International Conference on Computational Science (ICCS) 2007
Beijing, China
May 27-30, 2007

Active Learning with Support Vector Machines for Tornado Prediction

Theodore B. Trafalis[1], Indra Adrianto[1], Michael B. Richman[2]

[1]School of Industrial Engineering, University of Oklahoma, 202 West Boyd St, Room 124, Norman, OK 73019, USA.
ttrafalis@ou.edu, adrianto@ou.edu

[2]School of Meteorology, University of Oklahoma, 120 David L. Boren Blvd, Suite 5900, Norman, OK 73072, USA.
mrichman@ou.edu

Introduction

- Most conventional learning methods use static data in the training set to construct a model or classifier.
- Active learning has an ability to update the model dynamically using new incoming data.
- The objective of active learning for classification is to choose the instances or data points to be labeled and included in the training set.
- In many machine learning tasks, collecting data and/or labeling data to create a training set is costly and time-consuming.

Introduction (cont.)

- In tornado prediction, labeling data is considered costly and time consuming since we need to verify which storm-scale circulations produce tornadoes in the ground.
- The tornado events can be verified from facts in the ground including photographs, videos, damage surveys, and eyewitness reports.
- Based on tornado verification, we then determine and label which circulations produce tornadoes or not.

Introduction (cont.)

- Applying active learning for tornado prediction to minimize the number of instances and use the most informative instances in the training set in order to update the classifier would be beneficial.

Objectives

- To investigate the application of active learning with SVMs for tornado prediction using the Mesocyclone Detection Algorithm (MDA) and Near-Storm Environment (NSE) data.
- To compare this method to passive learning with SVMs where the next instances to be added to the training set are randomly selected using these data.

MDA & NSE

- Mesocyclone Detection Algorithm (MDA) (Marzban and Stumpf, 1996)
  - The MDA attributes measure radar-derived velocity parameters that describe various aspects of the mesocyclone.
- Near-Storm Environment (NSE) (Lakshmanan et al., 2005)
  - The NSE data described the pre-storm environment of the atmosphere on a broader scale than the MDA data, as the MDA attributes are radar-based.
  - Information on wind speed, direction, wind shear, humidity lapse rate and the predisposition of the atmosphere to accelerate air rapidly upward over specific heights were measured in the NSE data.

Data and Analysis

- The original data set was comprised of 23 attributes taken from the Mesocyclone Detection Algorithm (MDA) data set.
- Incorporate 59 attributes from the Near-Storm Environment (NSE) data to the MDA data set.
- Therefore, the MDA+NSE data consist of 82 attributes + 1 class attribute.

