Support Vector Machine on Analysis of Temperature and Precipitation Data

Huey-Ru Wu, Ting-Huai Chang, Chia-Hao Liu

Meteorological Information Center

Central Weather Bureau

Abstract

Extreme weather becomes an important topic in recent years, while how to predict the occurrence of extreme weather is still a tough task. We tried to introduce a quite powerful classification tool in data mining area to find a way to deal with this task. A support vector machine (SVM) tool uses a nonlinear mapping to transform the original data into a higher dimension, in order to separate the data into the targeted two parts. We used the SVM tool to analysis years of observation data to classify out the data with high temperature in summer, low temperature in winter and precipitation. Then, we analyze the results with some various performance measurements.

Key words: support vector machine, extreme weather

1. Introduction

Heavy rain, strong winds, heat waves and other abnormal weathers, because of difficulties in prediction, and hard in prevention, often result in relatively serious economic losses.[4] Through this research, we hoped to find some extreme weather forecasting method, which could be applied to the actual weather forecast later.

In this work, we picked out observation data of two seasons from two weather stations, and used the statistical method, stepwise regression, to choose related attributes from the data. The selected data are then used in the classification tool, support vector machine, for analysis. Finally, the results are analyzed and discussed.

2. Data

The data used in this work was the observation data from Keelung and Penghu weather stations of Central Weather Bureau. The reason of chosen data from these two stations was their weather styles were in two quite different types. We wanted to make some comparison between these two sets of experiments. As regards precipitation, Keelung has the highest rate of rainy data while Penghu has almost the least. As to temperature, Keelung is in the northern end of Taiwan Island, while Penghu is in the south, thus shows the different style.

The data of summer and winter from year 1995 (June, 1995) to 2008 (February, 2009) was used, where we chose data from June to August as summer and form December to February of next year as winter. The pre-gathered or processed weather attributes of each location and period were station pressure (pr), temperature (te), dew point (td), relative humility (rh), latitudinal and longitudinal direction wind (wu and wv) and precipitation (pp). The attributes used for SVM training were then chosen based on the result of the “stepwise regression” analysis.

"Stepwise regression" is a statistical method used for choosing the relatively effective variables from the dataset contains many variables. The repeatedly interleaving forward and backward searching procedures are done to develop the "best" subset of variables step by step. The forward procedure starts on model without any variable, checking each variable in turn. One variable per time is included into the subset when it is statistically most and enough significant. Then, the backward procedure starts with the full selected subset, testing and excluding each variable when it is no more significant in the renewed attribute combination. The interleaving procedures repeat until no more change occurred on the subset.[3] The chosen attributes in our work were shown in Table 1.

Table 1. Attributes used for SVM training.

station	season	class	pr	te	td	rh	wu	wv	pp
KL	summer	pp	V	V	V	V	V	V	
KL	summer	te	V		V	V	V		V
KL	winter	pp		V	V	V	V	V	
KL	winter	te	V		V	V		V	V
PH	summer	pp	V	V	V	V	V	V	
PH	summer	te			V	V	V	V	V
PH	winter	pp	V	V	V	V	V	V	
PH	winter	te	V		V	V	V	V	V

The classification target in precipitation was focus on the data noted as raining, and the one in temperature was to classify data with high temperature (>32°C) in the summers or with low temperature (<14°C) in the winters. The target data amount analysis was as shown in Figure 1. The amount of data used in the trainings was 2208 in each summer and 2160 or 2184 (leap year) in each winter. Differences between the seasons and the weather stations of the selected two target classification parameters can be easily seen. The data rate of the classification target over all data under tests in each dataset was analyzed as in Figure 2, which values ranging from near 0% to 40%. We would exam the influences of the varying target data rate in the results analysis in later works.

