ADDRESSING FAIRNESS ISSUES IN IMBALANCE DATASETS

Yen Chun Lin

American University, Client: Ahmad Mousavi

ABSTRACT

Imbalanced datasets pose a significant challenge to machine learning models, often leading to a bias toward the majority class while overlooking critical information in the minority class. This study focuses on addressing fairness issues in such datasets by employing various Support Vector Machine (SVM) models. The approach involves creating more flexible decision boundaries and generating Universum points to enhance classification performance for the minority class while maintaining overall accuracy. Experiments conducted on real-world datasets demonstrate that these methods effectively mitigate the effects of imbalance and improve predictive performance. This research provides valuable tools for achieving balanced and fair decision-making in scenarios with imbalanced data.

Index Terms— Support Vector Machine, Universum points, Imbalanced Datasets

1. INTRODUCTION

The class imbalance problem arises from unequal sample sizes between classes in a dataset. This issue is common in many real-world scenarios, such as disease diagnosis and fraud detection. Typically, the minority class is of greater interest to analysts, as it often holds critical information. However, standard machine learning algorithms tend to favor the majority class, leading to overfitting or the neglect of the minority class—precisely the class that requires more focus.
This study aims to address the class imbalance problem by leveraging variants of Support Vector Machine (SVM) models. These models incorporate more flexible boundaries and generate Universum points between the two classes, providing additional information to enhance the classifier's ability and improve classification performance. By employing these approaches, the study seeks to explore effective methods for addressing real-world class imbalance challenges.
If you have any questions, please contact Yen Chun Lin at 346-907-1418 or via email at yl5496a@american.edu.

2. RELATED WORK

2.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) [1] is a machine learning method for classification, it aims to identify an optimal separating hyperplane that effectively divides data points into two classes by maximizing the margin between them [2,3]. The foundational work by Cortes and Vapnik [1] introduced SVM and established its ability to handle linearly separable datasets effectively. The SVM hyperplane is determined by minimizing the following objective function:

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

Subject to:

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

Where C is a non-negative penalty parameter and ξ_i represents slack variables to account for noises.
The present study builds upon this foundational work by exploring variants of SVM, such as QSVM and TSVM, which address limitations of traditional SVM, particularly when dealing with non-linear and imbalanced datasets.

2.2. Quadratic Support Vector Machine (QSVM)

QSVM is an extension of the standard SVM that introduces a quadratic term into the objective function, resulting in a quadratic decision boundary. This modification makes QSVM more flexible and capable of handling more complex classification problems, particularly with non-linear datasets. By incorporating the quadratic term, QSVM can create a more effective classifier and improve the accuracy rate for the data. The QSVM optimization problem is formulated as:

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

Subject to:

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

This work extends prior research by evaluating QSVM's effectiveness in imbalanced datasets and comparing its performance to other SVM variants, particularly when integrated with Universum points.

2.3. Universum Support Vector Machine (USVM)

The integration of Universum points into SVM represents an advancement in improving classifier generation. Universum points are data points that do not belong to either class in the dataset but are relevant to the domain [4,5]. They can provide valuable information to improve the classifier and enhance the accuracy rate. Weston et al. [4] formulated the USVM framework as:

$$\frac{1}{2}\|w\|^2 + C_1\sum_{i=1}^{1}\xi_i + C_2\sum_{i=1}^{1}\eta_i$$

Subject to:

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

$$-\varepsilon - \eta_s \leq (w^T x_s + b) \leq \varepsilon + \eta_s, \eta_s \geq 0, s = 1, \ldots u$$

Where *C1* and *C2* are non-negative penalty parameter, ξ_i accounts for slack in target data points, and η_s regulates Universum points. The Universum points (u), should ideally lie within the epsilon (ε) band, positioned between the positive (minority) and negative (majority) class. This study builds on this by focusing on how Universum points enhance classification performance, particularly in imbalanced datasets.

