ECAI 2006
G. Brewka et al. (Eds.)
IOS Press, 2006

Efficient Learning from Massive Spatial-Temporal Data through Selective Support Vector Propagation

Yilian Qin and **Zoran Obradovic**[1]

Abstract. In the proposed approach, learning from large spatial-temporal data streams is addressed using the sequential training of support vector machines (SVM) on a series of smaller spatial data subsets collected over shorter periods. A set of representatives are selected from support vectors corresponding to an SVM trained with data of a limited spatial-temporal coverage. These representatives are merged with newly arrived data also corresponding to a limited space-time segment. A new SVM is learned using both sources. Relying on selected representatives instead of propagating all support vectors to the next iteration allows efficient learning of semi-global SVMs in a non-stationary series consisting of correlated spatial datasets. The proposed method is evaluated on a challenging geoinformatics problem of aerosol retrieval from Terra satellite based Multi-angle Imaging Spectro Radiometer instrument. Regional features were discovered that allowed spatial partitioning of continental US to several semi-global regions. Developed semi-global SVM models were reused for efficient estimation of aerosol optical depth from radiances with a high level of accuracy on data cycles spanning several months. The obtained results provide evidence that SVMs trained as proposed have an extended spatial and temporal range of applicability as compared to SVM models trained on samples collected over shorter periods. In addition, the computational cost of training a semi-global SVM with selective support vector propagation (SSVP) was much lower than when training a global model using spatial observations from the entire period.

1. INTRODUCTION

With recent advances in remote sensing technologies, scientists collect a much larger volume of spatial-temporal observations than previously. For example, since 2000, the Multi-angle Imaging SpectroRadiometer (MISR) instruments onboard Terra satellite of NASA's Earth observing system have been used to observe solar radiation through nine cameras each in four spectral bands with 1.1 km spatial resolution and global coverage of Earth in about 9 days [1]. Such a resolution results in quick accumulation of terabytes of high dimensional spatial-temporal observation data. The objective of learning nonlinear relationships from such non-stationary data can in principle be addressed by artificial neural networks (ANN) and support vector machines (SVM) [2]. While theoretically such prediction models are more accurate with large training datasets, the computational cost of learning from high volume data prohibits the training and retraining of global models using the entire historical data.

In a recent study [3], a neural network ensemble procedure was developed to learn ANN models from massive datasets. In this project, neural networks of low complexity were learned from small datasets first, and the component networks were incrementally combined into an ensemble while progressively increasing the size of the training dataset. This allowed for a regulation of the computational costs of developing nonlinear regression models from large datasets through an automated identification of an appropriate number of component networks, their complexity and a sample size for training the component.

For learning nonlinear relationships in high dimensional data, SVM are frequently preferred when both stability and high accuracy are required [4]-[6]. However, the training and prediction time of SVM increases with the increased number of support vectors. Therefore, both training and prediction time may be prohibitively long when learning with an SVM for massive spatial-temporal datasets that typically require a large number of support vectors. Approaches considered to address this situation include reducing the size of datasets [7], improving the performance of SVM through incorporating known invariance of the problems [8], approximating the decision surface and reformulation of the training problem that yields the same decision surface using a smaller number of basis functions [9][10], using a proximal SVM [11], and exploiting partial/merge k-mean method with parallel computation [12]. Nevertheless, as long as the time-consuming kernel matrix operations are required, it is prohibitively costly to develop a global SVM for massive spatial-temporal datasets. An efficient method uses SVMs trained locally with a small number of instances [13]. However, such an SVM is only valid locally while in many cases global and semi-global properties that are difficult to capture with local models are of primary interest.

Inspired by previous projects, we propose efficient development of semi-global SVM models valid within extended spatial regions and usable over a longer time span than those of local models. This approach is based on the observation that regional features frequently exist in spatial-temporal datasets over certain areas due to a large number of factors such as vegetation, topology, air pollution, humidity, snow, ice, and cloud coverage. Such factors are also strongly time dependent, and consequently induce semi-stationary temporal features within certain time intervals. Therefore, sub-models developed from regional datasets might remain

[1] Temple University, USA, email: zoran@ist.temple.edu

accurate for a reasonably long time or over a large area, and so can be frequently reused for reduced computational cost. The proposed method for training and reusing the semi-global SVM models will be discussed first followed by a summary of the experimental evaluations and discussion.