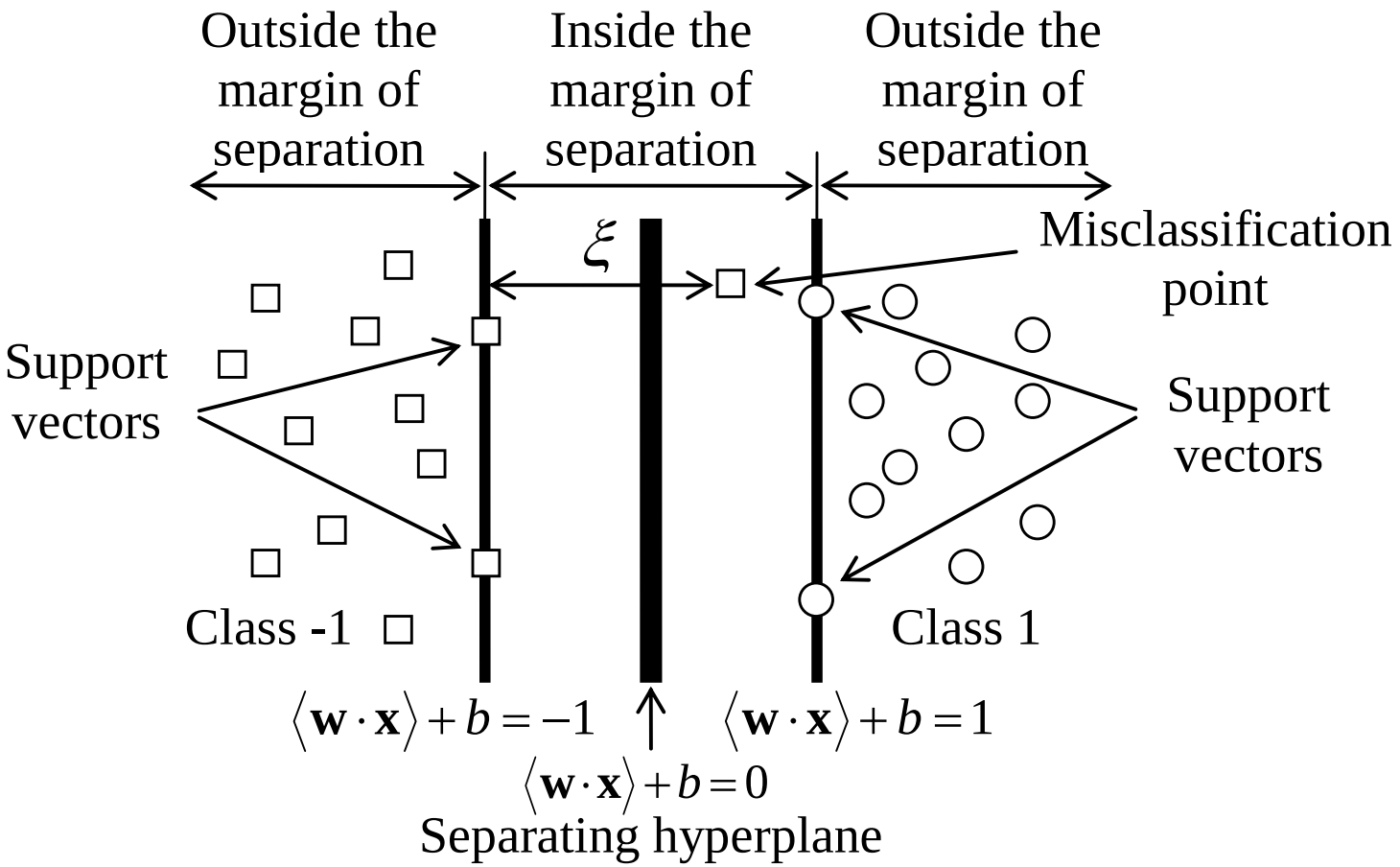
Methodology

- Support Vector Machines (SVMs)
- Active Learning with SVMs
- Measuring the Quality of the Forecasts for Tornado Prediction

Support Vector Machines (SVMs)

- The SVM algorithm was developed by Vapnik and has become a powerful method in machine learning (Boser et al., 1992; Vapnik, 1995, 1998).
- The objectives of SVMs (the primal problem) are to maximize the margin of separation and to minimize the misclassification error.

SVMs (cont.)



- **Fig 1.** Illustration of SVMs.

Active Learning with SVMs

- Several active learning algorithms with SVMs have been proposed by Campbell et al. (2000), Schohn and Cohn (2000), and Tong and Koller (2001).
- Campbell et al. (2000) suggested that the generalization performance of a learning machine can be improved significantly with active learning.
- Using SVMs, the basic idea of active learning algorithms is to choose the unlabeled instance for the next query closest to the separating hyperplane in the feature space.