2.4. Twin Support Vector Machine (TSVM)

The Twin Support Vector Machine (TSVM) is another variant of the traditional SVM [6], modifies traditional SVM by generating two non-parallel hyperplanes. Each closer to one class while maximizing the margin from the other class. The TSVM model is formulated as:

For Class 1:

$$min\left(\frac{1}{2}\|w_1\|^2 + C1\sum_{i=1}^{n}\xi_i\right), \qquad \xi_i > 0$$

Subject to:

$$y_i(w_1^T x_{1i} + b_1) \geq 1 - \xi_i\,, \qquad \forall x \in C_2$$

For Class 2:

$$min\left(\frac{1}{2}\|w_2\|^2 + C2\sum_{i=1}^{n}\eta_i\right), \qquad \eta_i > 0$$

Subject to:

$$y_i(w_2^T x_{2i} + b_2) \geq 1 - \eta_i, \qquad \forall x \in C_1$$

Where *C1* and *C2* are non-negative penalty parameter, ξ_i and η_i are non-negative slack variables. The data points of Class 2 should be one unit away from the hyperplane of Class 1, while the data points of Class 1 should also be one unit away from the hyperplane of Class 2. This study examines the application of TSVM to imbalanced datasets, comparing it with other SVM models in terms of classification metrics.

This study integrates Universum points into the SVM, QSVM, and TSVM frameworks to address challenges in imbalanced classification tasks. Unlike previous research, which examined each method independently, this work combines their strengths and evaluates them comprehensively. By employing detailed performance metrics, such as overall accuracy and recall, the study demonstrates improvements in both minority class classification and overall performance, highlighting the enhanced applicability of these methods to real-world datasets.

3. EXPERIMENT

3.1. Experiment Setting:

Using 5 balanced real-world datasets, each containing 1000 majority class samples and 1000 minority class samples, we introduce an imbalance by randomly selecting 800 majority class samples without replacement and 400 minority class samples (without replacement) for the training set. The remaining 200 samples from each class are designated as the test set. Various SVM models are trained on the imbalanced training set, and their performance is evaluated on the test set. This experimental setup allows us to explore how SVM models, and the use of Universum points, can enhance classification performance in imbalanced scenarios.

3.2. Data Description:

In this study, we utilized five real-world datasets sourced from UCI Machine Learning Repository and Kaggle [7-11]. The class imbalance ratio (IR) was fixed at 2 for the imbalanced setup. The imbalance ratio is calculated as the size of the majority class divided by the size of the minority class. For further information, please refer to Tables 1 and 2.

Table 1: Balance datasets information

ID	Dataset	Majority (-1)	Minority (1)
1	Face Recognize (B)	White (1000)	Asian (1000)
2	Animal (B)	Horse (1000)	Deer (1000)
3	Brain Tumor (B)	Glioma (1000)	Meningoma (1000)
4	Breast Cancer (B)	Bengin (1000)	Malignant (1000)
5	Waste (B)	Plastic (1000)	Paper (1000)

Table 2: Imbalanced datasets information

ID	Dataset	Majority (-1)	Minority (1)
6	Face Recognize (Im)	White (800)	Asian (400)
7	Animal (Im)	Horse (800)	Deer (400)
8	Brain Tumor (Im)	Glioma (800)	Meningoma (400)
9	Breast Cancer (Im)	Bengin (800)	Malignant (400)
10	Waste (Im)	Plastic (800)	Paper (400)

3.3. Accuracy metrics:

3.3.1. Accuracy Rate:

The overall accuracy rate in SVM models reflects how well the model predicts both classes in the dataset. However, in cases of class imbalance, the accuracy rate may disproportionately favor the majority class. This means it might not accurately represent the model's ability to predict instances from each class. A high accuracy rate in such scenarios could indicate potential overfitting, particularly for imbalanced datasets.

3.3.2. Recall:

A confusion matrix is a table that illustrates the performance of a classification model [12]. It provides information such as precision, recall, F1 score, and support. In this study, we will mainly focus on recall, a metric that indicates the correct classification rate. Specifically, recall shows how many instances of the majority and minority classes are correctly classified.

3.4 Models:

This study introduces four SVM models (SVM, TSVM, QSVM, and QTSVM) for both balanced and imbalanced datasets, as well as four additional models incorporating Universum points specifically for imbalanced datasets (USVM, UTSVM, UQSVM, and UQTSVM).

SVM: SVM is a supervised learning algorithm that finds the optimal hyperplane to separate data into distinct classes. It works well for balanced datasets but struggles with imbalanced ones.