2. METHODOLOGY

In the simplest case where the patterns are linearly separable, an SVM looks for a separation plane which maximizes the margin. For linearly non-separable patterns, the attributes are nonlinearly mapped to a higher dimensional space such that the patterns in the new feature space are linearly separable [6]. An SVM is represented by a kernel matrix, or equivalently a set of support vectors and parameters specifying the kernel functions. A kernel matrix in a quadratic size of the number of support vectors is involved in the SVM training processes.

The support vectors of an SVM are the most important instances in the datasets from which the SVM is trained. Therefore, merging the support vectors of an SVM trained using dataset one with raw examples of another dataset results in a merged dataset which to some extent carries the properties of both datasets. The number of instances in the merged dataset would be much smaller than the total number of instances in the combined raw datasets. This observation, the starting point in this project, is aimed at the construction of semi-global training datasets of limited size for efficient learning of a series of SVMs on large spatial-temporal data series.

In spatial-temporal situations, data properties often remain semi-stationary over certain space and time intervals. Therefore, SVMs trained from neighboring local spatial datasets in a time series might frequently be similar. In such cases, merging the support vectors of a previous SVM to the adjacent raw data might be sufficient to learn a new SVM that preserves statistical properties of both regions. Based on this idea, a coarse-grained pseudo code of the proposed algorithm for learning SVMs with selective support vector propagation (SSVP) is presented below:

```
1. S is initialized to an empty support vector set
2. For each spatial dataset $D_i$, i=1,2,…, in a series
3.     D = Union of S and $D_i$
4.     Train a support vector machine $M_i$ with D
5.     S’ = support vector set of $M_i$
6.     S” = set of instances in S’ misclassified by $M_i$
7.     S = set of instances in S’ but not in S”
8. End For
```

While the number of support vectors in each individual dataset D_i can be small, propagating all support vectors for merging with new data is likely to enlarge the merged data in time because of the inclusion of outliers that are considered support vectors. With a support vector selection method described in lines 5, 6, and 7 in the above pseudo code, the number of support vectors propagated to the next stage will be reduced significantly. More specifically, in each iteration of the training process, all the misclassified support vectors of the resulting SVM are excluded from the support vector set to be propagated to the next training step. The effectiveness of this method will be demonstrated and discussed in the experimental results section below.

3. DATA DESCRIPTION

3.1 Synthetic Data

A series of simple synthetic datasets were generated with four cluster centers, namely (0, 1), (1, 0), (0, -1), and (-1, 0) on the 2-D plane. For each dataset, the first two clusters are labeled as positive and the other two are negative. Using different noise levels, we generated a low noise and a high noise dataset. For the low noise dataset, a randomly generated noise of normal distribution with a standard deviation of 0.1 was added to the dataset. Each dataset in the low noise series contained 100 instances with 25 instances drawn from each of four clusters. While the properties of the series, such as the location of cluster centers and the standard deviations, did not vary with time, the dataset may still be considered a spatial-temporal one in that the noise is injected in time. For the high noise dataset, the same configuration of the datasets was adopted with a standard deviation of 0.3 in noise level. A series of 500 datasets were generated for low noise and high noise cases respectively.

3.2 MISR Data

The spatial-temporal data are obtained from the MISR instrument aboard the Terra satellite of NASA’s Earth observation system. MISR views Earth with nine cameras pointed at different viewing angles, each with four spectral bands. Terra orbits the Earth about 15 times per day and completes a data cycle in 16 days. For each 16-day cycle, there are 233 distinct MISR paths (orbits) [http://eosweb.larc.nasa.gov/MISRBR]. We analyzed four such cycles covering the entire continental U.S. from 07/17/2002 to 01/08/2003 where each cycle consisted of data over 47 paths.

The MISR Level 1B2 radiance and Level 2 aerosol optical depth (AOD) data were used for SVM model learning. We extracted 36 radiances and supplemented these with additional 82 geometric parameters resulting in 118 dimensional patterns. The learning target was the green band regional mean AOD, which was converted to binary classes of low (thin) vs. high (thick) aerosol situations as compared to the mean AOD (0.18) for the entire datasets. So, the problem of aerosol estimation from radiances was reduced to a classification problem of identifying low vs. high aerosol situations given two classes of about the same size.

The observations under cloudy conditions were removed since the existence of cloud would significantly affect the correctness of the AOD retrieval. Cases with low-quality radiance (radiometric data quality index less than 2) and with missing AOD were also removed. Therefore, the number of remaining instances used in our experiments is space and time dependent. The total number of instances in the constructed dataset of each cycle is shown in Table 3.1.