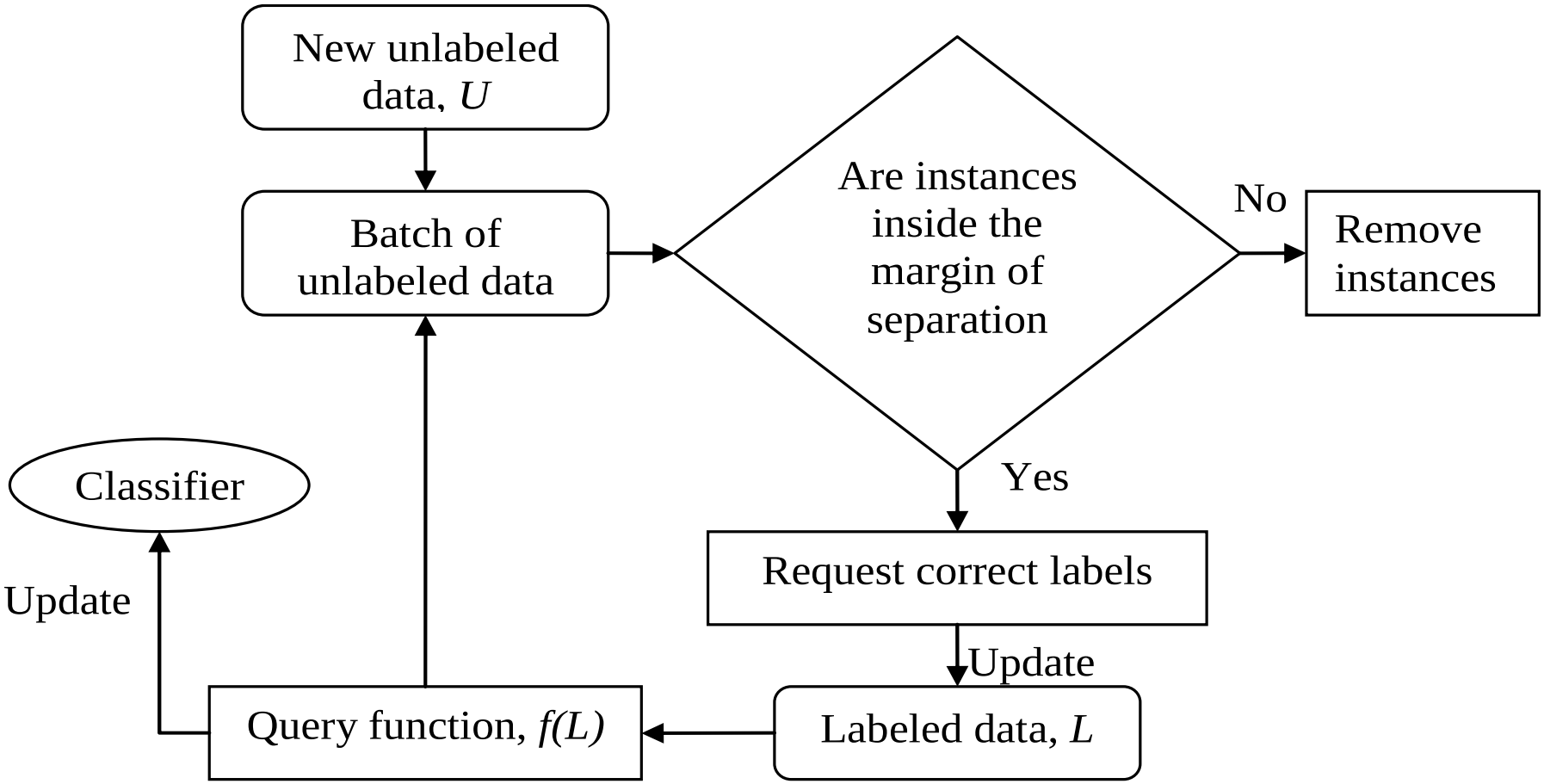
Active Learning with SVMs (cont.)

- In this paper, we choose the instances that are inside the margin of separation to be labeled and included in the training set.
- Since the separating hyperplane lies in the middle of the margin of separation, these instances will have an effect on the solution.
- Thus, the instances outside the margin of separation will be removed.

Active Learning with SVMs (cont.)

