Chapter 2

Related work

In this chapter, the literature related to robust statistics in SVM and its variants is discussed.

2.1 Robust Statistics based SVM

Since the SVM origin, researchers had a keen interest in this classifier/regressor, which originated several new researches in this area. They moved toward testing the robustness of this classifier on various applications. On finding the limitation of sensitivity toward outliers, researchers started proposing algorithms that can make the model robust like the one given by Le Thi Hoai and Tao [38] in which they proposed a truncation algorithm that can be used for both binary and multiclass classification. That method of truncation can be applied to any convex unbounded loss function that usually suffers because of outliers in the dataset.

Similarly, another technique was proposed to deal with outliers in the dataset, which considered the distance of each data point from the center of the class to evaluate the adaptive margin and then added the averaging technique into the classical SVM [39]. This addition helped them in getting a robust SVM, and it also controlled the amount of regularization. It again proved to be very useful as the number of support vectors

obtained was also significantly less. Song et al. [39] concentrated on bullet hole classification and applied their algorithm over it. This was the time when robust statistics started developing in the field of SVM. Several outlier detection algorithms were also proposed, like an outlier detection method for industrial data, which is otherwise difficult to find solving regression problems [40]. The primary objective of that technique is to find outliers even in high-dimensional datasets and with rank deficient datasets.

Several non-convex optimization problems were also proposed to get the robust regressors like the one given by Zhao and Sun, which also worked in robust regression using a non-convex loss function for SVR. They made use of Huber loss [41]. This non-convex optimization problem was turned into the convex function using CCCP, and Zhao et al. [41] solved this convex problem with a Newton-type algorithm to get robust SVR.

In 2017, another robust SVM for multi-class classification was proposed (multi-class SVM can be referred from Angulo et al. [42]. This method was based on ramp loss K-support vector classification-regression [43]. This method was solved using the CCCP procedure as the final optimization problem after the ramp loss function was non-differentiable and non-convex.

Recently, Zhu et al. [44] extended the application of pinball SVM to SVM+ and named it as PINSVM+. SVM+ and SVM have a difference that SVM+ consider additional information which is hidden in the training samples. This makes the model robust to outliers and also more stable for re-sampling. Again, both the PINSVM+ and SVM+ had the same computational complexity. The optimization problem corresponding to PINSVM+ is given in [44].

2.2 Robust Statistics based Variants of SVM

It was observed from the work of Suykens et al. [45] that the standard LS-SVM has two significant limitations of sparseness and robustness in the proposed framework.

These limitations were observed and solved by Suykens et al. by presenting weighted LS-SVM [45]. To obtain the robust version of LS-SVM, error variables ξ_i were weighted using weighting factor ν_i and the resultant optimization problem is [45]:

$$\begin{aligned} \min_{w,b,\xi} \ & \frac{1}{2}\|w\|^2 + \frac{\lambda}{2}\sum_{i=1}^{m}\nu_i\xi_i^2 \\ \text{s.t.} \ & y_i(w^T x_i + b + \xi_i), \ i = 1, 2, \ldots, m. \end{aligned} \tag{2.1}$$

Suykens et al. also proposed a technique to make the weighted LS-SVM sparse using pruning methods [45]. They suggested that pruning a relatively small amount of least meaningful data points can make a sparse approximation [45]. They used this technique for heavy-tailed Non-Gaussian error distribution and observed that their approach added robustness to the regression problems.

Later, this weighted LS-SVM approach was used and extended by various researchers to solve problems like overfitting and noise in the dataset. Although SVR had excellent robustness properties against noise, when the parameters indulged in the optimization problem are not selected correctly, overfitting may occur. The inclusion of outliers in the dataset may also lead to some severe overfitting. Chuang et al. [46] handled this situation by proposing a novel approach to robustify the SVR network. In that work, Chuang et al. have combined two methods: robust learning theory and SVR to form robust SVR (RSVR) networks. There were two phases proposed for the RSVR networks: The initial phase where the network structure and the initial network weight assigned were determined using SVR theory as shown:

$$\hat{y} = \sum_{i=1}^{s} wK(\bar{x}_i, \bar{x}) + b, \tag{2.2}$$

where $\hat{y}$ is the output variable, s denotes the number of support vectors and $K(\bar{x}_i, \bar{x})$ is the kernel function.

The proposal of LS-SVM opened the room for various recent researches. LS-SVM was also robustified in many different ways. Several loss functions were also added with LS-SVM, like ramp loss function to make it robust to outliers and noise [47]. The advantages of ramp loss function observed are:

(a) Ramp loss LS-SVM controlled the sparseness of LS-SVM.

