ADDRESSING FAIRNESS ISSUES IN IMBALANCE DATASETS

Ting Yi Chuang

American University

ABSTRACT

Imbalanced data sets are an ongoing problem in machine learning, leading to model bias and reduced prediction accuracy. Addressing these fairness issues is critical to developing more general models. This project investigates the integration of Universum data into the training process of different Support Vector Machines (SVMs). By incorporating Universum data, the goal is to enhance SVM's ability to learn more precise decision boundaries, thereby improving the model's performance and fairness when dealing with imbalanced data. The findings show that using Universum data in certain datasets can be effective in reducing bias and predicting more accurate classes.

Keywords — Imbalanced Datasets, Fairness Issues, SVM, Universum Data, Machine Learning

1. INTRODUCTION

Imbalanced data sets pose a constant challenge to machine learning, a problem that often leads to model bias and reduced prediction accuracy. This imbalance may hinder the model's ability to generalize. Ensuring that machine learning models are fair and reduce bias is critical to achieving reliable and fair results. In recent years, various techniques have been developed to solve this problem, ranging from data resampling and feature selection to advanced algorithm modifications.

Support vector machines (SVM) and its variants are widely used in classification tasks due to their effectiveness in finding optimal decision boundaries [1]. However, traditional support vector machines encounter difficulties when processing unbalanced data sets. Usually, the results are biased towards the majority category, and other techniques need to be introduced to balance this problem. A recent article introduced the concept of Universum data, which consists of two types of data. By integrating Universum data into the training process, SVM models can learn more robust decision boundaries, thereby reducing bias and improving performance [2, 3].

Several studies have explored the use of Universum points in machine learning. For example, Vapnik et al. [4] proposed Universum SVM (USVM), which incorporates Universum points into the optimization framework. Extensions such as Twin SVM (TSVM) [5] and Quantum SVM (QSVM) [6] have also been adapted to include Universum data, resulting in models such as UTSVM [7], UQSVM [8] and UQTSVM [9]. Furthermore, recent innovations such as Adaptive Regularization Algorithm-Based UTSVM (ARABUTSVM) [10] and Kernel Weighted Regularized UTSVM (KWRUTSVM) [11] further refine the utilization of Universum data to improve classification in imbalanced environments.

This project studies the integration of Universum data in various SVM models, including SVM, USVM, TSVM, UTSVM, QSVM, UQSVM, QTSVM, and UQTSVM. Through extensive experiments, the study evaluates the efficacy of these models in improving fairness and accuracy. In order to solve the challenge of imbalanced data sets, transfer learning technology is adopted in combination with methods in deep learning [12, 13]. By leveraging these methods, the aim is to enhance the model's ability to identify precise decision boundaries, thereby promoting fair classification results.

The remainder of this paper is structured as follows. Section 2 describes the methodology and experimental setup, including the integration of Universum data and the training process of various SVM models. The results are presented and analyzed in Section 3, followed by a discussion of implications and limitations in Section 4. This is followed by the main conclusions in Section 5.

2. METHDOLOGY

This work employs SVM-based algorithms, including SVM, QSVM, USVM, and TSVM, for binary classification tasks. This method uses VGG-16 pre-trained on ImageNet for 4096-dimensional feature extraction and uses PCA for dimensionality reduction to simplify and improve the computational efficiency of the model.

2.1 Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised learning algorithm for binary classification. It seeks to find the optimal hyperplane that separates two types of data points with the maximum separation. The slack variable ξ is introduced to handle misclassification, and the penalty parameter C balances the trade-off between maximizing the margin and minimizing the classification error. The SVM objective function is defined as:

$$\min\frac{1}{2}\left||w|\right|^2 + c\sum_{i=1}^{n}\xi_i$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i\,, \qquad \xi_i \geq 0,$$

where w is the weight vector, b is the bias term, and y_i is the class label.