Table 3.1. Cloud-free MISR radiance and aerosol data over entire continental US (period and number of instances).

Cycle	Start Time	End Time	# of Instances
1	07/17/2002	08/01/2002	76449
2	08/01/2002	08/17/2002	65620
3	08/18/2002	09/02/2002	51869
4	12/24/2002	01/08/2003	30451

4. RESULTS AND DISCUSSION

The method of combining and reusing the SVM proposed in Section 2 was applied to both synthetic and MISR dataset series described in Section 3. A linear classifier was used for experiments on synthetic datasets and RBF classifier with unity cost of constraint violation and width parameter for experiments with MISR datasets.

4.1 Results from Synthetic Data

The first objective was to explore the rate at which the number of support vectors might accumulate in time if all vectors were propagated to the next stages. For this experiment, the proposed method for merging relevant support vector representatives with new arriving data was applied without the proposed vector selection step by setting S" to be an empty set in line 6 of the algorithm in Section 2. The algorithm was applied to 50,000 data points partitioned into a series of 500 spatial datasets each consisting of 100 instances. The experiment was performed for low as well as for high noise synthetic data described in Section 3.1.

For a low noise situation, very few support vectors were propagated for merging with the newly arriving data independent of time. At the final 500-th time step, 10 support vectors were propagated for merging with 100 newly arrived data points and the resulting SVM was almost 100% accurate. In contrast, for high noise data, the number of propagated support vectors increased in time and soon become excessively large. In particular, unless pruned properly, about 2,200 support vectors were collected throughout the SSVP process of the initial 499 stages and they dominated over the newly arriving 100 data points. The number of the support vectors propagated at each of the 500 stages for merging with the next dataset is plotted at the top panel of Fig. 4.1. The linear trend of support vectors accumulation in this experiment was due to outliers identified as support vectors.

The selection step proposed for pruning support vectors (lines 5-7 of the algorithm described at Section 2) resolved this problem as shown at the bottom panel of Fig. 4.1. The simple selection method eliminated propagation of outliers as they were produced such that the number of propagated support vectors was much smaller (less than 140 throughout the entire process) as necessary for efficient applications in large spatial-temporal data series. While the accuracies of the final SVM obtained with and without support vector pruning on the entire dataset were similar (98.3% and 98.2%, respectively) the training and prediction with the pruning step was much faster than that without pruning.

At the bottom panel of Fig. 4.1, the slope of the curve corresponding to the number of propagated support vectors as a function of the time step is typically larger when the number of support vectors is smaller. This is expected since the position of the separation plane of an SVM is affected by the fraction of support vectors propagated from previous SVMs. The "inertia" of the separation plane, or the influence of previous datasets, increases with the fraction of propagated support vectors. On the other hand, the probability that a support vector from an earlier dataset is removed increases monotonically with time, such that the effect of a previous dataset becomes less important in the training process with the addition of new datasets. Therefore, the time interval, in which an SVM can be validly trained with the SSVP method, is determined by the competition of these two effects.

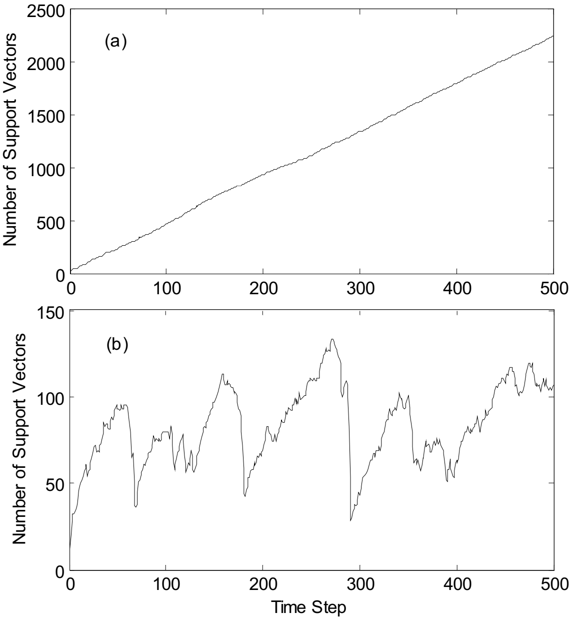

Figure 4.1. Effect of SSVP on the size of the combined datasets for high noise synthetic data. Without pruning the number of propagated support vectors in each of 500 time steps monotonically increased (top panel) while it stays small when the proposed selection step is used (bottom panel).