(b) As the ramp loss function is robust to outliers, ramp LS-SVM could incorporate noise and outliers suppression explicitly.

The optimization problem corresponding to ramp LS-SVM [47] is:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 + \frac{\lambda}{2} R_{\varepsilon,t}(y_i f(x_i) - 1) \tag{2.3}$$

where $R_{\varepsilon,t}$ is the ramp loss function, defined by

$$R_{\varepsilon,t} = \begin{cases} (t-\varepsilon)^2, & |z| > t, \\ (|z|-\varepsilon)^2, & \varepsilon \leq |z| \leq t, \\ 0, & |z| < \varepsilon \end{cases} \tag{2.4}$$

The resultant non-convex problem of ramp LS-SVM was solved using CCCP. It was also observed that the proposed approach allows better parallelization.

Like the weighted version of LS-SVM, fuzzy weighted SVR was also proposed by Chuang [48] which can be used further to make fuzzy SVR more robust to outliers or noise. Fuzzy SVM was also combined with an outlier detection algorithm so that the resultant fuzzy SVM can be made insensitive to outliers as suggested by Lee et al. [49]. That approach was based on four steps: First, they applied the outlier detection algorithm on the given training data. Second, they calculated the membership value by using the fuzzy sigmoid model. Third, kernel parameter estimation was performed, which proceeded toward the last step of applying fuzzy SVM [49]. The advantage

of that approach is the outlier detection algorithm which made the model robust to outliers and gave better results than the earlier proposed methods. Fuzzy SVR was also made robust toward Gaussian noise by making use of the fuzzy triangular technique to represent fuzzy membership values as suggested by Wu and Law [50]. Similarly, based on c-means and even Mahalanobis distance, SVMs were also made robust using fuzzy logic [51].

Although weighted versions proved their robustness towards noise in the data sets, the efficient method of weight assignment has always been discussed. In 2013, researchers used association rules to assign weights. This weights assignment method was better than the previous one as the use of association rules prevented bias to the majority class [52]. This method was suitable for both noisy data as well as imbalanced data.

In the above discussed approaches, it was very much required to adequately select the parameters like weight parameters and fuzzy membership values. If these parameters are not chosen correctly, the algorithm may not perform as per expectation.

2.3 Robust Loss Functions in SVM and its variants

Various researchers have proposed novel approaches to add robustness to their algorithms by making some hinge loss function changes. The hinge loss function is convex, unbounded and its results get adversely affected because of outliers in the data sets. Therefore, the researchers have proposed various robust formulations in which robust loss functions are used to make the model insensitive towards noise or outliers. Towards this direction, the first is the introduction of the truncated hinge loss function, which is non-convex and can also handle outliers in the dataset [53]. This made the optimization problem, non-convex minimization problem. They used a non-convex optimization approach, Difference of Convex (DC) functions, to solve non-convex problems via a sequence of convex subproblems.

Similarly, another robust loss function, ramp loss function, was added to SVM/SVR to make the optimization problem robust. In 2008, Wang et al. also used the same technique of changing a non-convex function to the convex one using CCCP, but the non-convex function they used was ramp loss function [36]. The ramp loss function is insensitive to outliers. Wang et al. [36] used the Newton-type algorithm to solve the primal optimization and compared their results with existing robust algorithms using classification datasets. They got this idea of training SVM in the primal form from Chapelle et al. [54]. Huang et al. [55] also worked in the same direction as above mentioned. They also made use of the ramp loss function [56] to make the SVM more robust than previously existing methods but with a new addition of L_1 norm penalty to the optimization problem. The properties of the ramp loss function are that it is a non-convex smooth loss function and, at the same time, insensitive to outliers. They paired the ramp loss function with L_1 penalty to induce sparsity and named their method as Ramp-LPSVM (where LP stands for linear programming). They found that their algorithm is more robust than the standard Hinge loss function and ramp-SVM [36].

The next part discusses the use of another loss function with SVM, the pinball loss function. We have seen the hinge loss used with SVM in the classification problems. Although SVM with hinge loss works efficiently but hinge loss is not robust to outliers. Therefore, it was required to either change or modify the loss function to make it robust to outliers. The pinball loss function was initially used for regression problems. In 2014, the pinball loss function was used for the first time with SVM classifier [37]. Huang et al. named it pin-SVM. They proposed the optimization problem [37]:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 + \lambda \sum_{i=1}^{m} L_\tau(1 - y_i(w^T x_i + b)), \tag{2.5}$$

where L_τ is the pinball loss function and it is defined as

$$L_\tau(u) = \begin{cases} u, & u \geq 0, \\ -\tau u, & u < 0. \end{cases} \tag{2.6}$$