- Suppose we are given an unlabeled pool *U* and a set of labeled data *L*. The first step is to find a query function *f(L)* where, given a set of labeled data *L*, we need to determine which instances in *U* to query next.
- This idea is called the pool-based active learning.

Active Learning with SVMs (cont.)



- **Fig. 2.** Active learning with SVMs scheme.

Measuring the Quality of the Forecasts for Tornado Prediction

- In order to measure the performance of a tornado prediction classifier, it is important to compute scalar forecast evaluation scores such as the Critical Success Index (CSI), Probability of Detection (POD), False Alarm Ratio (FAR), Bias, and Heidke Skill Score (HSS), based on a “confusion” matrix or contingency table (Table I).

Measuring the Quality of the Forecasts for Tornado Prediction (cont.)

- **Table 1.** Confusion matrix.

		Observation	
		Yes	**No**
Forecast	**Yes**	hit a	false alarm b
	No	miss c	correct null d

Measuring the Quality of the Forecasts for Tornado Prediction (cont.)

- CSI = a/(a+b+c)
- POD = a/(a+c)
- FAR = b/(a+b)
- Bias = (a+b)/(a+c)
- HSS = 2(ad-bc)/[(a+c)(c+d)+(a+b)(b+d)]

Experiments

- The data were divided into two sets: training and testing.
- In the training set, we had 382 tornadic instances and 1128 non-tornadic instances.
- In order to perform online setting experiments, the training instances were arranged in time order.
- The testing set consisted of 387 tornadic instances and 11872 non-tornadic instances.

Experiments (cont.)

- For both active and passive learning experiments, the initial training set was the first 10 instances consisted of 5 tornadic instances and 5 non-tornadic instances.
- At each iteration, new data were injected in a batch of several instances.
- Two different batch sizes, 75 and 150 instances, were used for comparison.

Experiments (cont.)