QSVM (Quadratic SVM): QSVM extends SVM by employing a quadratic decision surface instead of a linear hyperplane, allowing it to capture more complex patterns. It is particularly useful when the data distribution is nonlinear.

TSVM (Twin SVM): TSVM constructs two nonparallel hyperplanes, each closer to one class and farther from the other. This design improves computational efficiency and enhances classification in certain scenarios.

QTSVM (Quadratic Twin SVM):QTSVM combines the concepts of TSVM and QSVM by using quadratic surfaces for the twin hyperplanes. It is designed to handle complex and nonlinear data distributions effectively.

USVM (Universum SVM): USVM incorporates Universum points between classes to provide additional information for training. In imbalanced datasets, this could be helpful to improve generalization and robustness.

UQSVM (Universum Quadratic SVM): UQSVM combines Universum points with QSVM, leveraging quadratic decision boundaries for better performance in imbalanced or nonlinear data. The additional information from Universum points enhances its classification capability.

UTSVM (Universum Twin SVM): UTSVM integrates Universum points into TSVM, improving the twin-hyperplane model's ability to generalize. This adaptation helps mitigate the challenges posed by imbalanced datasets.

UQTSVM (Universum Quadratic Twin SVM): UQTSVM combines the strengths of Universum points, quadratic decision boundaries, and twin hyperplanes. It offers a powerful approach for tackling nonlinear and imbalanced classification problems.

3.5. Analysis Results

This section focuses on evaluating the classification performance of various SVM models to examine how Universum points and employing flexible decision boundaries can enhance classification performance.

3.5.1. Performance on real world balance datasets

We apply SVM, TSVM, QSVM, and QTSVM to both balanced and imbalanced datasets, aiming to investigate whether flexible boundaries improve classification performance.

Table 3: Classification performance of balance datasets.

ID	Data	SVM	TSVM	QSVM	QTSVM
1	Face	78.5% (78,79)	77.5% (80,75)	**80.17% (86,74)**	73.67% (74,73)
2	Animal	83.33% (87,80)	**88.33% (90,87)**	**88.33% (88,87)**	88.33% (93,83)
3	Brain	87% (86,88)	88.17% (86,90)	**92.33% (90,95)**	90% (85,95)
4	Breast	78.33% (87,70)	**81.67% (88,75)**	80.83% (80,82)	71.76 (67,77)
5	Waste	91.33% (94,89)	92.33% (93,91)	**94% (94,94)**	91% (84.98)

Table 3 display classification performance across balanced datasets using different SVM variants. Each column displays the accuracy (%) and recall score (%) for the majority class (-1) and minority class (1), with ranges indicated in parentheses. In balanced datasets, the overall classification performance for each dataset was moderately good. All SVM models recognized over 75% of the data points from both classes. Furthermore, the recall score for each class was also balanced, which indicates that the overfitting issue in

balanced datasets was not significant. The SVM models did not favor either class in the dataset.

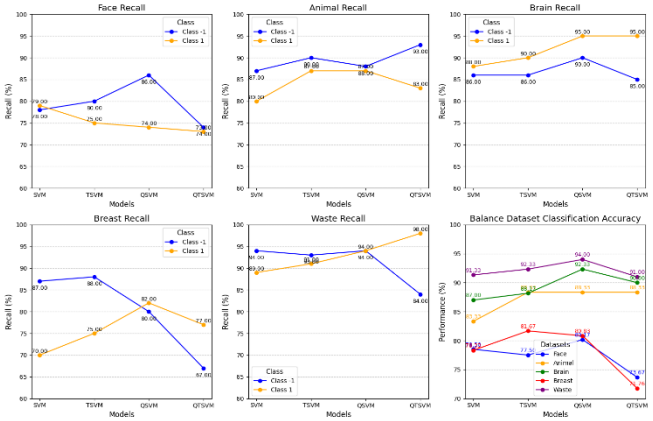

Fig 1: Classification performance of 5 balance datasets

TSVM and QSVM performed better. The overall accuracy predicted by TSVM and QSVM was the highest, and based on the recall scores, both TSVM and QSVM recognized more data points for each class compared to the other SVM models. Using a more flexible boundary improved classification performance.

3.5.2. Performance on real world imbalanced datasets

For imbalanced datasets, four additional models were applied: USVM, UTSVM, UQSVM, and UQTSVM. The aim was to investigate whether Universum points improve classification performance.

