The Effects of Outliers on Support Vector Machines

Josh Hoak

Portland State University

13 April 2010

Previously...

1. Recall that a Fuzzy Support Vector Machine (FSVM) is a SVM that takes *weighted* training examples.

$$(y_1, \mathbf{x}_1, s_1), \ldots, (y_l, \mathbf{x}_l, s_l), \quad \sigma \leq s_i \leq 1, \quad (\sigma > 0)$$

2. Idea: Outliers receive lower confidence (weight) than non-outliers.

Previously...

Table 1: Error rates for SVMs and FSVMs using KT and k-NN

	SVM	KT	k-NN
Banana	11.5 $\pm$ 0.7	*10.4 $\pm$ 0.5	11.4 $\pm$ 0.6
B. Cancer	26.0 $\pm$ 4.7	25.3 $\pm$ 4.4	*25.2 $\pm$ 4.1
Diabetes	23.5 $\pm$ 1.7	*23.3 $\pm$ 1.7	23.5 $\pm$ 1.7
German	23.6 $\pm$ 2.1	*23.3 $\pm$ 2.3	23.6 $\pm$ 2.1
Heart	16.0 $\pm$ 3.3	*14.2 $\pm$ 3.1	23.6 $\pm$ 2.1
Image	3.0 $\pm$ 0.6	*2.9 $\pm$ 0.7	-
Ringnorm	*1.7 $\pm$ 0.1	-	-
F. Solar	*32.4 $\pm$ 1.8	32.4 $\pm$ 1.8	32.4 $\pm$ 1.8
Splice	*10.9 $\pm$ 0.7	-	-
Thyroid	4.8 $\pm$ 2.2	*4.$\pm$ 2.3	-
Titanic	22.4 $\pm$ 1.0	*22.3$\pm$0.9	*22.3 $\pm$ 1.1
Twonorm	3.0 $\pm$ 0.2	*2.4 $\pm$ 0.1	2.9 $\pm$ 0.2
Waveform	*9.9 $\pm$ 0.4	9.9 $\pm$ 0.4	-

13 data sets from the UCI, DELVE and STAT- LOG

Note: We optimize the SVM first and then choose the parameters for the FSVM!

Often, the error of the SVM is within the conf. interval for the FSVM error.

Why does the FSVM do only slightly better?

Some questions to consider:

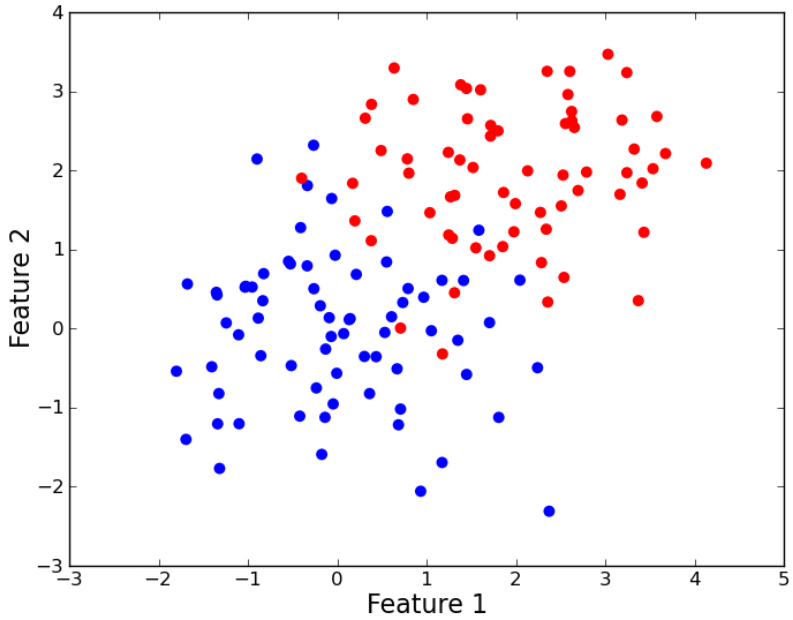
- To what degree do outliers affect SVM models?
- By how much can we improve our SVM models in ideal and real-world situations?
- What methods should we use to deal with outliers?

