1 A Quadratic Mean Based Supervised Learning Model for Managing Data Skewness

This chapter reviews a novel approach for addressing data skewness in supervised learning models, presented in the paper "A Quadratic Mean Based Supervised Learning Model for Managing Data Skewness." The authors propose a framework called QMLearn, which introduces a new way of calculating empirical risk using the quadratic mean, aimed at improving model robustness on imbalanced datasets.

1.1 Introduction to the Problem of Data Skewness

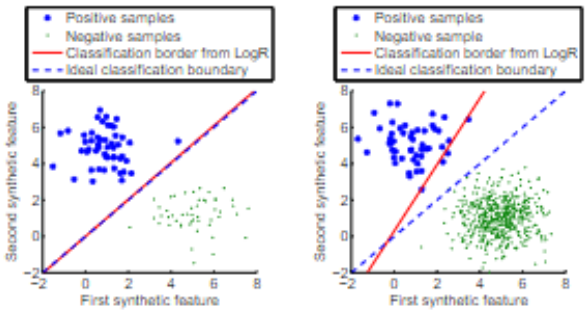
Data skewness, or imbalance, is a prevalent issue in supervised learning where the distribution of the dependent variable is uneven. This imbalance often causes traditional models to become biased towards the majority class, resulting in poor performance on the minority class. This paper identifies the limitations of traditional empirical risk minimization methods and introduces QMLearn to better handle skewed data.

1.2 Limitations of Traditional Learning Models

Traditional learning models, such as logistic regression and SVMs, typically minimize an empirical risk function defined as the arithmetic mean of the loss across all training examples:

$$R_{\text{emp}}(w) = \frac{1}{n} \sum_{i=1}^{n} l(x_i, y_i, w) \tag{1}$$

where n is the total number of training instances, x_i represents the feature vector, y_i the true label, and w the model parameters. This method often results in models that are biased towards the majority class, as illustrated in Figure 1.



(a) Class ratio = 1:1 (b) Class ratio = 1:10

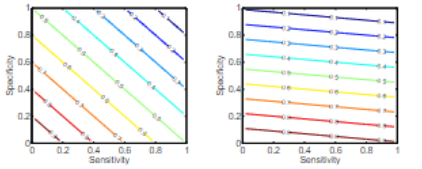
Figure 1: Classification boundaries using traditional logistic regression on balanced (left) and imbalanced (right) datasets.

1.3 Introduction to the QMLearn Framework

The QMLearn framework modifies the traditional empirical risk by using the quadratic mean rather than the arithmetic mean. This redefinition is designed to address the imbalance by equalizing the influence of both classes on the model's training process. The quadratic mean-based empirical risk function is defined as:

$$R^{Q}_{\text{emp}}(w) = \sqrt{\frac{\left(\frac{\sum_{i=1}^{n_1} l(x_i, y_i, w)}{n_1}\right)^2 + \left(\frac{\sum_{i=n_1+1}^{n} l(x_i, y_i, w)}{n_2}\right)^2}{2}} \tag{2}$$

where n_1 and n_2 are the number of instances in each class. This approach balances the error contributions from both classes, making the model more robust against skewed distributions.



(a) Original empirical risk on balanced data (b) Original empirical risk on imbalanced data (Pos:Neg = 1:10)

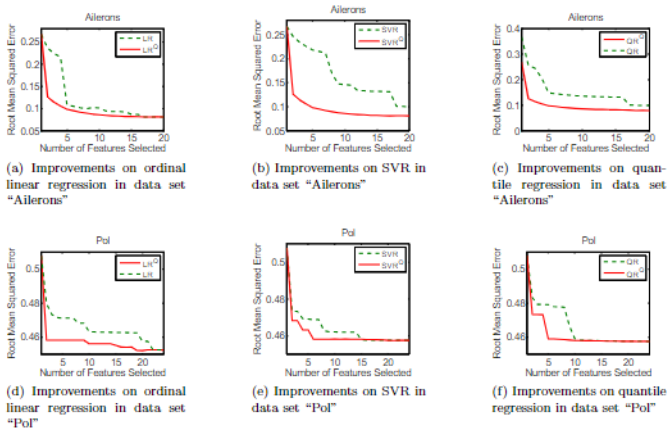
Figure 2: QMLearn-based empirical risk on balanced (left) and imbalanced (right) data.

1.4 Convex Optimization for Efficient Learning

The QMLearn method is formulated as a convex optimization problem, maintaining the convexity of the empirical risk function through the use of the quadratic mean. This allows for efficient computation and ensures that the solution is optimal, even for large-scale datasets.

1.5 Experimental Validation

Extensive experiments were conducted to validate the effectiveness of QMLearn compared to traditional models like logistic regression, SVMs, and quantile regression. The results, depicted in Figure 3, show that QMLearn consistently outperforms traditional methods, particularly on imbalanced datasets.



(a) Improvements on ordinal linear regression in data set "Ailerons"
(b) Improvements on SVR in data set "Ailerons"
(c) Improvements on quantile regression in data set "Ailerons"
(d) Improvements on ordinal linear regression in data set "Pol"
(e) Improvements on SVR in data set "Pol"
(f) Improvements on quantile regression in data set "Pol"

Figure 3: Performance improvements of QMLearn-based models in regression tasks on datasets 'Ailerons' and 'Pol'.

1.6 Using the QMLearn Package

To implement the QMLearn framework in your own work, you can install the QMLearn package by following the instructions available at qmlearn.rutgers.edu/source/install.html. For more details on installation and usage, please refer to the official QMLearn documentation.

1.7 Conclusion

The QMLearn framework offers a robust solution for managing data skewness in supervised learning models. By redefining the empirical risk function using the quadratic mean, QMLearn ensures balanced error consideration across classes, improving model performance on imbalanced datasets. Future research could explore the application of QMLearn to other types of models and further investigate its theoretical underpinnings.