Chin. J. Astron. Astrophys. Vol. 7 (2007), No. 3, 441–447
(*http://www.chjaa.org*)

Chinese Journal of
Astronomy and
Astrophysics

Support Vector Machine combined with K-Nearest Neighbors for Solar Flare Forecasting *

Rong Li, Hua-Ning Wang, Han He, Yan-Mei Cui and Zhan-Le Du

National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012; *lirong@bao.ac.cn*

Received 2006 September 1; accepted 2006 November 21

Abstract A method combining the support vector machine (SVM) the K-Nearest Neighbors (KNN), labelled the SVM-KNN method, is used to construct a solar flare forecasting model. Based on a proven relationship between SVM and KNN, the SVM-KNN method improves the SVM algorithm of classification by taking advantage of the KNN algorithm according to the distribution of test samples in a feature space. In our flare forecast study, sunspots and 10 cm radio flux data observed during Solar Cycle 23 are taken as predictors, and whether an M class flare will occur for each active region within two days will be predicted. The SVM-KNN method is compared with the SVM and Neural networks-based method. The test results indicate that the rate of correct predictions from the SVM-KNN method is higher than that from the other two methods. This method shows promise as a practicable future forecasting model.

Key words: Sun: flare — Sun: sunspot — Sun: activity — Sun: magnetic fields

1 INTRODUCTION

It is well known that the occurrence of solar X-ray flares is closely related to sunspots. So, a succession of flare forecasting methods based on this relationship has been proposed. McIntosh (1990) revised sunspot classification by categorizing sunspots group with modified Zurich class and two other parameters. Based mainly on the McIntosh classification, a specially dedicated system called Theophrastus was developed and adopted in 1987 as a tool in the daily operations of the Space Environment Services Center (McIntosh 1990). Gallagher et al. (2002) at Big Bear Solar Observatory developed a flare prediction system which estimated the probability for each active region to produce C-, M-, or X-class flares using historical averages of flare numbers according to the McIntosh classifications. At Beijing Astronomical Observatory, Zhang & Wang (1994) developed a multi-discrimination method for flare forecast by using observations of sunspots, 10 cm radio flux and longitudinal magnetic fields. Zhu & Wang (2003) presented a verification for this method. Recently, Wheatland (2004) suggested a Bayesian approach to flare prediction, in which the flaring record of an active region together with phenomenological rules of flare statistics are used to refine an initial prediction for the occurrence of a large flare during a subsequent period of time.

The methods mentioned above mainly rely on traditional statistical techniques. Neural networks (NN), as an important branch of artificial intelligence, has been applied to some space weather forecasting, such as geomagnetic storms forecasting (Lundstedt 1997) and proton event alert (Wang 2000; Gong et al. 2004). Without enough statistical theory support, NN's general ability is limited and a number of problems can be caused including overfitting and local minima in the back-propagation network (Vapnik 1995). Learning Vector Quantity (LVQ), as a new technique of NN, was evolved from the self organization feature map network. Unlike the traditional NN methods which minimize the empirical training error, LVQ is a method

* Supported by the National Natural Science Foundation of China.

based on reference points. It has the advantages, among others, of being easily fulfilled and a good generalization capability (Wu 2000). Meanwhile, support vector machine (SVM), proposed first by Vapnik (1995), has become a widely used technique of machine learning due to its strong basis in statistical theory and successful performance in various applications. Its algorithm has been applied to forecasting geomagnetic substorm, which demonstrates a promising performance in comparison with NN (Gavrishchaka & Ganguli 2001).

Even though the classifying ability of SVM is better than that of other pattern recognition methods, some problems still exist in its application, such as a low classifying accuracy in complicated applications and difficulty in choosing the kernel function parameters. In an attempt to solve these problems, a simple and effective improved SVM classifying algorithm was proposed by Li et al. (2002), which combines SVM with the K-nearest neighbor (KNN) classifier. This new algorithm has demonstrated to give excellent performance in various applications, especially in complicated ones (Li et al. 2002).

In Sections 2 and 3 we give an introduction to the SVM-KNN classifier and apply it to flare forecasting. In Section 4, a series of test results is presented and it is shown that SVM-KNN is better in performance than SVM or NN-based method. Our conclusions and a discussion are given in Section 5.