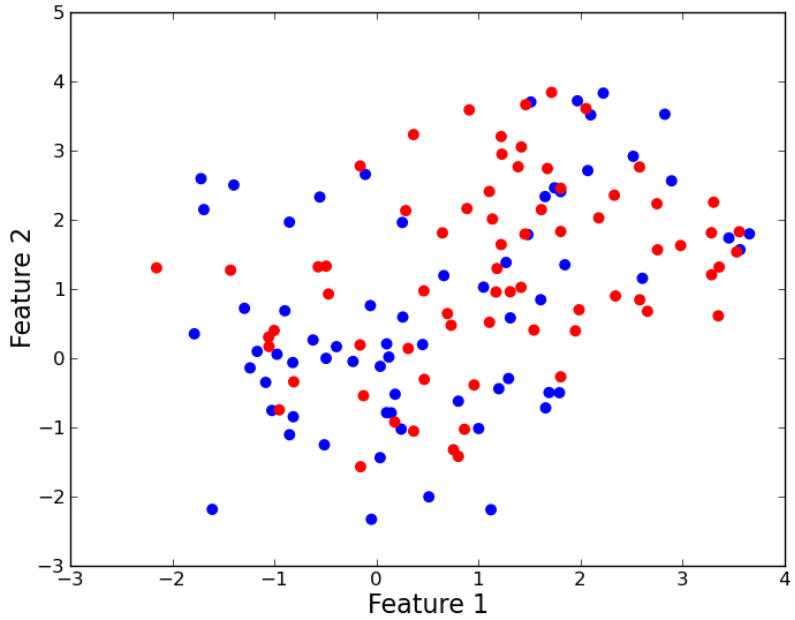
My Research

My research consisted of three parts:

1. Artificial Data – Outlier and kernel experiments
   - Created normally-distributed data using R.
   - Using LIBSVM, trained with different fractions of the examples set as outliers, which were made outliers by flipping the class.
   - Tried different settings of outliers and kernels
   - Tested on data without outliers.
2. Artificial Data – Outlier, kernel, and weight experiments.
   - LIBSVM has a *weighted* training tool. I set the weight w of each outlier to be between 0 and 1.
   - For my experiments, all outliers received the same weight for a given round.
3. Real-world Data – Outlier, kernel, and weight experiments.
   - Used several two-class data sets available on the web and then performed the same tests as above.

