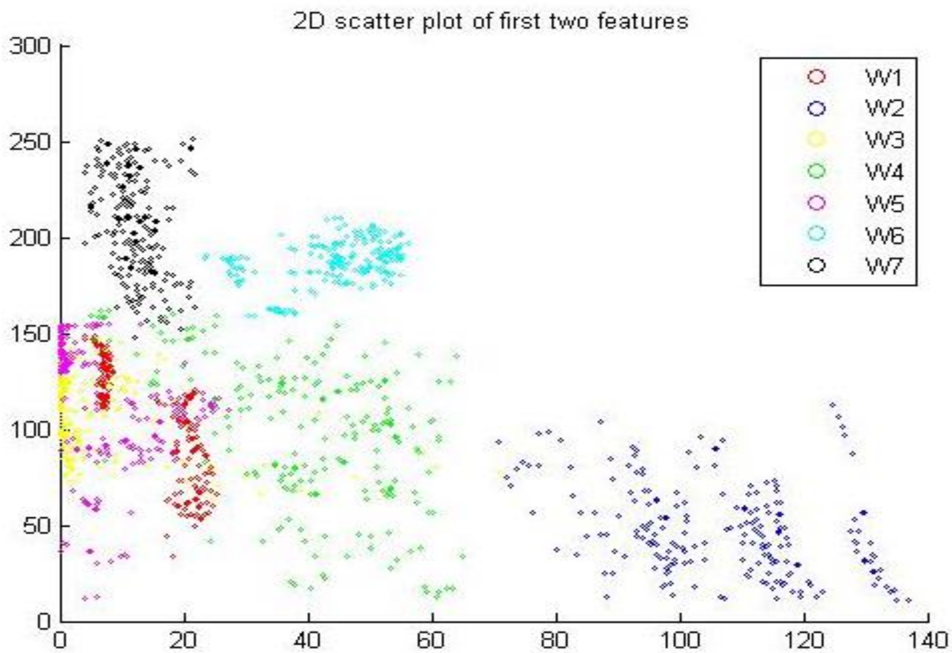

Figure 3. Two-dimensional scatter diagram of MLP method

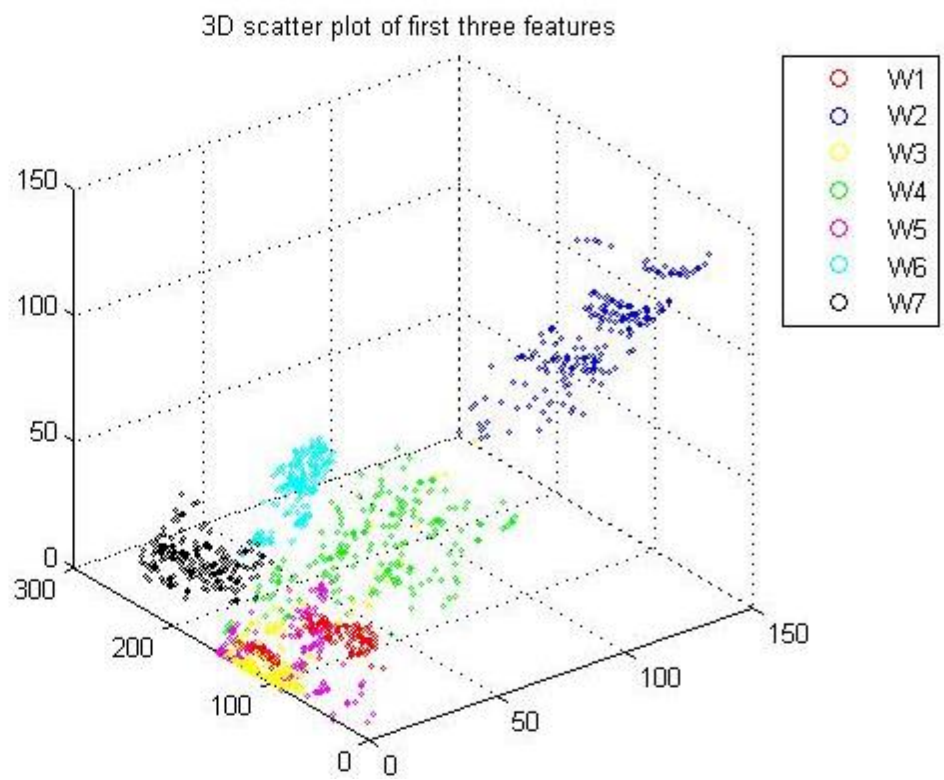

Figure 4. 3D scatter chart of MLP method

The above three diagrams fully justify the confusion matrix. According to the graphs, the characteristics of the three classes 3 and 5 overlap, so it is difficult to distinguish them from each other, as well as overlapping classes 4 and 6. But the graphs also show that the characteristics of class two do not overlap with other classes, and the confusion matrix also confirms the correctness of this content.