2.2 Universum Support Vector Machine (USVM)

USVM contains Universum points, which are domain-related but do not belong to any one category. These points guide hyperplane selection and improve classification accuracy. The USVM objective function is expressed as:

$$min\frac{1}{2}\left||w|\right|^2 + C_1\sum_{i=1}^{n}\xi_i + C_2\sum_{i=1}^{n}\eta_i,$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i\,, \qquad \xi_i \geq 0,$$

where η_i represents slack variables for Universum points, and C_1, C_2 are penalty parameters.

2.3 Twin Support Vector Machine (TSVM)

TSVM aims to solve two SVM-like optimization problems to produce two non-parallel hyperplanes, each closer to a class while maximizing separation from each other. The optimization of these two categories is defined as:

For Class1:

$$min\frac{1}{2}\left||w_1|\right|^2 + C1\sum_{i=1}^{n}\xi_i,$$

subject to:

$$y_i(w_1^T x_i + b_1) \geq 1 - \xi_i\,, \qquad \xi_i \geq 0.$$

For Class2:

$$min\frac{1}{2}\left||w_2|\right|^2 + C2\sum_{i=1}^{n}\eta_i,$$

subject to:

$$y_i(w_2^T x_i + b_2) \geq 1 - \eta_i\,, \qquad \eta_i \geq 0,$$

where C_1, C_2 are non-negative penalty parameter, ξ_i and η_i are non-negative slack variables.

2.4 Quadratic Support Vector Machine (QSVM)

QSVM extends traditional SVM by adding a quadratic term to the objective function, allowing it to model nonlinear decision boundaries more effectively. This flexibility is particularly beneficial for complex data sets. The QSVM objective function is expressed as:

$$min\sum_{i=1}^{m}\left||wx_i + b|\right|^2 + C\sum_{i=1}^{n}\xi_i$$

subject to:

$$y_i\left(\frac{1}{2}x_i^T w_{xi} + b^T x_i + c\right) \geq 1 - \xi_i, \qquad \xi_i \geq 0.$$

2.5 Quadratic Twin Support Vector Machine (QTSVM)

Quadratic Twin Support Vector Machine (QTSVM) is an enhanced version of Twin Support Vector Machine (TSVM) that incorporates quadratic terms in its formulation. By introducing these quadratic terms, QTSVM can produce more flexible and effective decision boundaries, which is particularly suitable for processing complex nonlinear data sets. QTSVM solves two optimization problems, each aiming to find a quadratic hyperplane that is closer to one class while maintaining a significant distance from the other class. The optimization of these two categories is defined as:

For Class1:

$$min\frac{1}{2}\left||w_1|\right|^2 + \frac{r_1}{2}\sum_{i=1}^{n}({x_i}^T w_1 + b_1)^2 + C1\sum_{i=1}^{n}\xi_i,$$

subject to:

$$y_i\left(\frac{1}{2}x_i^T w_1 x_i + b_1^T x_i + c_1\right) \geq 1 - \xi_i\,, \qquad \xi_i \geq 0,$$

where γ_1 is the regularization parameter for the quadratic term, C_1 is the penalty parameter, and ξ_i are slack variables.

For class 2:

$$min\frac{1}{2}\left||w_2|\right|^2 + \frac{r_2}{2}\sum_{i=1}^{n}({x_i}^T w_2 + b_2)^2 + C2\sum_{i=1}^{n}\eta_i,$$

subject to:

$$y_i\left(\frac{1}{2}x_i^T w_2 x_i + b_1^T x_i + c_2\right) \geq 1 - \eta_i , \qquad \eta_i \geq 0,$$

where γ_2 is the regularization parameter for the quadratic term, C_2 is the penalty parameter, and η_i are slack variables.

2.6 Evaluation Metrics

To evaluate the performance of the proposed SVM-based method, standard evaluation metrics such as accuracy, precision, recall, and F1 score were adopted. These metrics give you a comprehensive understanding of the model's efficiency and its ability to generalize to unseen data.