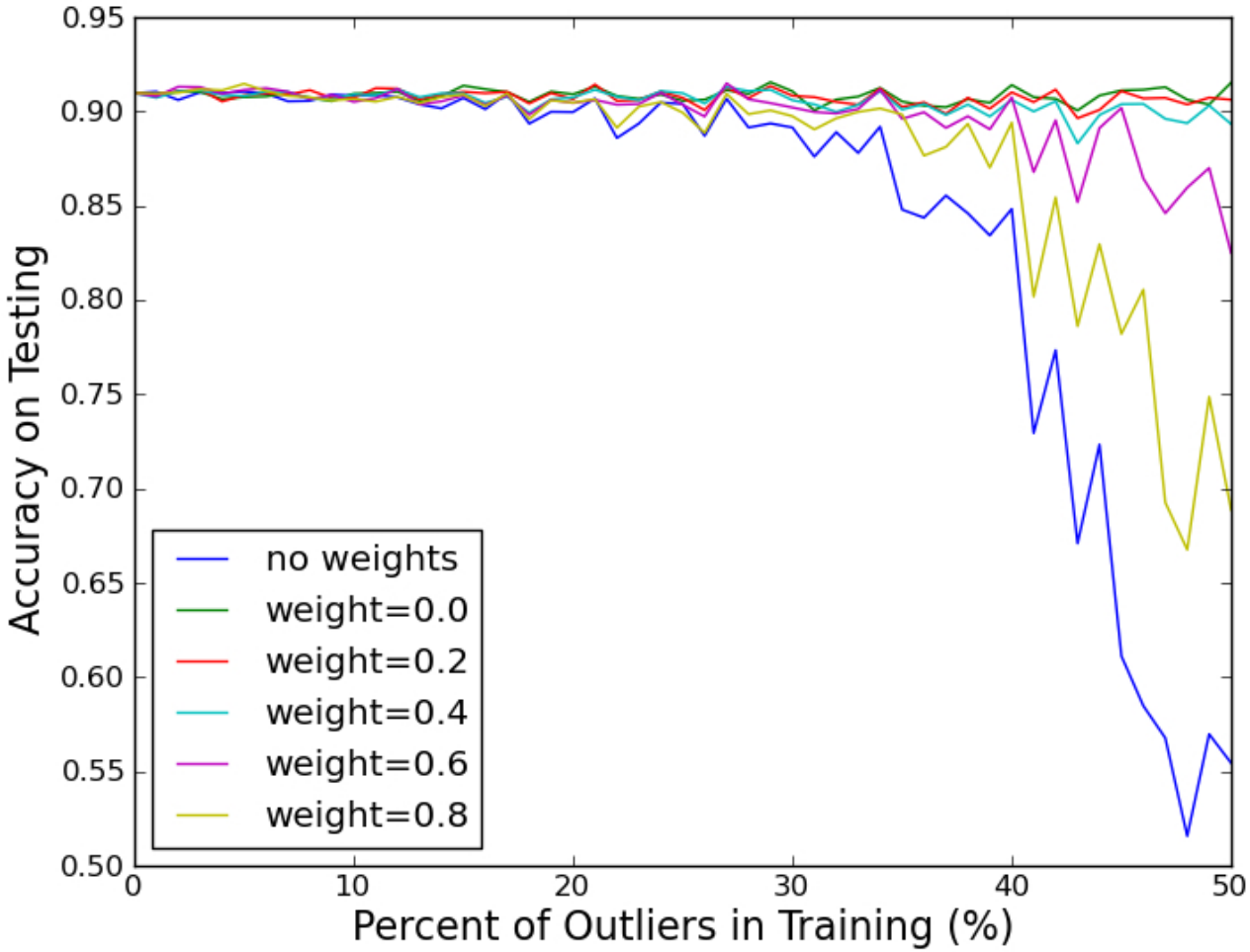
Initial Data





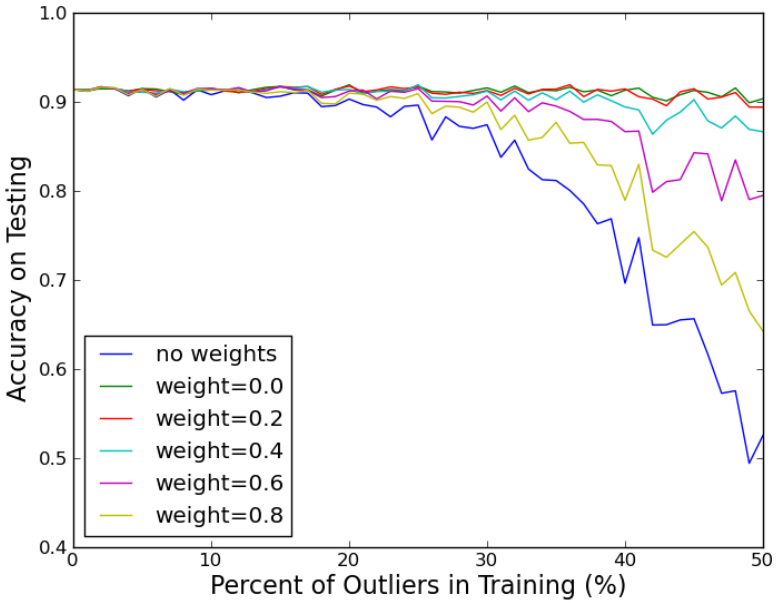
Two examples of initial data. On the left 0% of the examples are outliers, while on the right 30% of the examples are outliers.