4.2 Results from MISR Data

The proposed method with SSVP was then applied to the MISR AOD retrieval described in Section 3. In this challenging test data properties vary both spatially and temporally with highly complex regional and temporal features. To isolate the effect of regional features of the dataset on the training of SVMs, the first set of experiments was carried out using cycle 1 aerosol data over continental U.S. (Table 3.1).

First, 47 SVMs were trained using 47 individual paths without SSVP. Each SVM was applied to predict out-of-sample AOD at all 47 paths of cycle 1. It was evident that the SVMs trained on the Eastern region (paths 4 to 27) were not applicable to Western region (paths 28 to 50) and vice versa, whereas SVMs trained on a single path on one coast were somewhat applicable for out-of-path prediction over the same region. In particular, the average accuracy for 24 SVMs trained on a single eastern path when tested on out-of-sample data over the entire eastern region was 66%. Similarly, a single western-path-based SVM achieved 67% out-of-sample average accuracy on 23 region paths.

Next, the proposed method was applied to cycle 1 data by learning SVMs on training data corresponding to individual paths supplemented with SSVP from previously learned SVMs. In this experiment the training started from the western most path (number 50 which is California data) and proceeded east (to finish at path number 4 which is East Coast U.S. data). Again, 24 East Coast predictors were tested for out-of-sample accuracy on the entire eastern region while each of 23 West Coast predictors was tested on the entire western region. The average out-of-sample accuracy of the SVMs with SSVP on the eastern and western regions was 77% and 72% respectively, which is significantly higher than the accuracy obtained without SSVP.

The result of the first two sets of experiments provide evidence that SSVP is also beneficial for improving the prediction accuracy when data is high dimensional and the spatial relationship is complex. The next set of experiments on the MISR dataset was designed to explore the applicability of the proposed methods for modeling temporal variation in the properties of spatial-temporal datasets. For prediction on all cycles, we first applied the final predictor of the previous experiment obtained by learning local SVMs starting from the western most path (50) and moving east with SSVP. The accuracy of this predictor on all paths at all cycles is summarized at the top panel of Fig. 4.2.

As already observed, cycle 1 accuracy of this predictor was much better on the eastern region corresponding to path numbers lower than 28 (accuracy 73% vs. only 38% at west). The same predictor also had good accuracy on the eastern region when predicting at cycle 2 and 3 data (71% and 64%).. The prediction accuracy on cycle 4 was consistently much worse (40%) since this data corresponds to a completely different season as compared to training period (cycle 1). The identified temporal semi-stationary property of the MISR datasets provides opportunities for developing semi-global SVM models for periods covering multiple time cycles without discarding older observations of the same region.

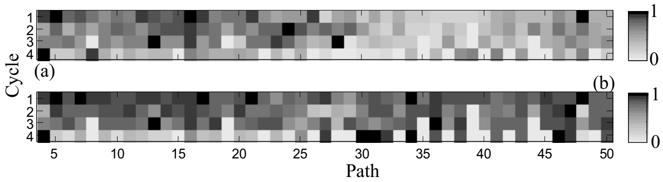

Figure 4.2. Out-of-sample classification accuracy (darker is more accurate) on Cycles 1-4 when trained on Cycle 1 data. Application of a single SVM obtained with SSVP (top panel) is compared to using 47 path-specific SVMs also constructed with SSVP (bottom panel).

Dependence of temporal semi-stationary behavior on the regional features is further examined in the experiments summarized at the bottom panel of Fig. 4.2. Here, an SVM trained on path k data of cycle 1 with SSVP from path k+1 at cycle 1 is used for out-of-sample classification at path k in all four cycles. So, instead of applying a single predictor for out-of-sample prediction, 47 path-specific predictors were trained with SSVP. The resulting accuracy was much improved as can be seen by comparing top and bottom panels of Fig. 4.2. In particular, for East Cost data (paths 4-27) average accuracy in cycle 1 and 3 is improved from 73%, 71% and 64% to 85%, 73% and 75%. In addition, West Coast (paths 28-50) accuracy was 80%, 71%, 71% and 82% for cycle 1-4, respectively. Therefore, model reuse in the MISR datasets was possible both spatially and temporally.

The performed experiments suggest that the SVMs trained with SSVP should only be extended to a spatial and temporal vicinity of the training datasets and not to extremely remote spatial regions or time periods. As a result, we developed a scheme for constructing a sequence of semi-global SVMs applicable for accurate prediction to moderately large spatial and temporal regions in an efficient way (such that the number of SVMs to cover the entire dataset is minimized).