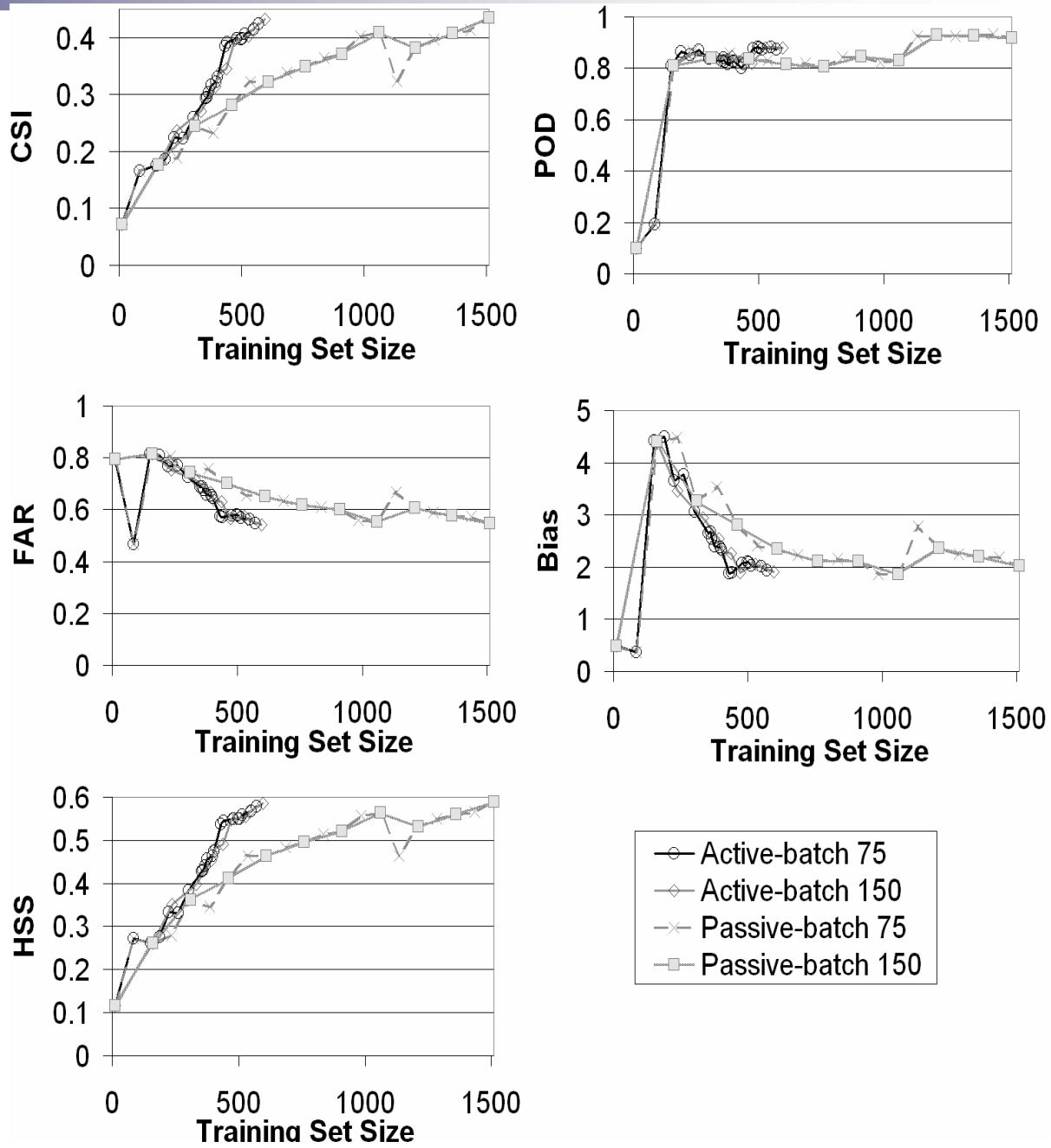
- In passive learning with SVMs, all incoming data were labeled and included in the training set.
- Conversely, active learning with SVMs only chooses the instances from each batch which are most informative for the classifier. Therefore, the classifier was updated dynamically at each iteration.

Experiments (cont.)