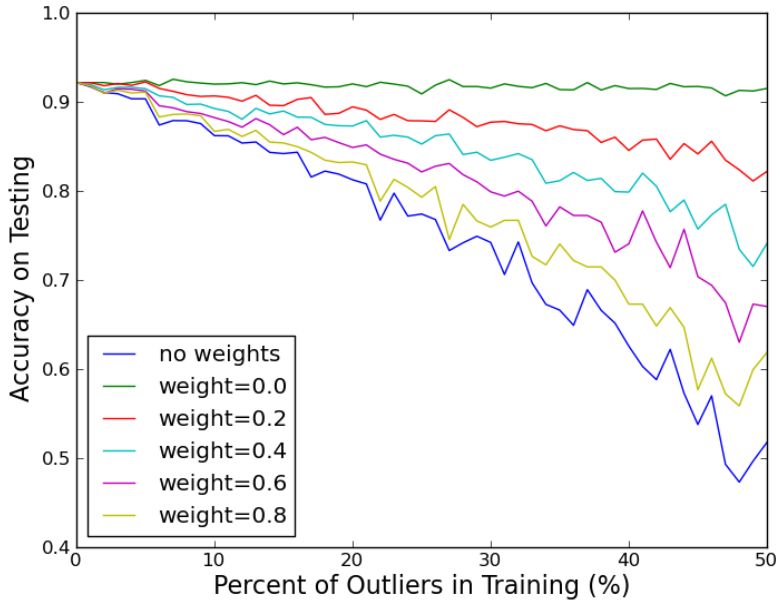
Linear Kernel: Results



Displayed are the averaged results of 20 data sets and using 51 different outlier percentages

Radial Basis Function and Polynomial (Cubic) kernels.





Left diagram uses an RBF kernel and right uses a cubic kernel. Displayed are the averaged results of 20 data sets and and using 51 different outlier percentages.