Figure 1. Target data amount analysis

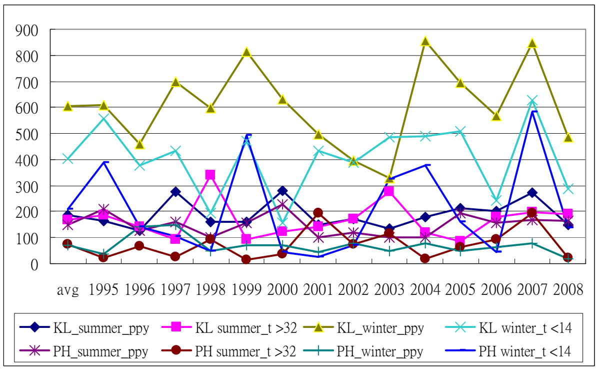
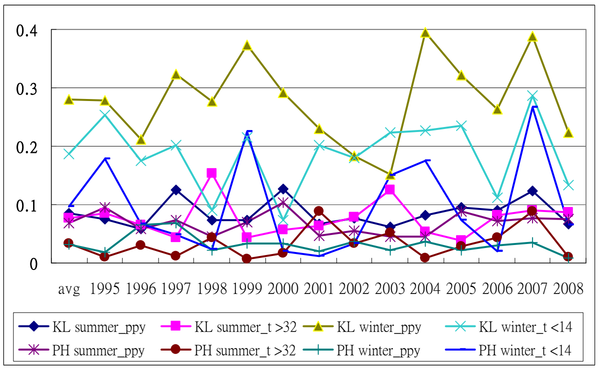


Figure 2. Target data rate analysis



3. Method

Classification is a method for analyzing data by separating it into two or more groups, in order to understand more about the characteristics of the data. A typical data classification is mainly a two-step process, training and predicting. In which a model is constructed through training process, then it is used to predict the categorical labels of data. Support vector machine is a relatively new and powerful method for data classification, both on linear and nonlinear ones.

SVM method uses a nonlinear mapping to transform the original training data into a higher dimension. Then, it searches for a "separating hyperplane", which is a boundary separating the data of one class from the other, in the new dimension. Data from two different classes can always be separated by a hyperplane with proper nonlinear transform into a high enough dimension. After finding the proper hyperplane with the maximal margins, the support vectors, which are transformed data lay on the margins, are noted in the model for later class prediction.[2] In this work, we tried to use SVM method on temperature and precipitation data analysis.

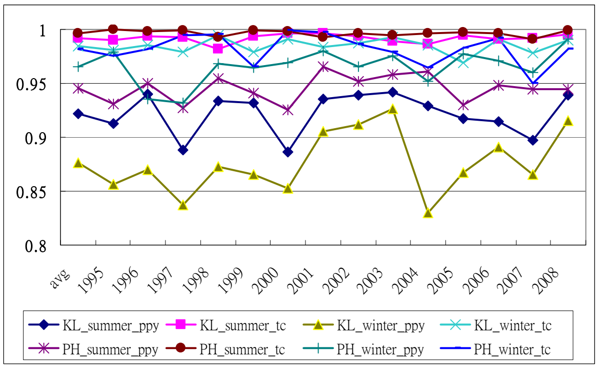
In practice, SVM tools are often designed based on the kernel functions rather than simply transform the data into higher dimensions. Some general kernel functions are proposed for SVM training, which work in rather low dimensions, but behave like the inner product in high dimensions. The LIBSVM (a LIBrary for Support Vector Machines) is a tool built based on some typical kernel functions.[1] The construction of SVM models and the following tests in our work were done using the functions of the LIBSVM. The parameters used in our SVM model training were the default values set in the LIBSVM.

The SVM model trainings were done in four data group, including of one 14-year training: 1995-2008 (a14), and three 7-year trainings: 1995-2001 (f7), 1999-2005 (m7), and 2002-2008 (r7). The SVM model was trained with every year data in the defined group for both precipitation and temperature classification, while one year of data per time was kept out for later test of prediction correctness. For example, in the KL_summer_ppy_m7_2001 training case, the model was trained with summer data in year 1999-2000 and year 2002-2005 of Keelung station for precipitation classification, and the data of year 2001 was used for testing the classifying ability of the generated SVM model. Due to the limit of document length, only the average values of trainings results in each year are shown in the following figures.

4. Results and Measurements

We first examined the accuracy of prediction in the experiments. As shown in Figure 3, the overall accuracy was more than 90% in average; and was more than 80% even in the worst case. This showed the possibility and power of the SVM tools.