By analyzing in more details cycle 1 classification accuracy of the predictor summarized at the top panel of Fig. 4.2, it can be seen that it was highly accurate in an East Coast region covering about 10 paths followed by deteriorated accuracy as it was applied west (see Fig. 4.3). Accuracy of the global SVM trained on cycle 1 and semi-global SVMs with SSVP over shorter regions was compared next. The results over 4 cycles and propagation over 47, 24, 16, 8 and 4 paths are reported in Fig. 4.4 (propagation over 47 paths at cycle 1 corresponds to Fig. 4.3 results). The overall highest accuracy, which was even slightly better than that of the global SVM, was achieved by SSVP over blocks of 8 neighboring paths. In these experiments, SVMs trained with SSVP after every 8 paths are applied for out-of-sample prediction at the corresponding 8 adjacent paths. Consequently, only six semi-global SVMs were required to cover each data cycle.

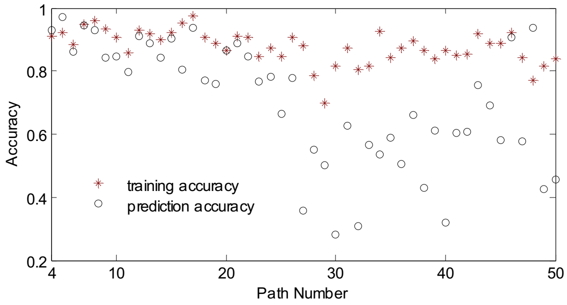

Figure 4.3. Path-specific training vs. prediction accuracy of the SVM trained with SSVP on Cycle 1. SSVP was from West to East (paths 50 to 4).

Finally, the computational efficiency of SVM training and prediction in a sequence of paths of MISR datasets with SSVP was compared to that of training a single SVM on the corresponding data. Time comparison experiments were performed on Cycle 1 data with the least number of missing values and the most balanced number of instances per path (see Table 3.1). A fixed number of 1500 instances were randomly sampled with replacement from each path (i.e. each instances may be included more than once). Both the training and the prediction CPU time was insensitive to the length of the data stream when learning in a series of local steps with

SSVP as proposed here. In contrast, for training a regular global model with merged data, the training time was a quadratic function of the number of paths N, and the prediction time increased linearly with N due to the increased size of training data (see Fig. 4.5).

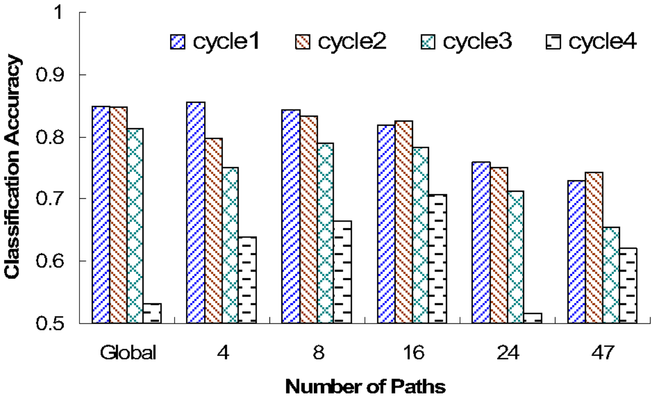

Figure 4.4. Classification accuracy on cycles 1 to 4 with global and semi-global SMVs trained on cycle 1 with SSVP over blocks of 4, 8, 16, 24 and 47 paths. Global SVM was trained with 70% instances of cycle 1.

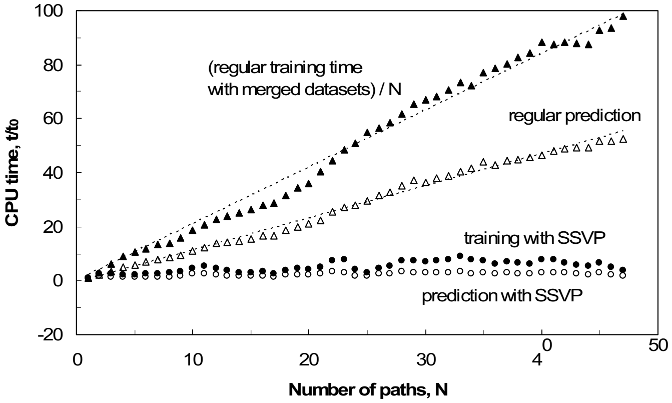

Figure 4.5. Comparison of normalized training and prediction CPU time, as a function of the number of paths N covered in training, for SVMs trained with SSVP vs. regular SVM trained using merged raw data from all paths. Training was from west to east (path 50 to 4) using the data of Cycle 1, and prediction was for each individual path only. CPU time for regular training with merged data is presented normalized by the number of paths as to show a quadratic growing curve together with a linear growing CPU time curve of regular prediction and near-constant curves obtained when training and testing with SSVP.