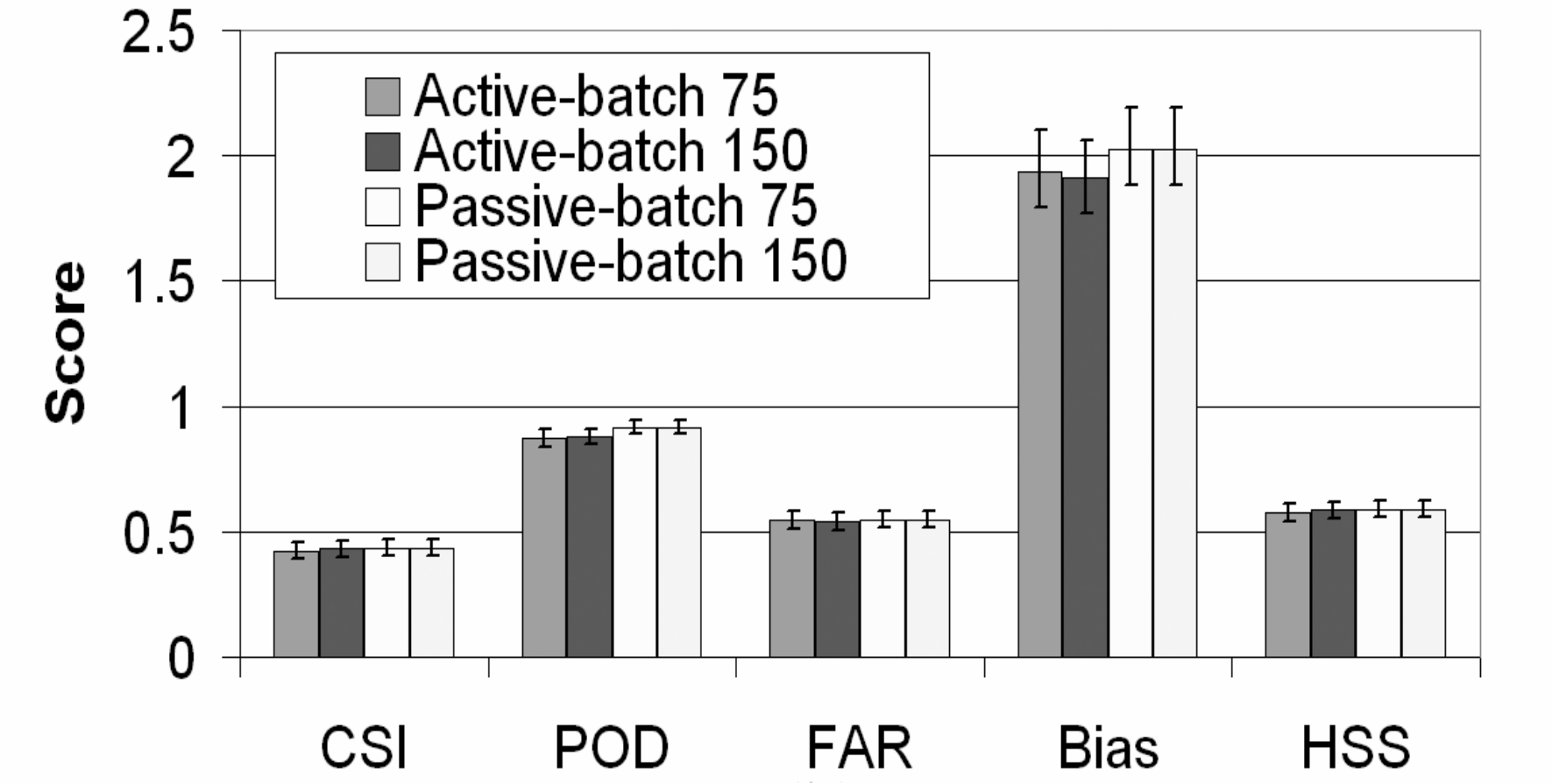
- The performance of the classifier can be measured by computing the scalar skill scores on the testing set.
- The radial basis function kernel with γ = 0.01 and C = 10 was used in these experiments.
- The experiments were performed in the Matlab environment using LIBSVM toolbox (Chang and Lin, 2001)

Results

- **Fig 3.** The results of CSI, POD, FAR, Bias, and HSS on the testing set using active and passive learning at all iterations.