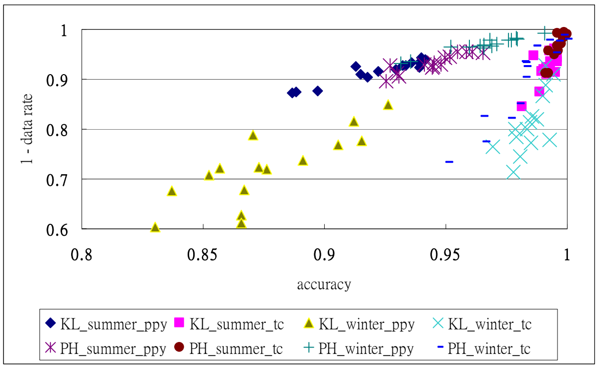
Figure 3. Accuracy of SVM training



However, when further analyzed the details of the results, we began to suspect the good performances of the trainings. The suspect increased as we look back on the target data rate analysis. With an initially 95% non-precipitation data as training input, the accuracy would be 95% even if the model predicts all of the test records as non-precipitation, though which is not the kind of model we want.

We further examined the prediction results and found that in all of the records noted as precipitation, the percentage of correctly prediction was not as high as the accuracy rate shown. The reason of getting the high accuracy while not perform that good in reality should be the influence of the data rate of the two classes in the tests, especially the high ratio of non-target class. Figure 4 shows the high relatedness between the accuracy and non-target data rate, the two values are proportional to each other.

Figure 4. Accuracy vs. 1－data rate



To avoid the possibly overoptimistic estimates of the results, some measurements based on confusion matrix were introduced, which are quite useful measurements in information retrieval area. The confusion matrix between classes of the observational data and the prediction results is defined as Table 2.

Table 2. Confusion matrix between the observed classes and the predicted classes

		Predicted class	
		Target	Non-target
Observed class	Target	True positives(TP)	False negatives(FN)
	Non-target	False positives(FP)	True negatives(TN)

The matrix separates both the observed class and the predicted class of data into target and non-target parts, and generates four sections in the results. In our earlier measurement, we used

accuracy = (TP+TN) / (TP+FP+TN+FN)

in which contains the performance of both “true positives” and “true negatives” results. While we are in fact more interested in the targeted part of data, some other measurements were chosen:

precision = TP / (TP+FP)

recall = TP / (TP+FN)

false alarm rate = FP / (FP+TN)

In which the “precision” measures the correct ratio over all records been predicted as “positive”. While the “recall” measures the ratio of successfully found targeting records from the data pool. And the “false alarm rate” checks the cost of wrong predictions.[2],[5] The comparison results of the new applied measurements were shown in Figure 5, Figure 6, and Figure 7. Based on the definition, the preferred value of precision and recall should be as high as possible, while the false alarm rate should be left as low as possible.