2.6.1 Accuracy: Accuracy is the indicator of the classifier making the correct prediction. It measures the proportion of correctly classified samples. The following equation describes the accuracy [14]:

$$Accuracy = \frac{TP + TN}{TN + FP + FN + TP}$$

where true positive, false negative, true negative and false positive are described by TP, FN, TN and FP, respectively.

2.6.2 Precision: Precision is an estimate of the ratio of true positive predictions to all positive predictions. High accuracy is associated with low False Positive Rate (FPR) as shown below [15]:

$$Precision = \frac{TP}{TP + FP}$$

2.6.3 F1 Score: A weighted average of precision and recall is combined into the F1 score to balance their trade-offs. It is used as a statistical measure to score the efficiency of a classifier. Therefore, this score considers both false positives and false negatives [16]:

$$F_1 = \frac{2(Recall)(Precision)}{(Recall + Precision)}$$

Each SVM variant was tested on a reduced feature set to determine the most efficient method for the classification task.

3. RESULTS

Obtain a balanced real-world cat and dog recognition [17], Fungi [18], Waste [19], QSAR [20], and Sentence [21] dataset composed of 1000 majority class samples and 1000 minority class samples from Kaggle and UCI. We introduce imbalance by randomly selecting 800 majority class samples without replacement and 400 minority class samples without replacement as the training set. The remaining 200 majority class samples and 200 minority class samples are used as the test set. Various SVM models are trained on the imbalanced training set and their performance is evaluated on the test set. Table 1 presents the test samples for all datasets.

Table 1 Number of training and testing samples for all the datasets used

ID	Dataset	Majority (-1)	Minority (1)	IR
1	Animal (Balance)	Dog (1000)	Cat (1000)	1
2	Animal (Imbalanced)	Dog (800)	Cat (400)	2
3	Fungi (Balance)	H6 (1000)	H1 (1000)	1
4	Fungi (Imbalanced)	H6 (800)	H1 (400)	2
5	Waste (Balance)	Metal (1000)	Glass (1000)	1
6	Waste (Imbalanced)	Metal (800)	Glass (400)	2
7	QSAR (Balance)	N (1000)	P (1000)	1
8	QSAR (Imbalanced)	N (800)	P (400)	2
9	Sentence (Balance)	N (1000)	P (1000)	1
10	Sentence (Imbalanced)	N (800)	N (400)	2

3.1 Performance on balanced datasets

We apply SVM, TSVM, QSVM, and QTSVM to three image-balanced datasets and two text-balanced datasets (Table 2, 3, 4, 5, 6), aiming to investigate whether flexible boundaries can improve classification performance.

Table 2 Classification performance of Cat and Dog Balanced data

	Accuracy	**Precision**	**F1 score**
SVM	95.88%	0.96	0.96
TSVM	95.62%	0.96	0.96
QSVM	95.62%	0.94	0.96
QTSVM	83.75%	0.99	0.81

Table 3 Classification performance of Fungi Balance data

	Accuracy	**Precision**	**F1 score**
SVM	96.17%	0.95	0.96
TSVM	96.00%	0.97	0.96
QSVM	95.12%	0.92	0.95
QTSVM	88.75%	0.94	0.88

Table 4 Classification performance of Waste Balance data

	Accuracy	**Precision**	**F1 score**
SVM	96.62%	0.96	0.97
TSVM	96.50%	0.95	0.97
QSVM	95.00%	0.93	0.95
QTSVM	95.17%	0.99	0.95

Table 5 Classification performance of QSAR Balance data

	Accuracy	**Precision**	**F1 score**
SVM	88.69%	0.82	0.90
TSVM	86.68%	0.78	0.87
QSVM	94.47%	0.92	0.95
QTSVM	74.87%	0.67	0.80

Table 6 Classification performance of Sentence Balance data

	Accuracy	Precision	F1 score
SVM	84.25%	0.87	0.84
TSVM	90.25%	0.92	0.90
QSVM	83.50%	0.78	0.85
QTSVM	91.75%	0.86	0.92