Table 4: Classification performance of Imbalance data

ID	Data	SVM	TSVM	QSVM	QTSVM
6	Face	72% (79,66)	73.5% (83,65)	74.25% (90,58)	**72.5% (66,63)**
7	Animal	87.5% (100,75)	60% (95,25)	82.5% (95,70)	**75% (75,75)**
8	Brain	86.25% (92,81)	89% (95,83)	**91.25% (97,85)**	81.25% (67,97)
9	Breast	67.5% (88,47)	71.5% (90,53)	70% (82,57)	**73.75% (72,75)**
10	Waste	**91% (90,92)**	92% (94,90)	92.5% (97,88)	76.5% (90,85)

Table 4 presents the classification performance across imbalanced datasets using different SVM models with flexible boundaries. Each column displays the accuracy (%) and recall score (%) for the majority class (-1) and the minority class (1).

In imbalanced datasets, the overall classification performance for each dataset was moderately good, with all SVM models recognizing over 67% of the data points. However, the accuracy rates suggest a potential overfitting issue. According to the recall scores, the SVM models could perfectly recognize data points belonging to the majority class, achieving 79–100% recall for this class. In contrast, the classification performance for the minority class was notably lower than that of the majority class.

The recall scores achieved by QTSVM demonstrated a more balanced prediction between the two classes. By utilizing a more flexible boundary, QTSVM was able to predict data points from both classes more equitably. The recall scores for the majority and minority classes were more balanced compared to other SVM models

Table 5: Classification performance of Imbalance data

ID	Data	USVM	UTSVM	UQSVM	UQTSVM
6	Face	74.25% (82,67)	74.25% (85,66)	73% (93,53)	80% (100,66)
7	Animal	82.5% (95,70)	87.5% (95,80)	87.5% (95,80)	60% (90,30)
8	Brain	89.25% (92,86)	89.25% (94,84)	90.25% (96,84)	85.05% (70,100)
9	Breast	73.75% (82,65)	70% (85,55)	70% (85,55)	40% (38,42)
10	Waste	92.5% (94,91)	93.5% (95,92)	93% (96,90)	87.5% (90,85)

Table 5 presents the classification performance across imbalanced datasets using different SVM models with Universum points. Each column shows the accuracy (%) and recall score (%) for the majority class (-1) and the minority class (1).

In imbalanced datasets, Universum points are generated to provide additional information to the classifier, with the aim of improving accuracy and enhancing the classifier's ability to recognize data points in the minority class. However, as shown in Tables 4 and 5, the inclusion of Universum points did not result in significant improvements in either the accuracy rate or the recall score for the minority class. In this case, Universum points did not have a substantial impact on the classifier's performance.

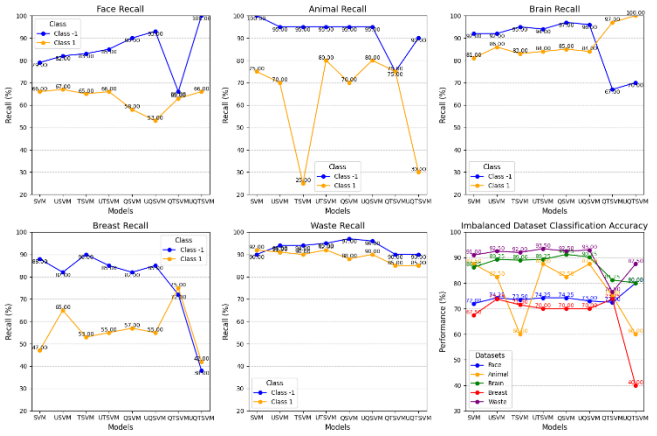

Fig 2: Classification performance of 5 imbalanced datasets

For imbalanced datasets, QTSVM can offer balanced classification performance across both classes, suggesting that using a flexible decision boundary can help identify more data points in the minority class. In this study, the impact of Universum points was not significant enough to enhance classification performance for the minority class.

4. DISCUSSION

In imbalanced datasets, generating Universum points or using flexible boundaries can improve classification performance. However, the analysis results in this study indicate that the improvement is only slight and not statistically significant. Several factors may explain this outcome.