Breast Cancer Data

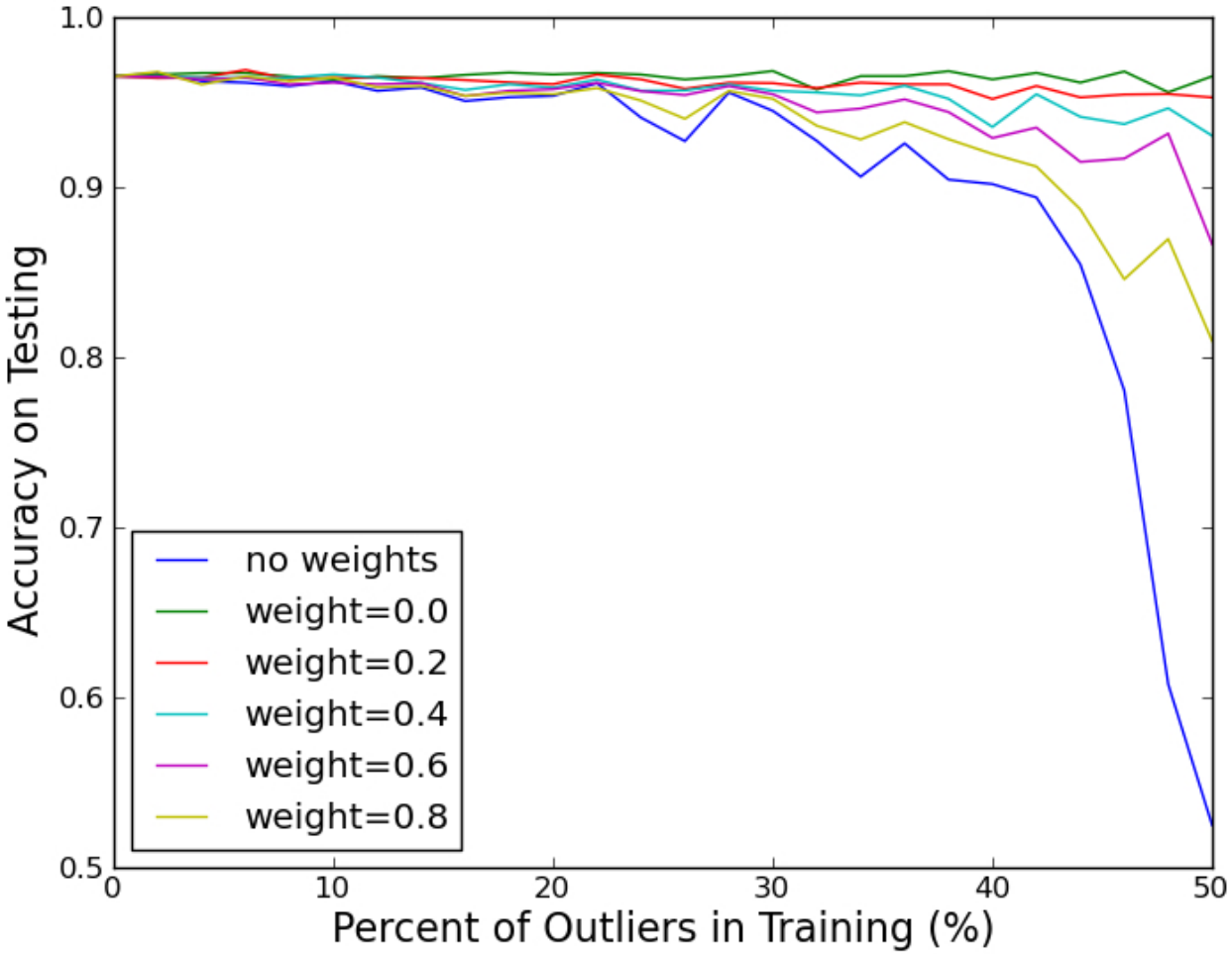
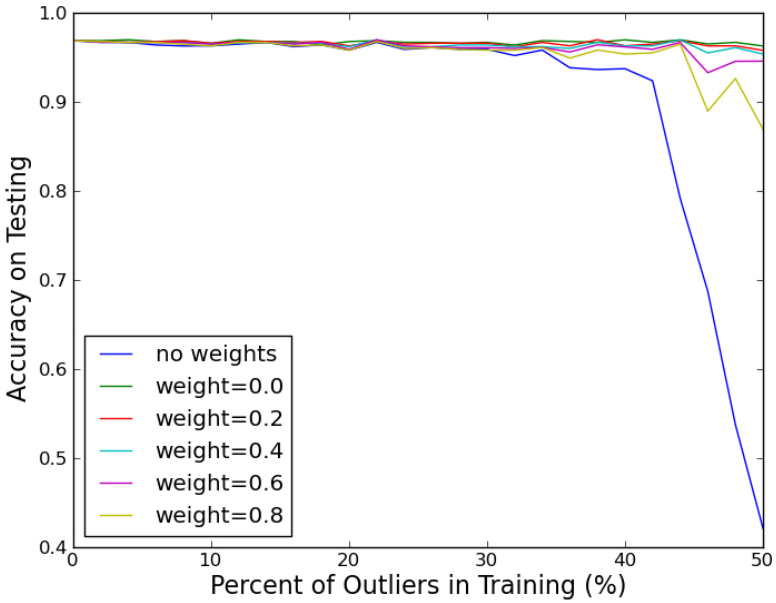
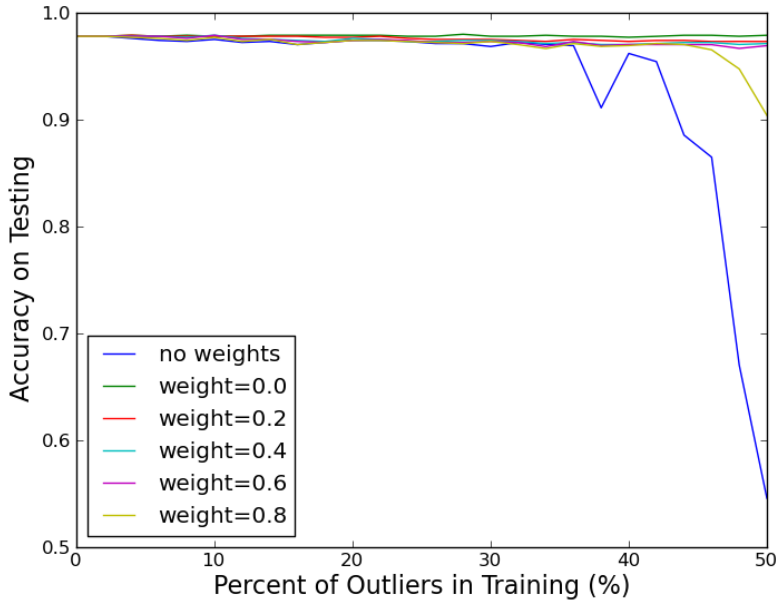


Diagram uses a linear kernel. Displayed are the averaged results of 5 data runs and using 21 different outlier percentages.