Figure 5. Precision of SVM training

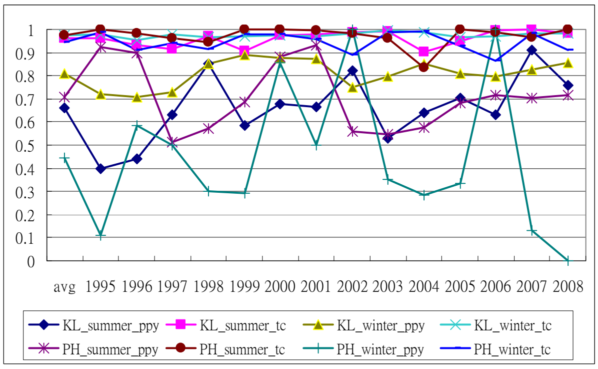


Figure 6. Recall of SVM training

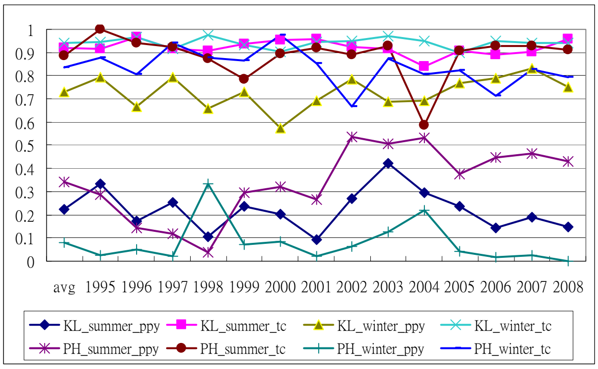
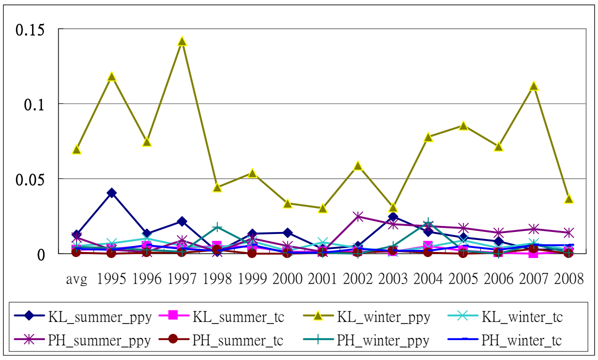


Figure 7. False alarm rate of SVM training



Comparing to the accuracy, we could see the obvious variance under these measurements. In the precision measure, some of the test performances decreased, and it became even lower under the measurement of recall. In the meanwhile, the false alarm rates of all the testing remained rather low. In practice, a combination of both precision and recall are often used for usually these two measurements have to be considered at the same time. We chose the F-score as the mixed measurement,

F-score = (2*precision*recall) / (precision+recall)

and confirmed it not been influenced by the varying data rate, as shown in Figure 8 and Figure 9. Thus the F-score seemed to be a more fair measurement than accuracy. As we checked the results, we could clearly see the different performances between experiments on temperature and precipitation. The same condition can also be seen in precision and recall measurements.

Figure 8. F-score of SVM training

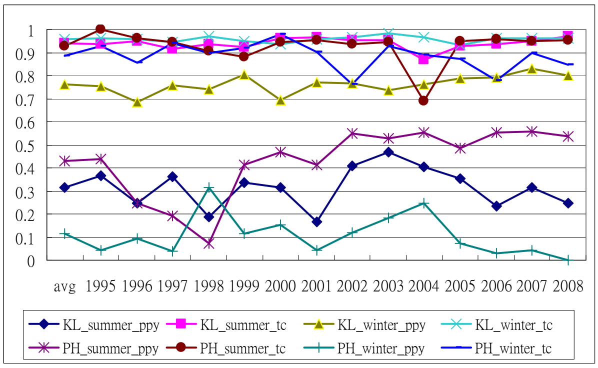
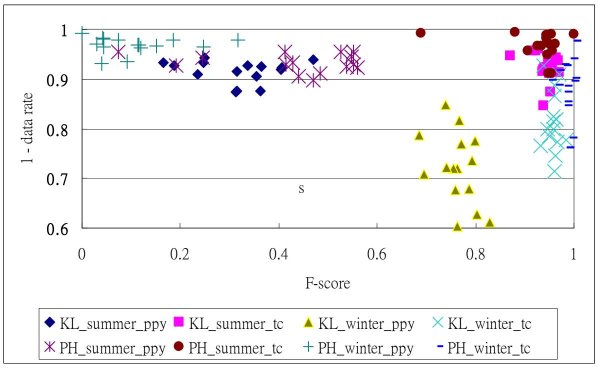
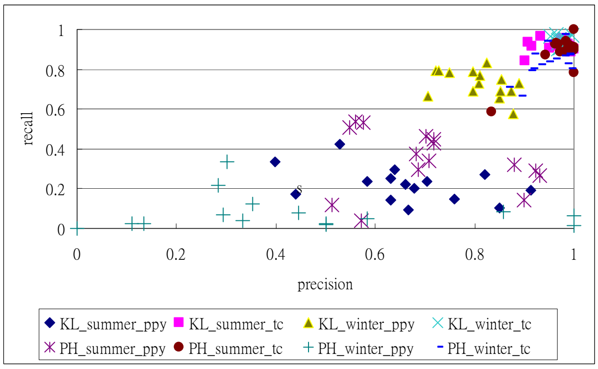