5. CONCLUSIONS

The proposed method of sequential SVM training with SSVP was developed for supervised learning in a non-stationary series consisting of correlated spatial datasets. The objective was to obtain accurate semi-global prediction models with low computation cost. A simple support vector selection method was applied at each propagation step to prevent monotonic growth of the number of propagated support vectors over long sequences. The validity of the proposed method was demonstrated on simple synthetic and challenging geoinformatics spatial-temporal datasets. It was found that to predict within larger spatial and temporal neighborhoods, the SVM trained on a sequence of local spatially correlated datasets with SSVP from a neighboring region can be superior to reusing local models trained with individual datasets.

Regional features discovered using the proposed approach allowed spatial partitioning of MISR aerosol data over the continental U.S. to western and eastern regions that provided more accurate prediction results. Furthermore, regional features were also found to be semi-stationary in time, such that the regional models trained on previous datasets were reusable for accurate regional predictions at future times. The proposed model reuse method was also demonstrated as a cost-effective alternative to merging raw data when learning SVMs from large spatial-temporal data streams.

REFERENCES

[1] Bothwell, G.W., Hansen, E.G., Vargo, R.E., and Miller K.C., ‘The Multi-angle Imaging SpectroRadiometer science data system, its products, tools, and performance’, *IEEE Trans. Geosci. Remote Sens.,* Vol. 40 No. 7, pp. 1467-1476, 2002.

[2] Vapnik, V., *Estimation of Dependences Based on Empirical Data: Nauka*, 1979.

[3] Peng, K., Obradovic, Z. and Vucetic, S., ‘Towards Efficient Learning of Neural Network Ensembles from Arbitrarily Large Datasets’, *Proc. 16th European Conf. Artificial Intelligence,* pp. 623-627, 2004.

[4] Boser, B. E., Guyon, I. M. and Vapnik, V. N., ‘A training algorithm for optimal margin classifiers’, *Proc. 5th Annual ACM Workshop on Computational Learning Theory,* pp. 144-152, 1992.

[5] Cortes, C. and Vapnik, V., *Support-Vector Networks: Machine Learning*, 20, pp. 273-297, 1995.

[6] Scholkopf, B. and Smola, A. J., *Learning with Kernels*, MIT Press, 2002.

[7] Vucetic, S. and Obradovic, Z., ‘Performance Controlled Data Reduction for Knowledge Discovery in Distributed Databases’, *Proc. 4th Pacific-Asia Conf. Knowledge Discovery and Data Mining,* pp. 29-39, 2000.

[8] Burges, C. J. C. and Schölkopf, B., ‘Improving the accuracy and speed of support vector learning machines’, in Mozer, M., Jordan, M. and Petsche, T., eds., A*dvances in Neural Information Processing Systems 9*, pp. 375-381, MIT Press, Cambridge, MA, 1997.

[9] Osuna, E. E. and Girosi, F., ‘Reducing the run-time complexity in support vector machines’, in Burges, C., Schölkopf B., and Smola, A. eds., *Advances in Kernel Methods: Support Vector Learning*, pp. 271–284. MIT Press, Cambridge, MA, 1999.

[10] Osuna, E. E., Freund, R. and Girosi, F., ‘An improved training algorithm for support vector machines’, *Proc. 1997 IEEE Workshop,* pp. 276-285, 1997.

[11] Fung, G. and Mangasarian, O. L., ‘Proximal Support Vector Machine Classifiers’, *Proc. 7th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining,* pp. 77-86, 2001.

[12] Nittel, S. and Leung, K. T., ‘Parallelizing Clustering of Geoscientific Data Sets using Data Streams,’ *Proc. 16th In’l Conf. Scientific and Statistical Data Base Management,* pp. 73-84, 2004.

[13] DeCoste, D. and Mazzoni, D., ‘Fast Query-Optimized Kernel Machine Classification Via Incremental Approximate Nearest Support Vectors’, *Proc. 20th Int’l Conf. Machine Learning,*pp. 115-122, 2003.