One possible reason is that the sample size is not large enough, and the imbalance ratio (800:400) is not particularly extreme. When the dataset size is limited, the model's ability to generalize from the training data is constrained. Universum points are designed to enhance the decision boundary by introducing additional prior knowledge, but their effectiveness may be restricted by the small number of target class samples. A larger dataset would allow the model to better capture the true data distribution, potentially resulting in a more significant impact of Universum points and flexible boundaries.

Moreover, with a small dataset, the added Universum points might not sufficiently refine the decision boundary, especially if the model is prone to overfitting due to the limited amount of target class data. While Universum points aim to provide useful information for boundary adjustments, their benefits are diminished when there is insufficient training data to fully leverage them.

On the other hand, the sample size ratio of 800:400, while imbalanced, is not severe enough to fully showcase the benefits of these techniques. In cases of moderate imbalance, the decision boundary may not be significantly biased, reducing the added value of Universum points and flexible boundaries. For datasets with more pronounced imbalance or a larger sample size, these methods might yield more substantial improvements.

5. CONCLUSION

Imbalanced datasets are common in real-world scenarios, such diseases diagnosis and fraud detection. It often led to overfitting or inaccurate analysis, as machine learning algorithms tend to favor the majority class. However, the minority class is typically the focus of interest and holds greater significance in many applications.

To tackle the challenges posed by imbalanced datasets, this study examined various SVM models aimed at improving classification performance. We employed more flexible hyperplanes, such as quadratic decision boundaries, to better distinguish between the classes. Additionally, we introduced Universum points between the two classes to provide supplementary information for the classifier.

The results demonstrate that both flexible boundaries and Universum points can enhance classification performance. For balanced datasets, models with flexible boundaries, such as TSVM and QSVM, exhibited improved classification performance. In contrast, for imbalanced datasets, QTSVM performed the best, although the introduction of Universum points did not result in significant improvements.

These approaches are particularly useful for addressing the challenges of imbalanced datasets, enabling machine learning models to better prioritize the minority class, which is often crucial in practical applications

6. REFERENCES

[1] Cortes, C., & Vapnik, V. (1995, September). Support-vector networks - machine learning. SpringerLink. https://link.springer.com/article/10.1007/BF00994018

[2] Cosma, G., Brown, D., Archer, M., Khan, M., & Pockley, A. G. (2016, November 9). A survey on Computational Intelligence Approaches for predictive modeling in prostate cancer. Expert Systems with Applications. https://www.sciencedirect.com/science/article/pii/S0957417416306297

[3] Pisner, D. A., & Schnyer, D. M. (2019, November 15). Support Vector Machine. Machine Learning. https://www.sciencedirect.com/science/article/pii/B9780128157398000067

[4] Weston, J., Collobert, R., Sinz, F., Bottou, L., & Vapnik, V. (2006, June 25). *Inference with the Universum: Proceedings of the 23rd International Conference on Machine Learning*. ACM Other conferences. https://dl.acm.org/doi/abs/10.1145/1143844.1143971

[5] Liu, D., Fu, S., Tian, Y., & Tang, J. (2024, January 9). Universum driven cost-sensitive learning method with asymmetric loss function. Engineering Applications of

Artificial Intelligence. https://www.sciencedirect.com/science/article/pii/S0952197624000071#b37

[6] Jayadeva, R. Khemchandani and S. Chandra, "Twin Support Vector Machines for Pattern Classification," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 5, pp. 905-910, May 2007, doi: 10.1109/TPAMI.2007.1068

[7] Antoreepjana. (2021, May 5). Animals Detection Images Dataset. Kaggle. https://www.kaggle.com/datasets/antoreepjana/animals-detection-images-dataset

[8] Chakrabarty, N. (2019, April 14). Brain MRI images for Brain tumor detection. Kaggle. https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection

[9] Nickparvar, M. (2021a, September 24). Brain tumor MRI dataset. Kaggle. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[10] Patel, V. (2020, November 6). Face recognition dataset. Kaggle. https://www.kaggle.com/datasets/vasukipatel/face-recognition-dataset

[11] UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/908/realwaste

[12] Singh, P., Singh, N., Singh, K. K., & Singh, A. (2021). Diagnosing of disease using machine learning. ScienceDirect Topics. https://www.sciencedirect.com/topics/engineering/confusion-matrix