Among the five balanced data sets, SVM and TSVM perform stably and well on most image data sets (such as cats and dogs, fungi, waste), and usually have the highest accuracy (such as SVM is 96.62%, TSVM is 96.50%). and F1 score (both 0.97), showing its stability and reliability in image classification tasks. QSVM showed obvious advantages on the QSAR data set, with an accuracy of 94.47% and an F1 score of 0.95, showing that the quadratic kernel can effectively handle nonlinear boundaries on this data set. However, QSVM performs slightly worse than SVM and TSVM on other image datasets, while QTSVM performs mediocre on most image datasets, especially on fungi and cat and dog datasets, with accuracy rates of only 88.75% and 83.75%, the F1 scores are 0.88 and 0.81 respectively, showing limited adaptability. Notably, QTSVM performed best on the sentence dataset with an accuracy of 91.75% and an F1 score of 0.92, demonstrating its potential in text classification tasks. Overall, SVM and TSVM show stable performance on image datasets, while QSVM and QTSVM show greater potential on specific data types such as QSAR and sentence datasets.

3.1 Performance on imbalanced datasets

We apply SVM, USVM, TSVM, UTSVM, QSVM, UQSVM, QTSVM and UQTSVM to three image-balanced datasets and two text-balanced datasets (Table 7, 8, 9, 10, 11), aiming to study whether flexible boundaries and Universum points can improve classification performance.

Table 7 Classification performance of Cat and dog Imbalance data.

	Accuracy	Precision	F1 score
SVM	97.00%	0.98	0.97
USVM	95.75%	0.96	0.96
TSVM	93.75%	0.95	0.94
UTSVM	48.75%	0.48	0.32
QSVM	95.50%	0.97	0.95
UQSVM	96.50%	0.97	0.96
QTSVM	92.25%	0.98	0.92
UQTSVM	76.25%	0.98	0.69

Table 8 Classification performance of Fungi Imbalance data

	Accuracy	Precision	F1 score
SVM	97.00%	0.97	0.97
USVM	97.25%	0.96	0.97
TSVM	96.00%	0.96	0.96
UTSVM	61.50%	1.00	0.37
QSVM	96.00%	0.93	0.96
UQSVM	95.75%	0.93	0.96
QTSVM	84.50%	0.95	0.82
UQTSVM	88.75%	0.86	0.89

Table 9 Classification performance of Waste Imbalance data

	Accuracy	Precision	F1 score
SVM	99.50%	1.00	0.99
USVM	99.75%	1.00	1.00
TSVM	99.75%	1.00	1.00
UTSVM	97.75%	1.00	0.98
QSVM	99.75%	1.00	1.00
UQSVM	99.75%	1.00	1.00
QTSVM	99.25%	1.00	0.99
UQTSVM	98.50%	1.00	0.98

Table 10 Classification performance of QSAR Imbalance data

	Accuracy	Precision	F1 score
SVM	94.00%	0.89	0.94
USVM	94.50%	0.90	0.95
TSVM	89.50%	0.83	0.90
UTSVM	91.50%	0.85	0.92
QSVM	96.00%	0.95	0.96
UQSVM	91.00%	0.92	0.91
QTSVM	74.00%	0.66	0.79
UQTSVM	93.50%	0.88	0.94

Table 11 Classification performance of Sentence Imbalance data

	Accuracy	Precision	F1 score
SVM	52.25%	0.58	0.26
USVM	51.50%	0.53	0.34
TSVM	54.50%	0.59	0.40
UTSVM	52.50%	0.55	0.38
QSVM	50.25%	0.51	0.19
UQSVM	51.75%	0.51	0.22
QTSVM	94.42%	0.77	0.76
UQTSVM	50.75%	0.50	0.61