2 SVM-KNN ALGORITHM

2.1 SVM Algorithm

As a successful implementation of the structural risk minimization principle and Vapnik-Chervonenkis(VC) dimension theory, SVM aims at minimizing an upper bound of the generalization error through maximizing the margin between the separating hyperplane and the data (Amari & Wu 1999). The optimal hyperplane can be derived and represented in feature space by means of a kernel function which expresses the dot products between mapped pairs of input points: $K(x^{'}, x) = \sum_i \phi_i(x^{'})\phi_i(x)$ (Cristianini et al. 1999), where $\phi_i(x)$ is a nonlinear mapping from the input space to a higher-dimensional feature space.

Then supposing in the case where the data are linearly separable, for the training set $(x_1, y_1) \cdots (x_l, y_l)$ belonging to two different classes $y \in (-1, +1)$, the problem of searching for the optimal hyperplane amounts to finding the adjustable coefficients α_i that maximize the Lagrangian function with constraints:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j k(x_i \cdot x_j), \tag{1}$$

$$0 \leq \alpha_i, i = 1, \cdots, n \quad \text{and} \quad \sum_{i=1}^{l} \alpha_i y_i = 0. \tag{2}$$

Those sample points having $\alpha_i > 0$ are called support vector located near the hyperplane. The separating rule is the following discriminant function:

$$f(x) = \mathrm{sgn}\Big(\sum_{i=1}^{l} y_i \alpha_i k(x_i \cdot x) - b\Big). \tag{3}$$

2.2 KNN Algorithm

The 1-Nearest Neighbor(1NN) classifier is an important pattern recognizing method based on representative points (Bian et al. 2000). In the 1NN algorithm, whole train samples are taken as representative points and the distances from the test samples to each representative point are computed. The test samples have the same class label as the representative point nearest to them. The KNN is an extension of 1NN, which determines the test samples through finding the k nearest neighbors.

2.3 SVM-KNN Algorithm

First, by analyzing the classifying process of SVM, a relationship between SVM and 1NN is found. This relationship is the theoretical basis of SVM-KNN and will be expatiated in Theorem 1.

[Theorem 1] SVM classifier is equal to a 1NN classifier which chooses one representative point for the support vectors in each class (a detailed proof can be found in Appendix A).

We examined distributions of wrong samples of SVM and found that they are almost always near the separating hyperplane. This prompts us that the information of hyperplane area should be used as much as we can in order to improve the classifying accuracy. We know that samples lying near the separating hyperplane area are basically support vectors. Instead of using SVM algorithm in which only one representative point is chosen for the support vector in each class and this representative point can not represent efficiently the whole class, we use KNN to classify algorithm in this case, in which each support vector is taken as a representative point. That means more useful information can be utilized.

Specifically, for samples far from the separating hyperplane (Region II in Fig. 1), the SVM classifying algorithm is available, while for samples close to the hyperplane (Region I), the KNN classifying algorithm is suitable. The main steps of the new classifying algorithm are as follows:

$$
\begin{aligned}
&\text{step1 } \text{ if } T_{\text{test}} \neq \Phi, \text{get } x \in T_{\text{test}}, \text{if } T_{\text{test}} = \Phi, \text{ stop;} \\
&\text{step2 } \text{ calculate } g(x) = \sum_i y_i \alpha_i k(x_i, x) - b; \\
&\text{step3 } \text{ if } |g(x)| > \varepsilon, \text{ calculate directly } f(x) = \text{sgn}(g(x)) \text{ as output;} \\
&\qquad\quad \text{ if } |g(x)| < \varepsilon, \text{ put it into KNN algorithm to classify;} \\
&\text{step4 } \ T \leftarrow T - x, \text{ go to step1.}
\end{aligned}
$$

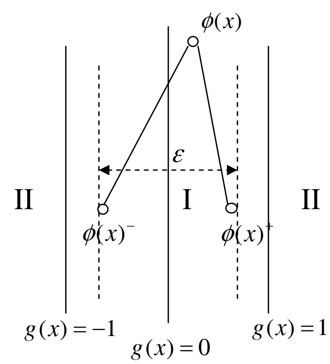

Fig. 1 The distances from the test sample $\phi(x)$ to two representative points $\phi(x)^+$ and $\phi(x)^-$ are calculated in a high dimension feature space, and the threshold ε and classifying algorithm are then decided.

