Screening rules for SVM

K Parimala ee15m17p100001@iith.ac.in
D Sri Krishna Priya cs14btech11010@iith.ac.in

Objective

The objective of this project is to implement fast and efficient screening rules to discard non support vectors by analyzing the dual problem of SVM via variational inequalities. The number of data samples entering into the optimization is reduced. This speeds up the training process of SVM.

1 Literature survey

The classifier for screening is determined only by support vectors of SVM. Removing the non-support vectors from the optimization may lead to substantial savings in the computational cost and memory.

1.1 SSNSV - Ogawa et al. (2013)

Safe Screening rule to identify non-support vectors for SVM. If the resulting classification model is different after applying data reduction for SVM, then that method is not safe. In order to run the screening, SSNSV needs to iteratively determine an appropriate parameter value and an associated feasible solution, which can be very time consuming.

2 Proposed Method

2.1 DVI (Jie Wang et al.(2014))

It is a set of novel, effective and efficient screening rules for SVM. It identifies the non-support vectors by estimating a lower bound of the inner product between each vector and the (unknown) optimal solution. As the optimal solution is unknown, the estimation is non-trivial.

The proposed framework to accurately estimate the optimal solution via the estimation of the "dual optimal solution" overcomes this difficulty as the primal and dual optimal solutions can be related by the KKT conditions.

2.2 Problem Formulation

SVM is a maximum margin classifier.
Primal :

$$\min_{w} \frac{1}{2} ||w| |^2 + C \sum_{i=1}^{l} [1 - w^T (y_i x_i)]_+ \tag{1}$$

x_i is the i^{th} input data instance.
$y_i \in$ {-1, 1} is the corresponding label
Margin:

$$f(x) = sign(w^T x)$$

Dual:

$$\min_{\theta}\{\tfrac{C}{2}||\bar{X}\theta||^2 - \textstyle\sum_{i=1}^{l}\theta_i : \theta_i \in [0,1], i = 1,\ldots,l\}$$

KKT conditions:

$w^*(C) = C\bar{X}(C)$, where $w^*(C)$ and $\theta^*(C)$ are the optimal primal and dual solutions of SVM.

2.3 Screening Rules

If i ∈ R = {k: $\left\langle C\bar{X}\theta^*(C)|\bar{x}_k\right\rangle > 1$}, then $[\theta^*(C)]_i$ = 0 and x_i can be removed from the optimization.
If i ∈ L = {k: $\left\langle C\bar{X}\theta^*(C)|\bar{x}_k\right\rangle < 1$}, then $[\theta^*(C)]_i$ = 1 and no need to compute $[\theta^*(C)]_i$
Here, $\theta^*(C)$ is unknown.

2.4 Estimation of dual optimal solution ($\theta^*(C)$)

If we can estimate a region for $\bar{X}\theta^*(C)$, we can still determine subsets of R and L. Suppose that $\bar{X}\theta^*(C) \subset \Xi$, then the following are the relaxed guidelines for screening out samples.

If $\min_{\xi}\{\{C\xi|\bar{x}_i\} : \xi \in \Xi\} > 1$, then i ∈ R, $[\theta^*(C)]_i$ = 0 and x_i can be removed from the optimization.

If $\min_{\xi}\{\{C\xi|\bar{x}_i\} : \xi \in \Xi\} < 1$, then i ∈ L, $[\theta^*(C)]_i$ = 1 and no need to compute $[\theta^*(C)]_i$

3 Datasets and Performance metrics

3.1 Datasets

1. IJCNN1 data set (Prokhorov, 2001)
2. Wine Quality data set (Cortez et al., 2009)
3. Forest Covertype data set (Hettich & Bay, 1999)
4. 3 Synthetic Data Sets : Toy1, Toy2, Toy3
   For each data set, we generate two classes.
   Each class has 1000 data points and is generated from $N(\{\mu,\mu\}^T, 0.75^2 I)$, where I ∈ R^{2x2} is the identity matrix. For the positive classes,mu = 1.5, 0.75, 0.5, for Toy1, Toy2 and Toy 3, respectively; and μ = -1.5, -0.75, -0.5, for the negative classes.
   Observations: As |μ| decreases, the two classes increasingly overlap and thus the number of data instances belong to the set L increases.

3.2 Performance Metrics using rejection rates

Rejection rate equals the ratio of the number of data instances whose membership can be identified by the rules to the total number of data instances.
Running time is also used as a metric to measure results.

4 Results

Using synthetic data sets, we need to show that DVIs are very effective in discarding non-support vectors even for largely overlapping classes.
Using Real Data Sets, we'll compare the performance of SSNSV, ESSNSV and DVIs in terms of the rejection ratio. Most import result is that DVIs rules identify far more non-support vectors than SSNSV and ESSNSV.

References

[1] Jie Wang ,Peter Wonka & Jieping Ye (2014) Scaling SVM and Least Absolute Deviations via Exact Data Reduction *Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):523-531, 2014.*

[2] Jieping Ye *Sparse Screening for Exact Data Reduction*