Among the five imbalanced data sets, SVM and USVM have the most stable and excellent overall performance, especially on the Cat and Dog, Fungi and QSAR data sets, with accuracy rates reaching 97.00%, 97.25% and 94.50% respectively, and F1 scores close to 0.97. Shows its good adaptability to imbalanced data. QSVM performed best on the QSAR dataset, with an accuracy of 96.00% and an F1 score of 0.96, indicating that quantum kernels have significant advantages in handling complex boundaries. However, QTSVM performs weakly on most image data (such as cats, dogs and fungi), with accuracy rates of only 92.25% and 84.50%, and F1 scores of 0.92 and 0.82 respectively, but it performs outstandingly on sentence datasets, with accurate the rate is as high as 94.42% and the F1 score is 0.76, showing its potential in text classification tasks. In contrast, UTSVM and UQTSVM perform poorly overall, especially in Fungi and sentence data, with accuracy rates below 62% and F1 scores much lower than other models, indicating their poor adaptability to imbalanced data. Notably, on the waste dataset, all models performed close to perfect, with accuracy and F1 scores close to 1.00, showing that this data feature may be very friendly to classification tasks. Overall, SVM and USVM are stable choices for imbalanced data classification, while QSVM and QTSVM have higher potential in specific data types, and UTSVM and UQTSVM require further improvements to improve their performance.

4. DISCUSSION OF THE RESULTS

The performance of classification models on balanced and unbalanced datasets highlights the different advantages and limitations of each approach. For balanced datasets, SVM and TSVM show consistent performance on all datasets to achieve high accuracy and balanced F1 scores. For example, SVM achieves 95.88% accuracy on the Cat and Dog dataset and 96.17% accuracy on the Fungi dataset, with F1 scores always around 0.96 or higher. These results demonstrate that SVM and TSVM are robust in the presence of uniform material distribution, as they effectively maximize class boundaries and minimize misclassification rates. QSVM also performs well on the balanced dataset, especially in handling nonlinear boundaries, although its performance is slightly lower than SVM and TSVM in some cases, as shown by the 95.12% accuracy on the Fungi dataset. In comparison, QTSVM struggled with balanced data, especially in the cat and dog datasets, where its accuracy dropped to 83.75% and F1 score of 0.81, indicating the challenges of effective generalization in simpler data distributions.

For imbalanced data sets, model performance varies significantly. SVM and USVM consistently show strong performance on multiple datasets, including Cat and Dog (accuracy of 97.00% and 95.75%, respectively) and Fungi dataset (accuracy of 97.00% and 97.25%, respectively), with close F1 scores 0.97. This demonstrates their ability to effectively handle data imbalances by maintaining robust decision boundaries. QSVM performed quite well in some cases, such as the QSAR dataset, where it achieved the highest accuracy (96.00%) with an F1 score of 0.96, reflecting its adaptability to nonlinear relationships. However, QTSVM shows mixed results, performing well on the Sentence dataset with an accuracy of 94.42% and an F1 score of 0.76, but performing poorly on other datasets such as QSAR (accuracy of 74.00% and an F1 score of 0.76). is 0.79). UTSVM and UQTSVM always perform poorly on imbalanced datasets, with lower accuracy and F1 scores, especially on the Fungi dataset (accuracy of 61.50% and 88.75%, respectively), which indicates the use of Universum in imbalanced environments. Limits of point and semi-supervised learning.

Overall, the results show that SVM and USVM are reliable choices for both balanced and unbalanced datasets, achieving consistently high performance. QSVM demonstrates strong potential for handling nonlinear material distributions, especially in imbalanced situations. However, QTSVM's inconsistent results suggest that its effectiveness depends on the data set type, performing well in textual data but poorly in others. UTSVM and UQTSVM need further optimization to improve their adaptability, especially for imbalanced data.

