ABSTRACT

The article presents a method for automatic recognition of cloud type by supervised learning. The method used is the Support Vector Machines (SVM) method, which was originally developed by Vapnik (Vapnik, 1998) and is nowadays increasingly used in pattern recognition and classification (Cristianini, 2000).

In addition to the existing well known cloud classifications from SEVIRI data, which are mainly based on threshold methods (Lutz et al., 2003; Derrien et al., 2003) an alternative method of supervised learning was tested and the results are presented in this article.

The Support Vector Machines method needs a dataset with SEVIRI measurements and reference labels for learning purposes. The dataset with SEVIRI measurements and classification made by nephanalyst was provided by Météo-France for the period from March to December 2003. Cloud classification by SVM was also verified on a selected synoptic case.

It has been established that the classification of the Support Vector Machines method using kernel function of Radial Basis Function gives better cloud classification than SVM using linear kernel function.

1. THE PRINCIPLES OF SUPPORT VECTOR MACHINES METHOD

Linear classification
The SVM algorithm for the purpose of classification creates a hyper-plane that separates the data into two classes with the maximum-margin between these two datasets. Given training examples labelled either "class1" or "class2", a maximum-margin hyper-plane splits the "class1" and "class2" training examples, so that the distance from the closest examples (the margin) to the hyper-plane is maximized.
The use of the maximum-margin hyper-plane is motivated by Vapnik (Vapnik, 1998) theory, which provides a probabilistic test error bound that is minimized when the margin is maximized. The parameters of the maximum-margin hyper-plane are derived by solving a quadratic programming (QP) optimization problem. There exist several specialized algorithms for quick solving of the QP problem that arises from SVM. The most common method which was used here for solving the QP problem is Platt's SMO algorithm (Cristianini, 2000).

Non-linear classification with the kernel
The original optimal hyper-plane algorithm proposed by Vladimir Vapnik (Vapnik, 1998) was a linear classifier. However, further upgrades of the original idea suggested a way to create non-linear classifiers by applying the kernel to maximum-margin hyper-planes. The resulting algorithm is formally similar, except that every dot product is replaced by a non-linear kernel function. This allows the algorithm to fit the maximum-margin hyper-plane in the transformed feature space. The transformation may be non-linear and the transformed space high dimensional; thus though the classifier is a hyper-plane in the high-dimensional feature space it may be non-linear in the original input space.

If the kernel used is a radial basis function, which was in our case, the corresponding feature space is a Hilbert space of infinite dimension, (Cristianini, 2000). Maximum margin classifiers are well regularized, so the infinite dimension does not spoil the results.

2. DATABASE FOR LEARNING

The Support Vector Machines method needs a dataset with the SEVIRI measurements and reference labels for learning purposes. The dataset with the SEVIRI measurements and classification made by nephanalyst was provided by Météo-France. From the period from March to December 2003 about 6,000 targets which consist of 5 x 5 SEVIRI pixels with 1 - 11 channels were used for training purposes.

Different selections of input features (SEVIRI channels and combinations of channels), different input parameters (to find optimal input parameters on training dataset) and different kernel functions (linear or RBF) of the Support Vector Machines method were evaluated. In addition, learning was performed separately on day time (sun zenith angle < 80°) and night time (sun zenith angle > 90°), since during night-time “solar” channels do not offer any additional information.

Diagram 1 illustrates schematically training and testing procedures. The database was first split in three random parts for each class, where two parts were used for training of the algorithm parameters and one part was used at the end for testing of the algorithm. The algorithm input parameters were adjusted in so that the margins between all classes were maximised and at the same time the number of wrongly classified targets was minimised on the training datasets. We have tested this
procedures for different input measurements (SEVIRI channels and combinations of SEVIRI channels) using both linear and nonlinear (Radial Basis Functions) kernels.

1. Diagram 1. Schematic procedure of SVM training and testing.

After establishing such classifier the results were tested on the testing part of database. The classification by SVM method was tested on some 2000 targets from reference database, where each target consists of 5 x 5 SEVIRI pixels. These testing targets were not included in the training algorithms and give independent evaluation of the method.