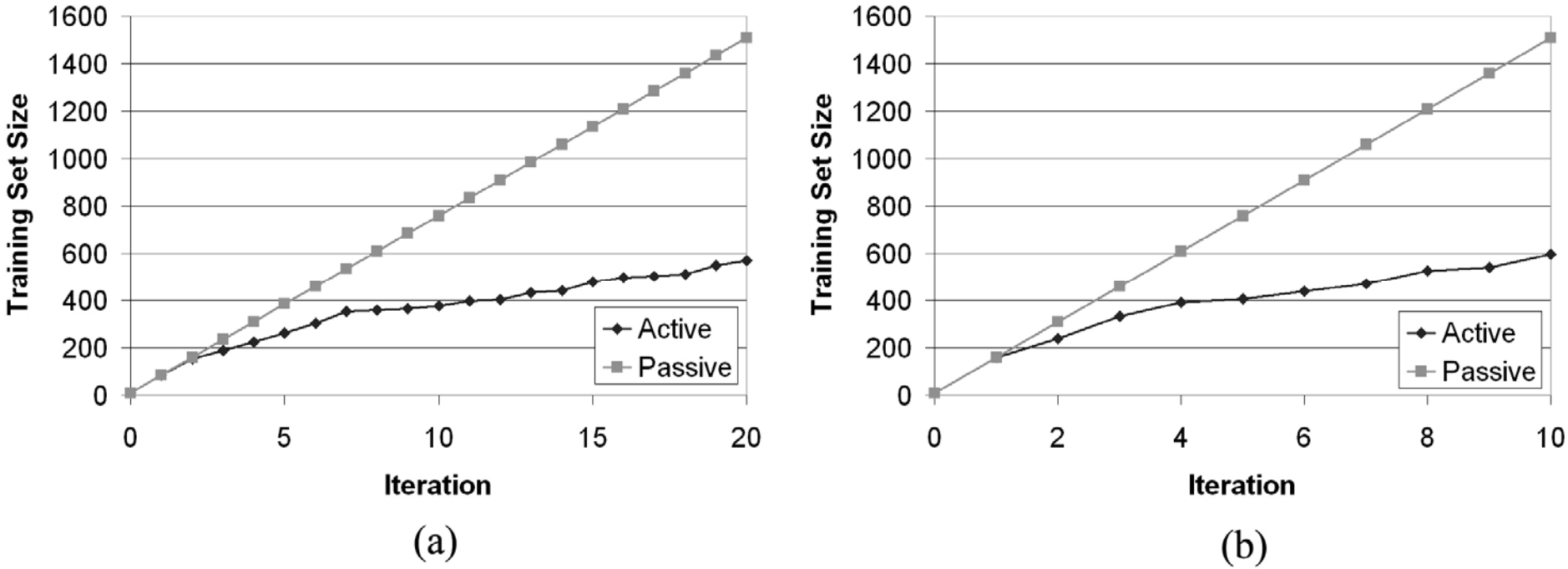


Results (cont.)



- **Fig 4.** The last iteration results with 95% confidence intervals on the testing set.

Results (cont.)



- **Fig 5.** Diagrams of training set size vs. iteration for the batch sizes of (a) 75 and (b) 150 instances.

Results (cont.)

- The results showed that active learning significantly reduced the training set size to attain relatively the same skill scores as passive learning.
- Active learning with SVMs reduces the training set size by 62.6% and 60.5% using the batch size of 75 and 150 instances, respectively.

Conclusions

- In this paper, active learning with SVMs was used to discriminate between mesocyclones that do not become tornadic from those that do form tornadoes.
- The preliminary results showed that active learning can significantly reduce the size of training set and achieve relatively similar skill scores compared to passive learning.
- Since labeling new data is considered costly and time consuming in tornado prediction, active learning would be beneficial in order to update the classifier dynamically.

Acknowledgments

- Funding for this research was provided under the National Science Foundation Grant EIA-0205628 and NOAA Grant NA17RJ1227.

References

- Marzban, C., Stumpf, G.: A neural network for tornado prediction based on Doppler radar-derived attributes. J. Appl. Meteorol. 35 (1996) 617-626
- Lakshmanan, V., Stumpf, G., Witt, A.: A neural network for detecting and diagnosing tornadic circulations using the mesocyclone detection and near storm environment algorithms. In: 21st International Conference on Information Processing Systems, San Diego, CA, Amer. Meteor. Soc. (2005) CD–ROM J5.2
- Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Haussler D (ed): 5th Annual ACM Workshop on COLT. ACM Press, Pittsburgh, PA (1992) 144-152
- Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer Verlag, New York (1995)
- Vapnik, V.N.: Statistical Learning Theory. Springer Verlag, New York (1998)
- Campbell, C., Cristianini, N., Smola, A.: Query learning with large margin classifiers. In: Proceedings of ICML-2000, 17th International Conference on Machine Learning. (2000)111-118
- Schohn, G., Cohn, D.: Less is more: Active learning with support vector machines. In: ICML Proceedings of ICML-2000, 17th International Conference on Machine Learning, (2000) 839-846
- Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2 (2001) 45-66
- Chang, C., Lin, C.: LIBSVM: a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (2001)