2-2 RBF

The normalized confusion matrix for RBF method is obtained as follows:

Table 5. Normalized confusion matrix for RBF

decision / real	ω1	ω2	ω3	ω4	ω5	ω6	ω7
ω1	0.9636	0	0	0.0182	0.0182	0	0
ω2	0	0.9818	0.0091	0.0091	0	0	0
ω3	0.0091	0	0.9182	0	0.0727	0	0
ω4	0	0.1364	0.1636	0.6091	0.0727	0.0182	0
ω5	0	0	0.2	0.0364	0.7637	0	0
ω6	0	0	0.0182	0.0636	0	0.9182	0
ω7	0	0	0.0091	0	0	0	0.9909

$CCR = 0.8779$

$P(Error) = 1 - CCR = 0.1221$

The histogram and scatter diagrams shown in the previous method justify the confusion matrix. Also, the confidence matrix and the required time for this method is shown in the following tables:

Table 6. The confidence matrix of the RBF method

decision / real	ω1	ω2	ω3	ω4	ω5	ω6	ω7
ω1	0.8048	0	0.1429	0	0	0	0
ω2	0	0.6570	0	0.1429	0	0	0
ω3	0	0	0.7397	0	0.0519	0	0
ω4	0	0	0	0.555	0.0357	0	0
ω5	0.1429	0	0.0714	0.0893	0.7522	0	0
ω6	0	0	0	0	0	0.5586	0
ω7	0	0	0	0	0	0	0.5176

Table 7. The required time for training and testing in RBF

	Training time	Testing time
RBF	2.0353sec	20.0783sec

2-3 Adaptive wavelet

The normalized confusion matrix for the adaptive wavelet method and CCR is obtained as follows:

Table 8. The normalized confusion matrix for adaptive wavelet

decision / real	ω1	ω2	ω3	ω4	ω5	ω6	ω7
ω1	0.6091	0	0.0273	0.0273	0.3364	0	0
ω2	0	1	0	0	0	0	0
ω3	0.0364	0	0.7364	0.0273	0.2	0	0
ω4	0.0091	0	0.0818	0.7091	0.1545	0.0364	0
ω5	0.0182	0	0.02364	0.0091	0.7364	0	0
ω6	0.0091	0	0.0182	0.3909	0	0.8818	0
ω7	0	0	0.0091	0	0	0	0.9909

$CCR = 0.9262$

$P(Error) = 1 - CCR = 0.0733$

The elements of this matrix in (1,5) and (6,4) are larger than other non-diagonal elements obviously which represents data in these classes are mistaken. Also, the confidence matrix for this method and the required time for training and testing are as follows:

Table 9. The confidence matrix of the adaptive wavelet method

decision / real	ω1	ω2	ω3	ω4	ω5	ω6	ω7
ω1	0.7997	0	0.0388	0.0211	0.0287	0	0
ω2	0	0.8309	0	0	0	0	0
ω3	0.003	0	0.6436	0.003	0.003	0	0
ω4	0.003	0.003	0.003	.7292	0.003	0.003	0
ω5	0.003	0	0.003	0.003	0.5953	0	0
ω6	0.003	0	0.003	0.003	0	0.6492	0
ω7	0	0	0.003	0	0	0	0.7394

Table 10. The required time for training and testing in adaptive wavelet

	Training time	Testing time
adaptive wavelet	208.0118 sec	1.3539sec

2-4 Gaussian wavelet

The normalized confusion matrix, CCR criterion, the confidence matrix and the required time for training and testing are presented below:

Table 11. The normalized confusion matrix of Gaussian wavelet

decision / real	ω1	ω2	ω3	ω4	ω5	ω6	ω7
ω1	0.6545	0	0.0364	0	0.3091	0	0
ω2	0	0.998	0	0	0	0	0
ω3	0.0545	0.0091	0.7818	0.0182	0.1364	0	0
ω4	0.0273	0.0364	0.0545	0.5273	0.2636	0.0909	0
ω5	0.0818	0	0.2182	0	0.7	0	0
ω6	0.0091	0	0.0273	0.2364	0.0091	0.7182	0
ω7	0	0	0	0	0.0273	0	0.9727