3. RESULTS ON TESTING PART OF DATABASE

Below are the results of the testing part of dataset for 8 classes: clear sea, clear land, snow, low clouds, middle clouds, high clouds, transparent clouds, transparent clouds over middle or low clouds. The verification measures were selected to be Probability of Detection (POD) and False Alarm Rate (FAR) for each class. Classification using linear and RBF kernel functions were tested but the results are better using RBF kernel (Iršič Žibert, 2004). The results are shown for SVM using RBF kernel function during day-time (Table 1) and night-time (Table 2). These results were better performed than those with linear kernel implemented. Tables 1
and 2 demonstrate that cloud free areas (sea, land and snow) are relatively well
classified, and so are low and high clouds. Middle and high transparent clouds (also
over lower clouds) have poorer scores, but still acceptable during daytime.
Exceptions are transparent clouds over lower clouds during night-time, which needs
further investigation.

Table 1. POD and FAR for testing part of database in daytime.

<table>
<thead>
<tr>
<th>class</th>
<th>sea</th>
<th>land</th>
<th>snow</th>
<th>low clouds</th>
<th>middle clouds</th>
<th>high clouds</th>
<th>transparent clouds</th>
<th>transparent over lower clouds</th>
</tr>
</thead>
<tbody>
<tr>
<td>POD</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
<td>0.95</td>
<td>0.73</td>
<td>0.93</td>
<td>0.84</td>
<td>0.53</td>
</tr>
<tr>
<td>FAR</td>
<td>0.06</td>
<td>0.03</td>
<td>0.02</td>
<td>0.05</td>
<td>0.23</td>
<td>0.06</td>
<td>0.18</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 2. POD and FAR for testing part of database during night.

<table>
<thead>
<tr>
<th>class</th>
<th>sea</th>
<th>land</th>
<th>snow</th>
<th>low clouds</th>
<th>middle clouds</th>
<th>high clouds</th>
<th>transparent clouds</th>
<th>transparent over lower clouds</th>
</tr>
</thead>
<tbody>
<tr>
<td>POD</td>
<td>0.97</td>
<td>0.97</td>
<td>0.93</td>
<td>0.59</td>
<td>0.96</td>
<td>0.82</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>FAR</td>
<td>0.08</td>
<td>0.08</td>
<td>0.06</td>
<td>0.13</td>
<td>0.09</td>
<td>0.23</td>
<td>0.36</td>
<td></td>
</tr>
</tbody>
</table>

The results are promising for the method used. It is important to stress that these
results cannot give final decision on cloud classification evaluation, since the
database consisted of measurements in the time when Meteosat-8 was not fully
operational, and the second important issue is that cloud classes are not equally
distributed in the selected database, consequently the interpretation of results has to
be taken with care. The one important specification of the SVM method is
independence of the NWP fields, since only satellite measurements were used as
input features.

4. SELECTED CASE

In addition to the scores on the test part of database we have also verified a synoptic
case, and thus also spatial consistency can be verified. We have selected 28 July
2003 from 10 – 18 UTC, but it is illustrated in this article only at 11:45 UTC (Fig. 1 -
3). For reference there are two multi-channels combinations presented in Figure 1
(VIS0.6+VIS0.8+IR10.8) and in Figure 2 (NIR1.6+VIS0.8+VIS0.6) where can be
seen: clear areas, low cloudiness, vertically developed cloudiness and high
transparent cloudiness.
Figure 1. Channel combination (VIS0.6+VIS0.8+IR10.8) at 11:45 UTC on 28 July 2003.

Figure 2. Multi-channel combination (NIR1.6+VIS0.8+VIS0.6) at 11:45 UTC on 28 July 2003.
Figure 3. SVM cloud classification results with RBF kernel at 11:45 UTC on 28 July 2003. Classes are in colour scale: dark blue - clear sea, blue - clear land, light blue - snow, green - low cloudiness, green/yellow - middle clouds, orange - high clouds, red - transparent clouds, dark red - transparent clouds over middle or low cloudiness.

In general, Figure 3 and other timeslots images (not shown) lead to the conclusion that overall performance of SVM classification is relatively good during daytime, but of course there is still room for improvements, specially during dawn and night-time.

5. CONCLUSIONS AND FUTURE PLANS

This paper presents the concept of supervised learning using Support Vector Machines method. For supervised learning the quality and completeness of reference database is of utmost importance. And since we used the first database of Meteosat-8 measurements from the early period (March – December 2003), it cannot be claimed that this SVM classification is in general very good, but for the selected database and a synoptic case we got very promising results when implementing RBF kernel.

We plan to verify results of SVM classification on newer and more complete database concerning equal distribution over different classes. In addition we plan to compare the results of SVM method with those of MPEF CLAI and/or NWCSAF CT products.
6. REFERENCES