In the steps described above, T_{test} refers to the test set and Φ represents the empty set. The distance threshold ε should satisfy $0 < \varepsilon < 1$. Note that distance used in this algorithm is calculated in a high dimension feature space. The distance formula used here is based on the kernel function and takes the following form:

$$\|\phi(x) - \phi(x_i)\|^2 = k(x, x) - 2k(x, x_i) + k(x_i, x_i). \tag{4}$$

3 APPLICATIONS

3.1 SVM-KNN Application Model

Applying the SVM-KNN algorithm to our problem of flare forecast is based on the understanding that this problem can be formalized to be one of pattern recognition. The input of the model includes the current daily data on solar active regions and the 10 cm radio flux data, which correspond to the feature vector $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ in Equation (3). The output refers to a classification of the importance class of solar flares occurring within the coming two days, and there will be two outcomes: larger than or equal to M if $f(x) = +1$ in Equation (3), or lower than M if $f(x) = -1$. In the training process for the SVM-KNN classifier, the inputs and outputs of all the samples are taken into Equation (1) to determine the coefficients α_i.

3.2 Data

The SEC solar active region data used in our forecast model span from 1996 January to 2004 December. The data were downloaded from the SEC web site: *http://sec.noaa.gov/ftpmenu/forecasts/SRS.html*. We count each observed active region on every day as one sample, and we have 19544 samples in total.

3.3 Predictors

In our study, the predictors including the area of the sunspot group, magnetic classification, McIntosh classification and 10 cm radio flux are divided into different groups, which were assigned numerical values (see Table 1) according to the relevant flare productivity (Zhang & Wang 1994).

Table 1 Classification and Flare Productivity Rates of Predictors

Area classification	$Sp \leq 200$	$200 < Sp \leq 500$	$500 < Sp \leq 1000$	$Sp > 1000$	Non-spot
Flare productivity rate	0.03	0.09	0.20	0.38	0.00
Magnetic classification	α	β	$\beta,\ \gamma$	$\beta\gamma\delta,\ \beta\delta,\ \delta(A)$	Non-spot
Flare productivity rate	0.05	0.20	0.34	0.47	0.00
McIntosh classification	[a] (a)	(b)	(c)	(d)	(e)
Flare productivity rate	0.00	0.08	0.31	0.68	0.81
10 cm radio flux level	[b] low	med	peak	fast	
Flare productivity rate	0.34	0.45	0.69	0.77	

[a] (a):non-spot; (b) Sunspot groups excluded from the MacIntosh classification; (c) Fso, Fko, Fri, Eac, Eko, Eao, Dhc, Dko, Dki, Dsc, Dac, Dho, Chi, Cko, Cki; (d) Fki, Fsi, Fai, Fhi, Fhc, Eki, Ehc, Ehi, Eai, Dkc; (e) Fkc, Ekc; [b] Low: Within 4 days before and after the minimum of the flux during the period of 27 days; Med: The medium period between the peak and the valley in the period of 27 days; Peak: Within 5 days before and after the maximum; Fast: The flux increases 15 sfu within 3 consecutive days.

4 TEST RESULTS

4.1 Test Methods and Parameter Setting

The data observed from 2001 January to 2004 December are grouped by year into four testing sets. Each of testing sets has a training set that begins in 1966 January and ends in December of the year before its beginning of the testing set .

Three methods are used and compared here: the SVM, the SVM-KNN, and the LVQ method. Our data contain far more non-flaring samples than flaring samples. Now LVQ requires the number of the two kinds of samples to be approximately equal, so we selected a random subset of the non flaring samples of the same size as the set of flaring samples. In our test, the three different classifying algorithms were applied to each of the testing sets.

For the LVQ algorithm, the number of initial reference points is set to be that of the flaring samples. For the SVM-KNN algorithm modified from LIBSVM described by Chang & Lin (2001), the distance threshold is set to be 0.8 and the number of nearest neighbors is set to be 1.0. The Gaussian Radial Basis function, given by $K(x, x_i) = \exp(-\frac{|x-x_i|^2}{\sigma^2})$, is used as the kernel function and its parameters were adjusted for optimal result separately in the SVM and SVM-KNN models. Test results for the 4 years are shown in Table 2.

4.2 Test Results

In Table 2, the first column, ‘Predic.’, gives the total number of predictions, the second column, ‘Observ.’, the total number of observations. The third column, ‘Equal’, is the number of correct predictions. The next two columns labeled ‘High’ and ‘Low’ are the number of false predictions: ‘High’ means that predicted class is larger than or equal to M when the observed class is lower than M, and conversely ‘Low’. The last three columns are ratios of ‘Equal’, ‘High’ and ‘Low’ to the total. As demonstrated in these tables, the SVM-KNN method gives the highest ‘Equal’ predictions and the lowest ‘High’ predictions among the three methods for all the four testing sets.