Figure 9. F-score vs. 1－data rate



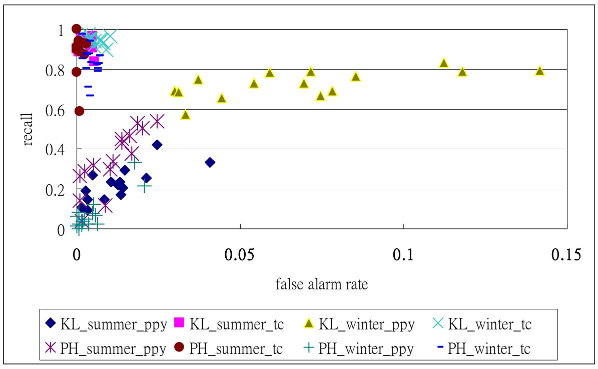
Then we examined the measurements in more detail. When checking the precision vs. recall condition, as shown in Figure 10, the experiments focused on temperature did quite well and lay at the up-right corner of the plot, which showed that both the precision and recall were as high as wanted. The ones focused on precipitation data were not as good, only the winter test of Keelung showed rather good results.

Figure 10. Precision vs. recall



As to the false alarm rate vs. recall analysis in Figure 11, in which the up-left corner with low false alarm rate and high recall was wanted, almost the same results were shown. Again the classification performances of temperature were much better. Further analyzed this figure we found out that the bad performances of precipitation tests were generated by low recalls, while the false alarm rates were in the acceptable low area.

Figure 11. False alarm rate vs. recall

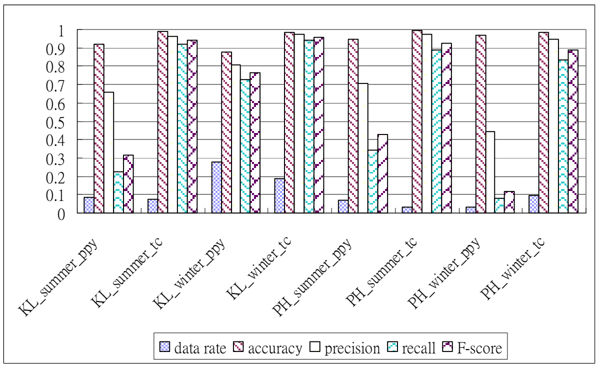


The performance of works on temperature did much better then the ones on precipitation under every measurement. Even if the target data rates of some precipitation tests were higher than the ones in temperature tests. Although the performances of precipitation tests were influenced clearly by the original data rate, that might not be the only reason. The data rate might become a major influence only when the support vector machine could not extract clear separate hyperplane with the limited records offered, not an overall truth.

5. Conclusions

The overall performances were summarized in Figure 12. It could easily be seen there are quite various results under different test groups. Some parts of the work got quite well performance, like the temperature ones, showed that the SVM tool might be usable in weather prediction, while the results of the other parts were not as satisfying, as many of the precipitation ones. These variance hints what we can work on next. Find methods which can gain the target data rate for training data to help the SVM in finding clearer separating hyperplane. Try different selection of attributes used as input training data which can be more related to the classification target. Or adjust the parameters of the LIBSVM to find some possibly better setting for model training.

Figure 12. SVM training average results



Reference

[1] Chih-Chung Chang, Chih-Jen Lin, 2001: “LIBSVM: a library for support vector machines”, Software is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[2] Jiawei Han, Micheline Kamber, 2007: Data Mining: Concepts and Techniques, Second Edition, Chapter 6, Morgan Kaufmann Publishers, San Francisco, CA, 337-344, 359-362

[3] John Neter, Michael H. Kutner, Christopher J. Nachtsheim and William Wasserman, 1999: Applied linear regression models, Third edition, Chapter 8.4, McGraw-Hill Co. Inc., Singapore, 348-353

[4] Lott, N., Ross, T., Houston, T., and A. Smith, 2008: Billion dollar U.S. weather disasters, 1980-2008. Factsheet, NOAA National Climatic Data Center, Asheville, NC, 2

[5] Tom Fawcett, 2006: “An introduction to ROC analysis”, Pattern Recognition Letters 27, 861-874