This work was extended to non-linear classification as well. The difference between hinge loss and pinball loss function is that the penalty was given to the correctly classified points by the pinball loss function. While both the loss functions had similar computational complexity, pinball is less sensitive to outliers, more stable for re-sampling than hinge loss function [37]. This work was also extended further by using truncated pinball loss function with SVM as proposed by Shen et al. [57]. Although robust to outliers, the Pinball loss function completely lost its sparseness with SVM, so a truncated pinball loss function was introduced so that the classifier can be both robust to outliers and much sparser than earlier approaches [57]. The resultant non-convex optimization problem was solved using CCCP. Truncating the loss functions was not a new concept in machine learning. Truncated least square loss function [58] (for regression problems) and truncated logistic loss [59] (for classification problems) also contributed to robust statistics.

The use of the robust loss function was extended to regression problems as well. In 2017, Yang et al. [60] also proposed a robust regression algorithm for non-convex loss functions based on Laplace Kernel-induced loss function. As the L_K loss function is non-convex, therefore, DC programming was applied first. The advantage of considering L_K-loss is its boundedness. In that paper, Yang et al. proposed new use of the L_K loss function. They decomposed it into two formulations: regression formulation, which was introduced as DC programming (LKRE), and the other one was obtained by weighting it with a suitable parameter. They named it a mixed loss function MLKRE. Researchers gave a more robust formulation of L_p-norm based least-squares SVR, which offers the non-convex optimization problem [61]. It provided a robust feature selection with faster and robust SVR compared to L_p-norm SVR and SVR method. Although the technique

proposed was effective, it was slower than L_1 SVR and LS-SVR. Another robust SVR approach was proposed using the least absolute deviation and to solve the optimization problem corresponding to it, and they used split Bregman iteration [62]. Cevikalp et al. [63] used the ramp loss function to make TSVM robust. This method can also be applied to large-scale data. The idea of using robust loss functions was extended to TWSVM as well. The concept of using various loss functions to the non-parallel support vector machines was first given by Mehrkanoon et al. [64] which was proposed for classifiers only. In that work, the objective function was comprised of a regularization term, a misclassification loss, and a scatter loss function. It was initially proposed for GEPSVMs, TWSVM, and also the least square variants of it. They considered the loss functions as were least-squares loss function and pinball loss function [64]. Because of the sensitivity to noise and instability for resampling [65], hinge loss was replaced. In 2017, pinball loss function was considered with TWSVM, which dealt with the quantile distance, and it is also less sensitive to noise [65].

Similarly, ramp loss function was also used to add robustness to TSVR [66]. It was observed that the use of ramp loss function not only adds robustness but also provides an excellent sparseness. For the two non-parallel hyperplanes of TSVR, ε_1-insensitive downbound function $(f_1(x) + \varepsilon_1)$ and ε_2-insensitive up-bound function $(f_1(x) - \varepsilon_2)$ provided two primal problems [66]:

$$\begin{aligned} \min_{w_1,b_1} \quad & \frac{1}{2}\|w\|^2 + \lambda_1 \sum_{i=1}^{m} R_{\varepsilon i}(|z_{1i}|) + \lambda_3 \sum_{i=1}^{m} R_{0i}(|z_{1i}|) \\ z_{1i} \quad & = f_1(x_i) + \varepsilon_1 - y_i \end{aligned}$$

and

$$\begin{aligned} \min_{w_2,b_2} \quad & \frac{1}{2}\|w\|^2 + \lambda_2 \sum_{i=1}^{m} R_{\varepsilon i}(|z_{2i}|) + \lambda_4 \sum_{i=1}^{m} R_{0i}(-z_{2i}) \\ z_{2i} \quad & = f_1(x_i) - \varepsilon_2 - y_i \end{aligned}$$

with training dataset with N data points. Both the ramp ε-insensitive loss functions are given below [66]:

$$R_{\varepsilon i}(|z|) = \begin{cases} 0, & |z| < \varepsilon, \\ |z| - \varepsilon, & \varepsilon \leq |z| \leq \nu_i, \\ \nu_i - \varepsilon, & |z| > \nu_i \end{cases} \tag{2.7}$$

and

$$R_{0i}(z) = \begin{cases} 0, & z < 0, \\ z, & 0 \leq z \leq \nu_i, \\ \nu_i, & z > \nu_i. \end{cases} \tag{2.8}$$

Similarly, OCSVM was also paired with various robust loss functions to add robustness to it. Motivated by the fact that the ramp loss function induces robustness to SVM, it is also used with OCSVM in 2017 [67]. Rescaled hinge loss function was also used with OCSVM to make the model robust towards noise and outliers [68].