$CCR = 0.9246$

$P(Error) = 1 - CCR = 0.0754$

Table 12. The confidence matrix of Gaussian wavelet

decision / real	ω1	ω2	ω3	ω4	ω5	ω6	ω7
ω1	0.5752	0	0.240	0	0.0414	0	0
ω2	0	0.7093	0	0	0	0	0
ω3	0.0026	0.0026	0.6642	0.0026	0.0026	0	0
ω4	0.0026	0.0026	0.0026	0.6439	0.0026	0.0026	0
ω5	0.0026	0	0.0026	0	0.6830	0	0
ω6	0.0026	0	0.0026	0.0026	0	0.7529	0
ω7	0	0	0	0	0.0026	0	0.6252

Table 13. The required time for training and testing for Gaussian wavelet

	Training time	Testing time
Gaussian wavelet	741.2517sec	0.1586sec

2-5 Comparison of four methods

In the table below, four methods are compared in terms of the time required for test and training data and the CCR rate.

Table 14. The comparison of four methods in terms of CCR rate and the required time for testing and training data

method	Testing time (sec)	Training time (sec)	CCR
RBF	1.7783	2.0353	0.8779
MLP	1.3574	2.0695	0.8506
Gaussian wavelet	5.1586	741.2517	0.9246
Adaptive wavelet	4.3539	208.0118	0.9262

It can be seen that the adaptive wavelet method has the highest CCR. The RBL and MLP methods take relatively less training time compared to Gaussian and Adaptive wavelet methods so maybe the adaptive wavelet method is the right choice in case that training time is not important for us.

2-6 Combination of four methods

After implementing four classification methods by using SVM, to achieve a better quality classification, combining these methods is evaluated. For this purpose, each test data is classified with the use of these four methods to determine the desired class of data, then the desired place is chosen as the class that has received the most votes among the four classification methods. In this simulation, if the numbers of votes for classes are equal, the class will be chosen randomly.

The normalized confusion matrix and CCR for a combination of methods are presented in below table:

Table 15. The normalized confusion matrix for combination of methods

Decision / real	ω1	ω2	ω3	ω4	ω5	ω6	ω7
ω1	0.9909	0	0	0	0.0091	0	0
ω2	0	1	0	0	0	0	0
ω3	0.0273	0	0.8455	0.0273	0.1	0	0
ω4	0.0273	0.0182	0.0636	0.8818	0.0545	0.0545	0
ω5	0.0273	0	0.2364	0.0455	0.8909	0	0
ω6	0.0091	0	0.0182	0.0455	0	0.9273	0
ω7	0	0	0.0091	0	0	0	0.9909

$CCR = 0.9866$

$P(Error) = 1 - CCR = 0.0134$

According to the CCR value, this combined method has performed better than any of the methods, but it should be noted that the execution time will be much longer in this combined method.

Conclusion

In this project, the SVM method with four different kernels (adaptive wavelet, RBF, MLP and Gaussian wavelet) are investigated. According to the results, the adaptive wavelet has the most amount of CCR because the kernel parameters are determined adaptively. The Gaussian and adaptive wavelet have better performance than RBF and MLP but based on performance time RBF and MLP are more efficient approaches. In addition the RBF method implements classification better than MLP. Also, to improve the classification performance, a combination of classifications has been used. This method has a lower CCR error than the cases where only one method is used, but the training and testing time is longer as expected.

References

[1] Andrew ng."*cs229 Lecture notes.* "Nature, pp. 378-406, (2000).
[2] The Wavelet Tutorial Part 1,
http://users.rowan.edu/~polikar/WAVELETS/WTpart1.html, 2014.

[3] Xinlu Zong1, Chunzhi Wang2 and Hui Xu3, “ Density-based Adaptive Wavelet Kernel SVM Model for P2P Traffic Classification”, *School of Computer Science and Technology, Hubei University of Technology*
Wuhan 430068, China.
[4] Yongkang Zheng,Li Zhu and Xing Zou, “Short-Term Load Forecasting Based on Gaussian Wavelet
SVM”, Chengdu, China, September 2011.
[5] https://en.wikipedia.org/wiki/Support_vector_machine.
[6] *http://www.support-vector.net/t*