5. CONCLUSION

This study investigates the performance of various SVM-based models on balanced and imbalanced datasets. SVM and TSVM consistently demonstrate strong and reliable performance on both balanced and imbalanced datasets, achieving high accuracy and F1 scores, making them reliable choices for general classification tasks. By utilizing Universum points to enhance minority class generalization, USVM shows strong adaptability in imbalanced environments, and its performance is almost the same as SVM. QSVM is good at handling nonlinear and complex boundaries, especially in imbalanced datasets such as QSAR, demonstrating its potential in processing datasets with complex feature relationships. However, the performance of QTSVM varies widely, performing well on literal datasets (e.g., sentence datasets) but performing poorly on most balanced and some unbalanced datasets, suggesting that it has limited generalizability and is specific to the material type. potential specialization. UTSVM and UQTSVM consistently struggle with balanced and unbalanced datasets, indicating the need for further improvements in Universum point utilization and semi-supervised learning strategies.

Overall, SVM and USVM emerged as the most reliable models across different material distributions, while QSVM provided advantages for complex nonlinear data sets. QTSVM shows great promise in special scenarios such as text data classification, but it still needs to be improved in the future to achieve wider applicability. Addressing the challenges faced by UTSVM and UQTSVM is crucial to improve their adaptability and performance in real-world imbalanced classification problems.

6. REFERENCES

[1] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[2] Vapnik, V. N. (2006). Estimation of Dependencies Based on Empirical Data. Springer.

[3] Weston, J., et al. (2006). Incorporating prior knowledge with SVMs using Universum data. NIPS.

[4] Vapnik, V., & Izmailov, R. (2015). Learning using privileged information: SVM+ and Universum. Machine Learning, 85(1), 3–30.

[5] Khemchandani, R., & Chandra, S. (2007). Twin support vector machines for pattern classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(5), 905–910.

[6] Schuld, M., Sinayskiy, I., & Petruccione, F. (2015). An introduction to quantum machine learning. Contemporary Physics, 56(2), 172–185.

[7] Tanveer, M., et al. (2019). Universum twin SVM for binary and multiclass imbalanced classification. Knowledge-Based Systems, 165, 235–246.

[8] Chen, Y., & Lin, C. (2021). Quantum-enhanced support vector machine for imbalanced datasets. Quantum Machine Intelligence, 3(1), 1–13.

[9] Qiu, J., et al. (2022). Universal quantum twin SVM for imbalanced classification problems. Expert Systems with Applications, 200, 117043.

[10] Xie, H., et al. (2023). Adaptive regularization algorithms for Universum twin SVMs. Pattern Recognition Letters, 172, 109–117.

[11] Singh, A., et al. (2023). Kernel weighted regularized twin support vector machines. Journal of Machine Learning Research, 24(6), 1–27.

[12] Krizhevsky, A., et al. (2012). ImageNet classification with deep convolutional neural networks. NIPS.

[13] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1–48.

[14] Long, M., Cao, Y., Cao, Z., et al.: 'Transferable representation learning with deep adaptation networks', IEEE Trans. Pattern Anal. Mach. Intell., 2018, 41, (12), pp. 3071–3085

[15] Long, M., Cao, Y., Cao, Z., et al.: 'Transferable representation learning with deep adaptation networks', IEEE Trans. Pattern Anal. Mach. Intell., 2018, 41, (12), pp. 3071–3085

[16] Long, M., Cao, Y., Cao, Z., et al.: 'Transferable representation learning with deep adaptation networks', IEEE Trans. Pattern Anal. Mach. Intell., 2018, 41, (12), pp. 3071–3085

[17] Will Cukierski. (2013). Dogs vs. Cats. Kaggle. https://kaggle.com/competitions/dogs-vs-cats

[18] Kotzias, D. (2015). Sentiment Labelled Sentences [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C57604.

[19] Hajati, F., Javier Pineda Sopo, C., Hajati, F., & Gheisari, S. (2021). DeFungi [Dataset]. UCI Machine Learning Repository. https://doi.org/10.48550/arXiv.2109.07322.

[20] QSAR androgen receptor [Dataset]. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C53317.

[21] Single, S., Iranmanesh, S., & Raad, R. (2023). RealWaste [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5SS4G.