Figure 2.1 shows the robust loss functions used with SVM or its variants to summarize the above-discussed work. As the surrogate loss functions were large in number,

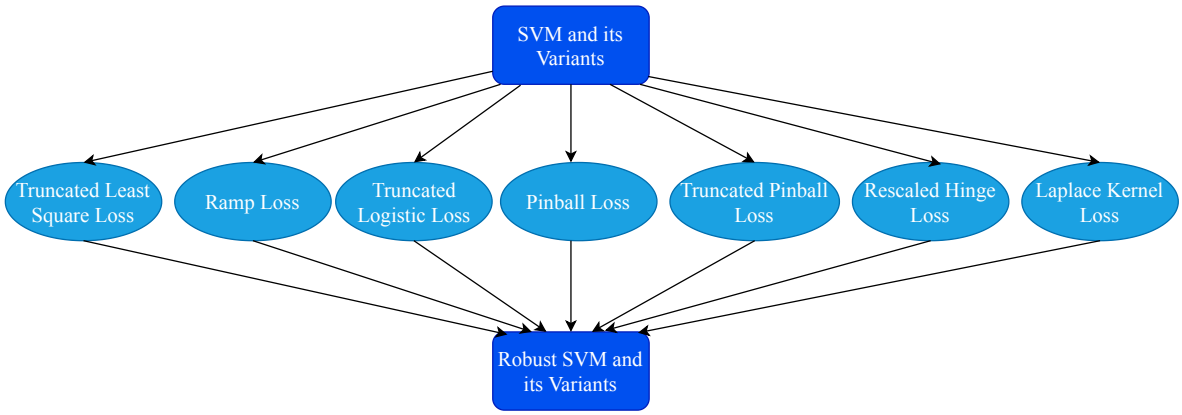

Figure 2.1: Robust Loss Functions Used with SVM and its Variants

another problem was to best find the surrogate loss function for the classification problem. To solve this problem, a method of importance reweighting was proposed by Liu and Tao [69] according to which any kind of surrogate loss function can be used with

a classification problem with label noise. It also solved the problem of obtaining the noise rate. This method depends on how accurately the conditional distribution P_{D_ρ} can be estimated where D_ρ is the distribution of corrupted variables. Both the efficiency and the robustness of the proposed approach were demonstrated experimentally with synthetic and real-world datasets [69].

2.4 Applications of Robust SVM

The robust SVM proposal has been applied in various areas, which can be an altogether different survey. This part of the chapter discusses some of the recent robust SVM applications, which can motivate the researchers to work in these directions.

In 2015, the SVM was used to formulate a two-class classification problem for voice activity detection (VAD). In that work, the SVM model was trained against various noise levels for speech and non-speech classification [70]. Similarly, in 2016, an SVM-based method was proposed to remove impulsive noise from grayscale images [71]. Roy et al. [71] used fuzzy filter-based classification, which classified all the test images as either noisy or non-noisy.

Robust SVM also played a significant role in the field of medical diagnosis. In 2018, Vijaylakshmi et al. proposed to deal with Ischemic stroke [72]. They performed magnetic resonance imaging (MRI) image analysis to detect the brain image's lesion tissue. Vijyalakshmi et al. [72] made use of Kernelized fuzzy c-means clustering with adaptive thresholding algorithm.

SVM was also used to detect noise in the speech signal [73]. That work made use of cumulative short-time Fourier transform for the classification of noise [73]. As various types of noises are present in the signal, SVM for multi-class classification was used.

When any classifier is used in digital image processing, its primary challenge is to remove noise from the images. SVM was also combined with low-rank matrix decomposition (LRMD) to denoise the image [1]. SVM was merged with LRMD because it

removes various types of noises from the image simultaneously. It was observed that the proposed method could remove both the noise and residual aliasing artifact from pMRI reconstructed noisy images [1].

The current applications of SVM include the detection of malicious Facebook posts through intelligent systems [74]. That research uses SVM as the classifier to accurately classify the posts into malicious and non-malicious.

Similarly, other recent SVM applications include the improvement in a human detection system for security and military applications [75]. Furthermore, the ensemble of SVM with kernel spherical k-means was also proposed, and this model was used for acute sinusitis classification [76]. The ensemble model of SVM was also used in the field of image communication like the one proposed by Wang et al. [77].

From the above literature, it can be concluded that the 'robust statistics based machine learning' is an active research field. In the next chapter, Chapter 3, the first contribution in this field is mentioned.