Breast Cancer Data





Left diagram uses an RBF kernel and right uses a cubic kernel. Displayed are the averaged results of 5 data runs and using 21 different outlier percentages.

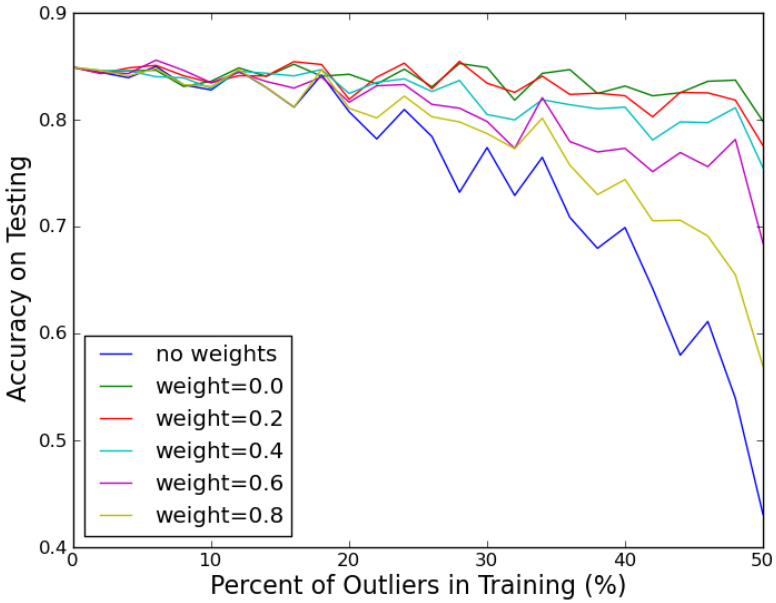
Heart Data

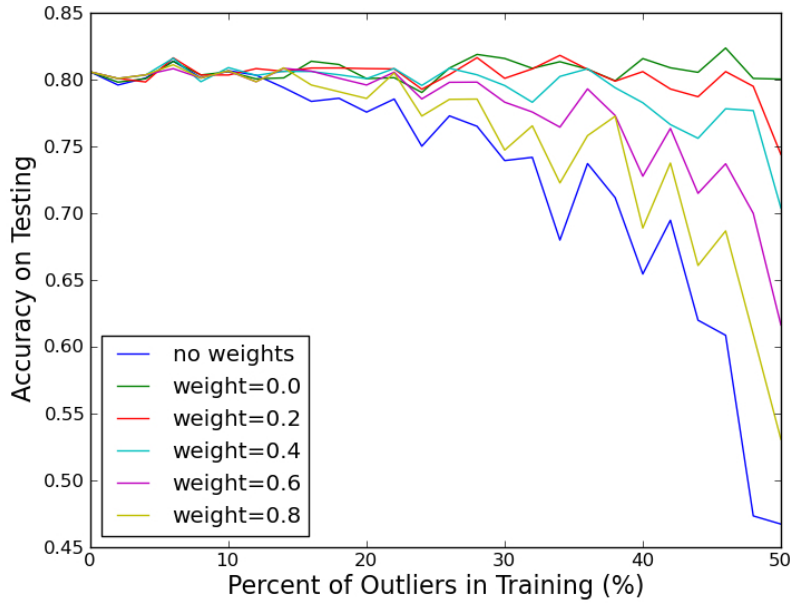
Recall from the FSVM results ...

	SVM	KT	k-NN
Heart	16.0 ± 3.3	$^*14.2 \pm 3.1$	23.6 ± 2.1

Note: The researchers were using an RBF kernel.

Heart Data





Left uses a linear kernel and right uses an RBF kernel. Averaging the results of 5 data runs and using 21 different outlier percentages.

End