Table 2 Test Results by Three Methods for the Years 2001–2004

Year	Methods	Predic.	Observ.	Equal	High	Low	Equal(%)	High(%)	Low(%)
2001									
	LVQ	3461	3461	2925	432	104	84.52	12.48	3.00
	SVM	3461	3461	3054	266	141	88.24	7.69	4.07
	SVM-KNN	3461	3461	3110	144	207	89.86	4.16	5.98
2002									
	LVQ	3514	3514	2984	410	120	84.92	11.67	3.41
	SVM	3514	3514	3062	307	145	87.14	8.74	4.12
	SVM-KNN	3514	3514	3152	180	182	89.70	5.12	5.18
2003									
	LVQ	2139	2139	1861	213	65	87.00	9.96	3.04
	SVM	2139	2139	1893	171	75	88.50	7.99	3.51
	SVM-KNN	2139	2139	1953	85	101	91.30	3.98	4.72
2004									
	LVQ	1306	1306	1059	188	59	81.09	14.40	4.51
	SVM	1306	1306	1086	153	67	83.15	11.72	5.13
	SVM-KNN	1306	1306	1130	88	88	86.52	6.74	6.74

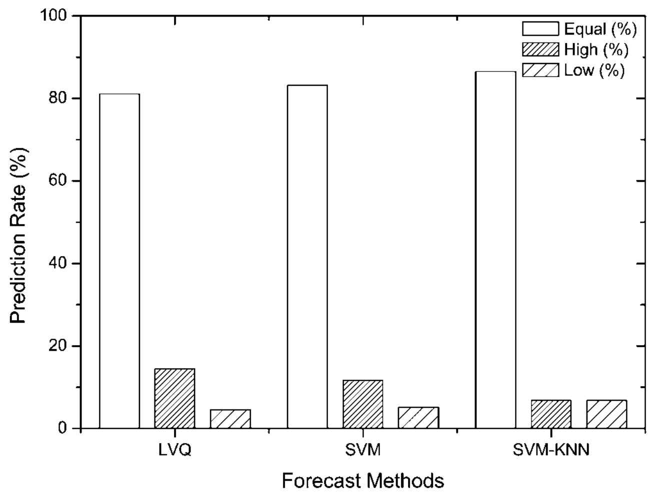

Fig. 2 Test result of 2004 obtained form LVQ, SVM and SVM-KNN method.

4.3 Results Analysis

Figure 2 demonstrates that SVM-KNN method offers certain advantages over other two methods. Since the test results for all four years are similar, Figure 2 only plots the test results of 2004. It can be seen that SVM-KNN method has the highest rate of 'Equal' and lowest rate of 'High'. On the other hand, the rate of 'Low' is slightly greater in SVM-KNN than in the other two. This fact can be explained as follows. The value range of non-flaring samples is larger than that of flaring samples, which means the non-flaring samples are more spread out in the feature space than the flaring samples. According to the SVM-KNN algorithm, samples near the separating hyperplane take part in the classification, the more spread-out distribution makes the non-flaring samples in the training set slightly more attractive to the samples in the testing set, which results in a slightly increase of 'Low' predictions.

5 CONCLUSIONS AND DISCUSSION

The SVM-KNN method is firstly applied to solar flare forecasting. Based on a proven relationship between SVM and KNN, this new method improves the SVM algorithm for classification by taking advantage of the KNN algorithm according to the distribution of test samples in a feature space and gives a higher prediction accuracy than the SVM or an NN-based method alone. At the same time, however, it also gives an increased rate of 'Low' predications, which is not always desirable. The present forecasting model is constructed on data from active regions, which means the number of non-flaring samples is larger than the number of flaring samples. The existence of a large number of non-flaring samples is also a contributing factor for the high prediction accuracy in our tests.

Our study involves only a two-class forecasting: whether the flare importance is or is not smaller than M. Experiments on multi-class flare forecast will be considered in our future research. Furthermore, some new predictors will be extracted from observational data of solar photospheric vector magnetic field (Cui et al. 2006).

Acknowledgements This work is supported by National Natural Science Foundation of China (NSFC) under Grants 10233050 and 10673017, by Chinese Academy of Sciences under grant KGCX3-SYW-403-10, and by National Ministry of Science and Technology under grant 2006CB806307. The authors are indebted to the anonymous referee for helpful suggestions, the GOES, and SOHO teams for providing the wonderful data.

Appendix A: PROOF OF THEOREM 1

Defining positive and negative support vectors as two representative points: $\phi(x)^+ = \frac{1}{C}\sum_{y_i=1,i=1}^{l}\alpha_i\phi(x_i)$ and $\phi(x)^- = \frac{1}{C}\sum_{y_i=-1,i=1}^{l}\alpha_i\phi(x_i)$, where $\sum_{y_i=1}\alpha_i = \sum_{y_i=-1}\alpha_i = C$ (from $\sum_{i=1}^{l}\alpha_i y_i = 0$).

For optimal solution w, we have

$$w = \sum_{i=1}^{l}\alpha_i\phi(x_i) = C[\phi(x)^+ - \phi(x)^-]. \tag{A.1}$$

For each positive sample, from Kuhn-Tucker condition:

$$\alpha_i\{y_i[(w,x_i) - b] - 1\} = 0, i = 1, \cdots, l, \tag{A.2}$$

we have $\alpha_i\{[w,\phi(x_i)] - b - 1\} = 0$, accordingly,

$$\begin{aligned} 0 &= \sum_{y_i=1}\alpha_i\{[w,\phi(x_i)] - b - 1\} \\ &= [w, \sum_{y_i=1}\alpha_i\phi(x_i)] - C\cdot b - C \\ &= C[\phi(x)^+ - \phi(x)^-, C\phi(x)^+] - C\cdot b - C \\ &= C\{C[\phi(x)^+ - \phi(x)^-, \phi(x)^+] - b - 1\}. \end{aligned} \tag{A.3}$$

Thus

$$b = C[\phi(x)^+ - \phi(x)^-, \phi(x)^+] - 1. \tag{A.4}$$

For each negative sample, similarly from Equation (A.1), the following equal can be acquired:

$$b = C[\phi(x)^+ - \phi(x)^-, \phi(x)^+] + 1. \tag{A.5}$$

Using [(A.3)+(A.4)]/2 yields:

$$b = \frac{C}{2}[\phi(x)^+ - \phi(x)^-, \phi(x)^+ + \phi(x)^-] = \frac{C}{2}[k(x^+,x^+) - k(x^-,x^-)]. \tag{A.6}$$

Putting the 1NN classified formula into the classified process of SVM we obtain the follow formula:

$$\begin{aligned} g(x) &= \|\phi(x) - \phi(x)^-\|^2 - \|\phi(x) - \phi(x)^+\|^2 \\ &= 2k(x,x^+) - 2k(x,x^-) + k(x^-,x^-) - k(x^+,x^+) \\ &= \frac{2}{C}\big\{\sum_i \alpha_i y_i k(x,x_i) + \frac{C}{2}[k(x^-,x^-) - k(x^+,x^+)]\big\} \\ &= \frac{2}{C}\big[\sum_i \alpha_i y_i k(x,x_i) - b\big]. \end{aligned} \tag{A.7}$$

References

Amari S., Wu S., 1999, Neural Networks, 12(6), 783
Bian Z. Q., Zhang X. G., 2000, Pattern Recognition, Beijing: TsingHua Univ. Press
Chang C. C., Lin C. J., 2001, LIBSVM: a library for support vector machines (Version 2.3.1), *http: //citeseer.ist.psu.edu / chang01libsvm.html*
Cristianini N., Campbell C., Shawe-Taylor J., 1999, Neural Networks (ESANN)
Cui Y. M., Li R., Zhang L. Y., He Y. L, Wang H. N., 2006, Sol. Phys., 237, 45
Gallagher P. T., Moon Y.-J., Wang H. M., 2002, Sol. Phys., 209, 171
Gavrishchaka Valeriy V., Ganguli S. B., 2001, J. Geophys. Res., 106, 29911
Li R., Ye S. W., Shi Z. Z., 2002, Chinese Journal of Electronics, 30(5), 745
Lundstedt H., 1997, Geophys. Monogr., 98, 243
Gong J. C., Xue B. S., Liu S. Q. et al., 2004, Chinese Astronomy and Astrophysics, 28, 174
McIntosh P. S., 1990, Sol. Phys., 125, 251
Wang J. L., 2000, Chinese Astronomy and Astrophysics, 24, 10
Wheatland M. S., 2004, AJ, 609, 1134
Zhang G. Q., Wang J. L., 1994, Progress in Geophysics, 9, 54
Zhu C. L., Wang J. L., 2003, Chin. J. Astron. Astrophys. (ChJAA), 3, 563
Vapnik V., 1995, The Nature of statistical Learning Theory, New York: Springer-Verlag
Wu M. R., 2000, The research on classifier design for pattern recognition problems of large scale, Ph.D, Beijing: Tsinghua University